# Testing High Dimensional Mean Under Sparsity

Xianyang Zhang*

This version: August 24, 2017

**Abstract** Motivated by the likelihood ratio test under the Gaussian assumption, we develop a maximum sum-of-squares test for conducting hypothesis testing on a high dimensional mean vector. The proposed test, which incorporates the dependence among the variables, is designed to ease the computational burden and to maximize the asymptotic power of the likelihood ratio test. A simulation-based approach is developed to approximate the sampling distribution of the test statistic. The validity of the testing procedure is justified under both the null and alternative hypotheses. We further extend the main results to the two-sample problem without the equal covariance assumption. Numerical results suggest that the proposed test can be more powerful than some existing alternatives.

**Keywords:** High dimensionality, Maximum type test, Simulation-based approach, Sparsity, Sum-of-squares test

## 1 Introduction

Due to technological advancement, modern statistical data analysis often deals with high-dimensional data arising from many areas such as biological studies. High dimensionality poses significant challenges to hypothesis testing. In this paper, we consider a canonical hypothesis testing problem in multivariate analysis, namely inference on a mean vector. Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d $p$-dimensional random vectors with $\mathbb{E}X_i = \theta$. We are interested in testing

$$H_0 : \theta = 0_{p\times 1} \text{ versus } H_a : \theta \neq 0_{p\times 1}.$$

When $p \ll n$, the Hotelling's $T^2$ test has been shown to enjoy some optimal properties for testing $H_0$ against $H_a$ [Anderson (2003)]. However, for large $p$, the finite sample performance of the Hotelling's $T^2$ test is often unsatisfactory. To cope with the high dimensionality, several alternative approaches have been suggested, see e.g Bai and Saranadasa (1996), Srivastava and Du (2008), Srivastava (2009), Chen and Qin (2010), Lopes et al. (2011), Cai et al. (2014), Gregory et al.

*Department of Statistics, Texas A&M University, College Station, TX 77843, USA. E-mail: zhangxiany@stat.tamu.edu. The author thanks conference participants at the Multivariate Statistics Workshop at Texas A&M University.

(2015) and references therein. In general, these tests can be categorized into two types: the sum-of-squares type test and the maximum type test. The former is designed for testing dense but possibly weak signals, i.e., $\theta$ contains a large number of small non-zero entries. The latter is developed for testing sparse signals, i.e., $\theta$ has a small number of large coordinates. In this paper, our interest concerns the case of sparse signals which can arise in many real applications such as detecting disease outbreaks in early stage, anomaly detection in medical imaging [Zhang et al. (2000)] and ultrasonic flaw detection in highly scattering materials [James et al. (2001)].

Suitable transformation of the original data, which exploits the advantages of the dependence structure among the variables, can lead to magnified signals and thus improves the power of the testing procedure. This phenomenon has been observed in the literature [see e.g. Hall and Jin (2010); Cai et al. (2014); Chen et al. (2014); Li and Zhong (2015)]. To illustrate the point, we consider the signal $\theta = (\theta_1, \ldots, \theta_{200})'$, where $\theta$ contains 4 nonzero entries whose magnitudes are all equal to $\psi_j\sqrt{\log(200)/100} \approx 0.23\psi_j$ with $\psi_j$ being independent and identically distributed (i.i.d) random variables such that $P(\psi_j = \pm 1) = 1/2$. Let $\Sigma = (\sigma_{i,j})_{i,j=1}^{200}$ with $\sigma_{i,j} = 0.6^{|i-j|}$, and $\Gamma = \Sigma^{-1}$. Figure 1 plots the original signal $\theta$ as well as the signal after transformation and studentization $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_{200})'$ with $\widetilde{\theta}_j = (\Gamma\theta)_j/\sqrt{\gamma_{jj}}$. It is clear that the linear transformation magnifies both the strength and the number of signals (the number of nonzero entries increases from 4 to 12 after the linear transformation). In the context of signal recovery, the additional nonzero entries are treated as fake signals and need to be excised, but they are potentially helpful in simultaneous hypothesis testing as they carry certain information about the presence of signal. In general, it appears to be sensible to construct a test based on the transformed data, which targets not only the largest entry [see Cai et al. (2014)] but also other leading components in $\widetilde{\theta}$. Intuitively such a test can be constructed based on $\bar{Z} = \Gamma\bar{X}$ with $\bar{X} = \sum_{i=1}^{n} X_i/n$ being the sample mean, which serves as a natural estimator for $\Gamma\theta$. For known $\Gamma$, a test statistic which examines the largest $k$ components of $\Gamma\theta$ can be defined as,

$$T_n(k) = n \max_{1 \leq j_1 < j_2 < \cdots < j_k \leq p} \sum_{s=1}^{k} \frac{\bar{z}_{j_s}^2}{\gamma_{j_s,j_s}}.$$

If the *original* mean $\theta$ contains exactly $k$ nonzero components with relatively strong signals, it seems reasonable to expect that $T_n(k)$ outperforms $T_n(1)$. Interestingly, we shall show that the test statistic $T_n(k)$ is closely related to the likelihood ratio (LR) test for testing $H_0$ against a sparse alternative on $\theta$.

Our derivation also provides insight on some methods in the literature. In particular, we show that the data transformation based on the precision matrix proposed in Cai et al.(2014) can be derived explicitly using the maximum likelihood principle when $\Theta_a$ is the space of vectors with exactly one nonzero component. We also reveal a connection between $T_n(k)$ and the thresholding test in Fan (1996).

The rest of the paper is organized as follows. Adopting the maximum likelihood viewpoint, we develop a new class of tests named maximum sum-of-squares tests in Section 2.1. In Section 2.2,

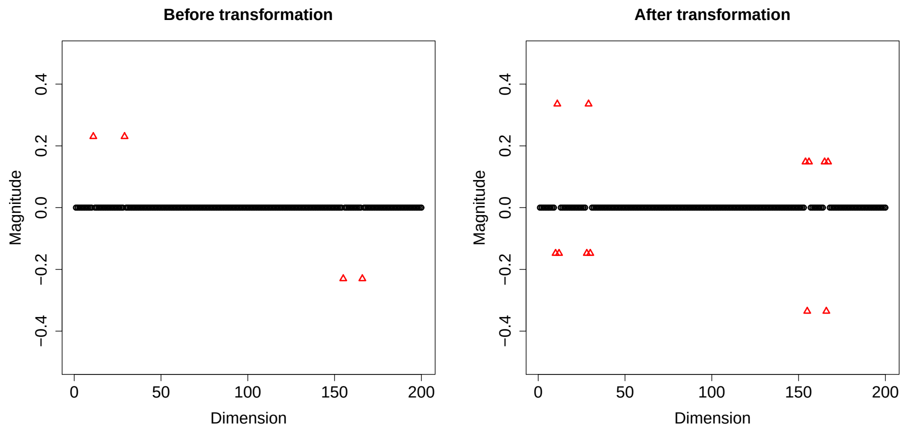



Figure 1: Signals before (left panel) and after (right panel) the linear transformation $\Gamma$. The non-zero entries are denoted by $\triangle$.

we introduce the feasible testing procedure by replacing the precision matrix by its estimator. A simulation-based approach is proposed to approximate the sampling distribution of the test. We describe a modified testing procedure in Section 2.3. Section 2.4 presents some theoretical results based on the Gaussian approximation theory for a high dimensional vector. We extend our main results to the two-sample problem in Section 3. Section 4 reports some numerical results. The method is illustrated via a data example arising from biological studies in Section 5. Section 6 concludes. The technical details are deferred to the technical appendix.

*Notation.* For a vector $a = (a_1, \dots, a_p)'$ and $q > 0$, define $|a|_q = (\sum_{i=1}^p |a_i|^q)^{1/q}$ and $|a|_\infty = \max_{1\le j\le p} |a_j|$. Set $|\cdot| = |\cdot|_2$. Denote by $||\cdot||_0$ the $l_0$ norm of a vector and let card$(\cdot)$ be the cardinality of a set. For $C = (c_{ij})_{i,j=1}^p \in \mathbb{R}^{p\times p}$, define $||C||_1 = \max_{1\le j\le p}\sum_{i=1}^p |c_{ij}|$, $||C||_2 = \max_{|a|=1}|Ca|$ and $||C||_\infty = \max_{1\le i,j\le p}|c_{ij}|$. Denote by diag$(C)$ the diagonal matrix diag$(c_{11}, c_{22}, \dots, c_{pp})$. For $S \subseteq \{1, 2, \dots, p\}$, let $C_{S,S}$ be the submatrix of $C$ that contains the rows and columns in $S$. Similarly, define $C_{S,-S}$ with the rows in $S$ and the columns in $\{1, 2, \dots, p\} \setminus S$. The notation $N_p(\theta, \Sigma)$ is reserved for the $p$-variate normal distribution with mean $\theta$ and covariance matrix $\Sigma$.

# 2 Main results

## 2.1 Likelihood ratio test

Let $X_i = (x_{i1}, \dots, x_{ip})'$ be a sequence of i.i.d $N_p(\theta, \Sigma)$ random vectors with $\Sigma = (\sigma_{ij})_{i,j=1}^p$. We are interested in testing

$$H_0 : \theta \in \Theta_0 = \{0_{p\times 1}\} \quad \text{versus} \quad H_a : \theta \in \Theta_a \subseteq \Theta_0^c. \tag{1}$$

Let $\Theta_{a,k} = \{b \in \mathbb{R}^p : ||b||_0 = k\}$. Notice that $\Theta_a \subseteq \Theta_0^c = \cup_{k=1}^p \Theta_{a,k}$. A practical challenge for conducting high dimensional testing is the specification of the alternative space $\Theta_a$, or in other words, the direction of possible violation from the null hypothesis.

Hypothesis testing for a high-dimensional mean has received considerable attention in recent literature. Although existing testing procedures are generally designed for a particular type of alternatives, the alternative space is not often clearly specified. In this paper, we shall study (1) with the alternative space $\Theta_a$ stated in a more explicit way (to be more precise, it is stated in terms of the $l_0$ norm). To motivate the subsequent derivations, we consider the following problem

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_{a,k} : \theta \in \Theta_{a,k}. \tag{2}$$

Given the covariance matrix $\Sigma$ or equivalently the precision matrix $\Gamma = (\gamma_{ij})_{i,j=1}^p := \Sigma^{-1}$, we shall develop a testing procedure based on the maximum likelihood principle. Under the Gaussian assumption, the negative log-likelihood function (up to a constant) is given by

$$l_n(\theta) = \frac{1}{2}\sum_{i=1}^n (X_i - \theta)'\Gamma(X_i - \theta).$$

The maximum likelihood estimator (MLE) under $H_{a,k}$ is defined as

$$\hat{\theta} = \arg\min_{\theta \in \Theta_{a,k}} \sum_{i=1}^n (X_i - \theta)'\Gamma(X_i - \theta). \tag{3}$$

To illustrate the idea, we first consider the case $\Gamma = I_p$, i.e., the components of $X_i$ are i.i.d $N(0, 1)$. It is straightforward to see that the $k$ nonzero components of $\hat{\theta}$ are equal to $\bar{x}_{j_s^*} = \sum_{i=1}^n x_{ij_s^*}/n$, where

$$(j_1^*, j_2^*, \dots, j_k^*) = \underset{1 \le j_1 < j_2 < \cdots < j_k \le p}{\arg\max} \sum_{s=1}^k \bar{x}_{j_s}^2.$$

Although the maximum is taken over $\binom{n}{k}$ possible sets, it is easy to see that $j_1^*, \dots, j_k^*$ are just the indices associated with the $k$ largest $|\bar{x}_j|$, i.e., we only need to sort $|\bar{x}_j|$ and pick the indices associated with the $k$ largest values. In this case, the LR test (with $\Gamma = I_p$ known) can be written as maximum of sum-of-squares, i.e.,

$$LR_n(k) = n \max_{S:\mathrm{card}(S) = k} |\bar{X}_S|_2^2.$$

The LR test is seen as a combination of the maximum type test and the sum-of-squares type test and it is designed to optimize the power for testing $H_0$ against $H_{a,k}$ with $k \ge 1$. The two extreme cases are $k = 1$ (the sparsest alternative) and $k = p$ (the densest alternative). In the former case, we have $LR_n(1) = n|\bar{X}|_\infty^2 = n \max_{1 \le j \le p} |\bar{x}_j|^2$, while in the latter case, $LR_n(p) = n|\bar{X}|^2 = n\sum_{j=1}^p |\bar{x}_j|^2$, where $\bar{X} = \sum_{i=1}^n X_i/n = (\bar{x}_1, \dots, \bar{x}_p)'$. Also see Liu and Shao (2004) and Bickel and Chernoff (1993) for LR tests in the case where the covariance is the identity.

We note that an alternative expression for the LR test is given by

$$n\sum_{j=1}^{p}\bar{x}_j^2\mathbf{1}\{|\bar{x}_j|>\delta\}, \tag{4}$$

for some $\delta > 0$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Thus $LR_n(k)$ can also be viewed as a thresholding test [see Donoho and Johnstone (1994); Fan (1996)]. In this paper, we focus on the regime of very sparse (e.g. the number of nonzero entries grows slowly with $n$) but strong signals (e.g. the cumulative effect of the nonzero entries of $\theta$ is greater than $\sqrt{2k\log(p)/n}$) and choose $\delta = |\bar{x}_{j^*_{k+1}}|$ with $|\bar{x}_{j^*_1}| \geq |\bar{x}_{j^*_2}| \geq \cdots |\bar{x}_{j^*_p}|$ in (4) (assuming that $|x_{j^*_k}| > |x_{j^*_{k+1}}|$). For weaker but denser signals, Fan (1996) suggested the use of $\delta = \sqrt{2\log(pa_p)/n}$ for $a_p = c_1(\log p)^{-c_2}$ with $c_1, c_2 > 0$. A more delicate regime is where the signals are so weak that they cannot have a visible effect on the upper extremes, e.g., the strength of signals is $\sqrt{2r\log(p)/n}$ for $r \in (0,1)$. In this case, the signals and noise may be almost indistinguishable. To tackle this challenging problem, the thresholding test with $\delta = \sqrt{2s\log(p)/n}$ for $s \in (0,1)$ was recently considered in Zhong et al. (2013). And a second-level significance test by maximizing over a range of significance levels (the so-called Higher Criticism test) was used to test the existence of any signals [Donoho and Jin (2004)].

It has been shown in the literature that incorporating the componentwise dependence helps to boost the power of the testing procedure [Hall and Jin (2010); Cai et al. (2014); Chen et al. (2014)]. The maximum likelihood principle provides a natural way to utilize such dependence. For general covariance structure, the LR test for testing $H_0$ against $H_{a,k}$ is given by

$$\begin{aligned} LR_n(k) &= \max_{\theta\in\Theta_{a,k}}\sum_{i=1}^{n}\left\{X_i'\Gamma X_i - (X_i-\theta)'\Gamma(X_i-\theta)\right\} \\ &= n\max_{S:\mathrm{card}(S)=k}(\Gamma_{S,S}\bar{X}_S + \Gamma_{S,-S}\bar{X}_{-S})'\Gamma_{S,S}^{-1}(\Gamma_{S,S}\bar{X}_S + \Gamma_{S,-S}\bar{X}_{-S}), \end{aligned}$$

where the second equality follows from Lemma 7.2 in the appendix and $a_S = (a_j)_{j\in S}$ for $a = (a_1,\ldots,a_p)$. Letting $Z = (z_1,\ldots,z_p)' = \Gamma\bar{X}$, a simplified expression is then given by

$$LR_n(k) = n\max_{S:\mathrm{card}(S)=k} Z_S'\Gamma_{S,S}^{-1}Z_S.$$

It is worth pointing out that $LR_n(k)$ is indeed the LR test for testing

$$H_0: \theta\in\Theta_0 \quad \text{against} \quad H_{a,1:k}: \theta\in\cup_{j=1}^{k}\Theta_{a,j},$$

because $\hat{\theta} = \mathrm{argmin}_{\theta\in\cup_{j=1}^{k}\Theta_{a,j}}\sum_{i=1}^{n}(X_i-\theta)'\Gamma(X_i-\theta) = \mathrm{argmin}_{\theta\in\Theta_{a,k}}\sum_{i=1}^{n}(X_i-\theta)'\Gamma(X_i-\theta)$. As an illustration, we consider the following two examples.

EXAMPLE **2.1** (Sparsest case). When $k = 1$, we have

$$LR_n(1) = n \max_{1 \le j \le p} \frac{|z_j|^2}{\gamma_{jj}},$$

which has been recently considered in Cai et al. (2014) for the two-sample problem. Cai et al. (2014) pointed out that "the linear transformation $\Gamma X_i$ magnifies the signals and the number of the signals owing to the dependence in the data". Although a rigorous theoretical justification was provided in Cai et al. (2014), the linear transformation based on $\Gamma$ still seems somewhat mysterious. Here we "rediscover" the test from a different perspective.

EXAMPLE **2.2** (Densest case). To test against the dense alternative $H_{a,p}$, one may consider

$$LR_n(p) = n\bar{X}'\Gamma\bar{X} = n \sum_{i,j=1}^{p} \bar{x}_i \bar{x}_j \gamma_{ij}$$

or its $U$-statistic version $LR_{n,U}(p) = \frac{1}{n-1}\sum_{i,j=1}^{p} \gamma_{ij} \sum_{k \neq l} x_{ki} x_{lj}$. In view of the results in Chen and Qin (2010), the asymptotic behavior of such a test is expected to be very different from $LR_n(k)$ with relatively small $k$. A serious investigation for this test is beyond the scope of the current paper.

We note that the test statistic $LR_n(k)$ involves taking maximization over $\binom{p}{k}$ tuples $(j_1, \ldots, j_k)$ with $1 \le j_1 < \cdots < j_k \le p$, which can be computationally very intensive if $p$ is large. To reduce the computational burden, we consider the following modified test by replacing $\Gamma_{S,S}$ with $\text{diag}(\Gamma_{S,S})$ which contains only the diagonal elements of $\Gamma_{S,S}$. With this substitution, we have

$$T_n(k) = n \max_{S:\text{card}(S)=k} Z_S' \text{diag}^{-1}(\Gamma_{S,S}) Z_S.$$

To compute the modified statistic, one only needs to sort the values $|z_{j_l}|^2/\gamma_{j_l,j_l}$ and find the indices corresponding to the $k$ largest ones, say $j_1^*, j_2^*, \ldots, j_k^*$. Then the test statistic can be computed as

$$T_n(k) = n \sum_{l=1}^{k} \frac{|z_{j_l^*}|^2}{\gamma_{j_l^*,j_l^*}}.$$

Therefore, the computation cost for $T_n(k)$ with $k > 1$ is essentially the same as $LR_n(1)$. Note that $T_n(k)$ is increasing in $k$, the sorting of the values $|z_{j_l}|^2/\gamma_{j_l,j_l}$ only needs to be done once; i.e. $T_n(k+1)$ can be computed by adding just one term to $T_n(k)$.

By the matrix inversion formula $\Gamma_{-j,j} = -\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}\gamma_{jj}$, $z_j/\gamma_{jj} = \bar{x}_j - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\bar{X}_{-j}$. From the above derivation, we note that $n|z_j|^2/\gamma_{jj}$ can be interpreted as the LR test for testing $\theta_j = 0$ given that $\theta_k = 0$ for $k \neq j$. This strategy is conceptually simple and can be conveniently implemented in practice. Also it can be generalized to other parametric models. We also mention that the form of $T_n(k)$ is related with the notion of 2-$k$ symmetric gauge norm, see Bhatia (1997).

REMARK **2.1**. One may employ the so-called graph-assisted procedure [see e.g. Jin et al. (2014); Ke et al. (2014)] to circumvent the expensive computation (of the order $O(p^k)$) required in $LR_n(k)$. Under the Gaussian assumption, $\Gamma$ defines a graph $(V, E)$ in terms of conditional (in)dependence, that is the nodes $i$ and $j$ are connected if and only if $\gamma_{ij} \neq 0$. Let $J(1), \ldots, J(q_0)$ be all the connected components of $(V, E)$ with size less or equal to $k$. Then an alternative test statistic can be defined as,

$$LR_{graph,n}(k) = n \max \sum_{i=1}^{k_0} Z'_{J(j_i)} \Gamma^{-1}_{J(j_i),J(j_i)} Z_{J(j_i)}, \tag{5}$$

where the maximization is over all $\{j_1, \ldots, j_{k_0}\} \subseteq \{1, 2, \ldots, q_0\}$ such that $\sum_{i=1}^{k_0} ||J(j_i)||_0 \leq k$. Under suitable assumptions on $\Gamma$, it was shown in Jin et al. (2014) that the number of all connected components with size less or equal to $k$ is of the order $O(p)$ (up to a multi-log$(p)$ term). A greedy algorithm can be used to list all the sub-graphs. Note that $T_n(k)$ corresponds to the case where $J(j) = \{j\}$ for $1 \leq j \leq p$. Thus $LR_{graph,n}(k)$ could be viewed as a generalized version of $T_n(k)$ with the ability to explore the dependence in $Z$ via the connected components of $(V, E)$.

## 2.2 Feasible test

We have so far focused on the oracle case in which the precision matrix is known. However, in most applications $\Gamma$ is unknown and thus needs to be estimated. Estimating the precision matrix has been extensively studied in the literature in recent years [see e.g. Meinshausen and Bühlmann (2006); Bickel and Levina (2008a; 2008b); Friedman et al. (2008); Yuan (2010); Cai and Liu (2011); Cai et al. (2011); Liu and Wang (2012); Sun and Zhang (2013)].

When $\Gamma$ is known to be banded or bandable, one can employ the banding method based on the cholesky decomposition [Bickel and Levina (2008a)] to estimate $\Gamma$. For sparse precision matrix without knowing the banding structure, the nodewise Lasso and its variants [Meinshausen and Bühlmann (2006); Liu and Wang (2012); Sun and Zhang (2013)] or the constrained $l_1$-minimization for inverse matrix estimation (CLIME) [Cai et al. (2011)] can be used to estimate $\Gamma$.

In this paper, we use the nodewise Lasso regression to estimate the precision matrix $\Gamma$ [Meinshausen and Bühlmann (2006)], but other estimation approaches can also be used as long as the resulting estimator satisfies some desired properties. Let $\widetilde{\mathbf{X}} := (\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n)' \in \mathbb{R}^{n \times p}$ with $\widetilde{X}_i = X_i - \bar{X}$. Let $\widetilde{\mathbf{X}}_j$ be the $j$th column of $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{X}}_{-j}$ be the $n \times (p-1)$ matrix without the $j$th column. For $j = 1, 2, \ldots, p$, consider

$$\widehat{\gamma}_j = \arg \min_{\gamma \in \mathbb{R}^{p-1}} (|\widetilde{\mathbf{X}}_j - \widetilde{\mathbf{X}}_{-j}\gamma|^2/n + 2\lambda_j|\gamma|_1), \tag{6}$$

with $\lambda_j > 0$, where $\widehat{\gamma}_j = \{\widehat{\gamma}_{jk} : 1 \leq k \leq p, \ k \neq j\}$. Let $\widehat{C} = (\widehat{c}_{ij})_{i,j=1}^p$ be a $p \times p$ matrix with $\widehat{c}_{ii} = 1$ and $\widehat{c}_{ij} = -\widehat{\gamma}_{ij}$ for $i \neq j$. Let $\widehat{\tau}_j^2 = |\widetilde{\mathbf{X}}_j - \widetilde{\mathbf{X}}_{-j}\widehat{\gamma}_j|^2/n + \lambda_j|\widehat{\gamma}_j|_1$ and write $\widehat{T}^2 = \text{diag}(\widehat{\tau}_1^2, \ldots, \widehat{\tau}_p^2)$ as a diagonal matrix. The nodewise Lasso estimator for $\Gamma$ is constructed as

$$\widehat{\Gamma} = \widehat{T}^{-2}\widehat{C}.$$

Given a suitable precision matrix estimator $\widehat{\Gamma} = (\widehat{\gamma}_{ij})_{j=1}^{p}$ (e.g. obtained via nodewise Lasso), our feasible test can be defined by replacing $\Gamma$ with its estimator, i.e.,

$$T_{fe,n}(k) = n \max_{S:\text{card}(S)=k} \widehat{Z}_S' \text{diag}^{-1}(\widehat{\Gamma}_{S,S})\widehat{Z}_S,$$

where $\widehat{Z} = \widehat{\Gamma}\bar{X} = (\widehat{z}_1, \dots, \widehat{z}_p)'$. Under suitable assumptions, it has been shown in Cai et al. (2014) that $T_{fe,n}(1)$ converges to an extreme distribution of Type I. To mimic the sampling distribution of $T_{fe,n}(k)$ for $k \geq 1$ under sparsity assumption, we propose a simulation-based approach which is related with the multiplier bootstrap approach in Chernozhukov et al. (2015). The procedure can be described as follows:

1. Estimate $\Gamma$ by $\widehat{\Gamma}$ using a suitable regularization method.
2. Generate $\widehat{Z}^* = (\widehat{z}_1^*, \dots, \widehat{z}_p^*)' = \widehat{\Gamma}\sum_{i=1}^{n}(X_i - \bar{X})e_i/n$, where $e_i \sim^{i.i.d} N(0,1)$ are independent of the sample.
3. Compute the simulation-based statistic as
$$T_{fe,n}^*(k) = n \max_{S:\text{card}(S)=k} \widehat{Z}_S^{*\prime} \text{diag}^{-1}(\widehat{\Gamma}_{S,S})\widehat{Z}_S^*.$$
4. Repeat Steps 2-3 a large number of times to get the $1-\alpha$ quantile of $T_{fe,n}^*(k)$, which serves as the simulation-based critical value.

## 2.3 Choice of $k$ and a modified test

In this subsection, we propose a data dependent method for choosing $k$ which is motivated from the power consideration. Consider the Hotelling's $T^2$ test $T_n^2 := n\bar{X}'\widehat{S}^{-1}\bar{X}$ with $\widehat{S}$ being the sample covariance matrix. Bai and Saranadasa (1996) showed that the asymptotic power function for the Hotelling's $T^2$ test under $p/n \to b \in (0,1)$ has the form

$$\Phi\left(-z_{1-\alpha} + \sqrt{\frac{n(1-b)}{2b}}\theta'\Gamma\theta\right), \tag{7}$$

where $\Phi$ is the distribution function of $N(0,1)$ and $z_{1-\alpha}$ is the $1-\alpha$ quantile of $N(0,1)$. Intuitively, for a set $S$ with $\text{card}(S) = k$ and $k < n$, one may expect that the asymptotic power function of $nZ_S'\Gamma_{S,S}^{-1}Z_S$ is determined by the term,

$$\sqrt{\frac{n-k}{2k}}(\Gamma\theta)_S'\Gamma_{S,S}^{-1}(\Gamma\theta)_S.$$

For known $\Gamma$, we note that $Z_S'\Gamma_{S,S}^{-1}Z_S - \frac{k}{n}$ is an unbiased estimator for $(\Gamma\theta)_S'\Gamma_{S,S}^{-1}(\Gamma\theta)_S$. From the power consideration, a natural test statistic can be defined as

$$\max_{1\leq k\leq M}\max_{\mathrm{card}(S)=k}\sqrt{\frac{n-k}{2k}}\left(Z_S'\Gamma_{S,S}^{-1}Z_S-\frac{k}{n}\right),$$

where $M$ is an upper bound. By replacing $\Gamma_{S,S}^{-1}$ with $\mathrm{diag}^{-1}(\Gamma_{S,S})$, a computationally feasible test is then given by

$$\begin{aligned}\widetilde{T}_n(M) &= \max_{1\leq k\leq M}\max_{\mathrm{card}(S)=k}\sqrt{\frac{1-k/n}{2k}}\left(nZ_S'\mathrm{diag}^{-1}(\Gamma_{S,S})Z_S-k\right)\\ &= \max_{1\leq k\leq M}\sqrt{\frac{1-k/n}{2k}}\left(T_n(k)-k\right).\end{aligned}$$

Substituting $\Gamma$ with $\widehat{\Gamma}$, we therefore propose the following test

$$\widetilde{T}_{fe,n}(M) = \max_{1\leq k\leq M}\sqrt{\frac{1-k/n}{2k}}\left(T_{fe,n}(k)-k\right).$$

To approximate its sampling distribution, we suggest the following modified simulation-based statistic in Step 3 above,

$$\widetilde{T}^*_{fe,n}(M) = \max_{1\leq k\leq M}\sqrt{\frac{1-k/n}{2k}}\left(n\max_{S:\mathrm{card}(S)=k}\widehat{Z}_S^{*\prime}\mathrm{diag}^{-1}(\widehat{\Gamma}_{S,S})\widehat{Z}_S^*-k\right),$$

where $\widehat{Z}^* = (\widehat{z}_1^*,\dots,\widehat{z}_p^*)' = \widehat{\Gamma}\sum_{i=1}^n(X_i-\bar{X})e_i/n$ with $e_i \sim^{i.i.d} N(0,1)$ that are independent of the sample.

## 2.4 Theoretical results

In this subsection, we study the theoretical properties of the proposed test and justify the validity of the simulation-based approach. To facilitate the derivations, we make the following assumptions. Denote by $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ the smallest and the largest eigenvalues of $\Sigma$ respectively. Let

$$d = \max_{1\leq j\leq p}||\{\gamma_{jk}: k\neq j, 1\leq k\leq p\}||_0.$$

Assumptions 2.1-2.2 below are common in the analysis of nodewise Lasso, see e.g. van de Geer et al. (2014).

Assumption **2.1**. *Suppose* $\max_{1\leq j\leq p}\sigma_{j,j} < c_1$ *and* $c_2 < \lambda_{\min}(\Sigma)$ *for some* $c_1, c_2 > 0$.

Assumption **2.2**. *Suppose* $d^2\log(p)/n = o(1)$.

Proposition 7.1 in the appendix states some consistency properties for $\widehat{\Gamma}$ under 2.1-2.2. We remark that any precision matrix estimators that satisfy (17)-(19) in the appendix can be used in the proposed tests. In the simulation studies, we recommend the use of nodewise square root Lasso which is tuning insensitive.

We are now in position to present the main results in this section. Define the quantity $\phi(\Gamma; k) = \min_{|v|=1,||v||_0 \leq k} v'\Gamma v$. Let $X_1^n = \{X_1, \ldots, X_n\}$.

THEOREM **2.1**. *Assume that* $k^2 d(\log(np))^{5/2}/\sqrt{n} = o(1)$ *and* $\phi(\Gamma; k) > c$ *for some positive constant* $c$*. Under Assumptions 2.1-2.2 and* $H_0$*, we have*

$$\sup_{t \geq 0} \left| P\left(T^*_{fe,n}(k) \leq t \middle| X_1^n\right) - P\left(T_{fe,n}(k) \leq t\right) \right| = o_p(1).$$

Theorem 2.1 justifies the use of the simulation-based approach in Section 2.2. The next theorem further establishes its validity for the modified test.

THEOREM **2.2**. *Assume that* $M^4 d(\log(np))^{5/2}/\sqrt{n} = o(1)$ *and* $\phi(\Gamma; M) > c$ *for some positive constant* $c$*. Under Assumptions 2.1-2.2 with* $k$ *replaced by* $M$*, and* $H_0$*, we have*

$$\sup_{t_M \geq t_{M-1} \geq \cdots \geq t_1 \geq 0} \left| P\left(\bigcap_{j=1}^{M} \{T^*_{fe,n}(j) \leq t_j\} \middle| X_1^n\right) - P\left(\bigcap_{j=1}^{M} \{T_{fe,n}(j) \leq t_j\}\right) \right| = o_p(1).$$

*As a consequence, we have*

$$\sup_{t \geq 0} \left| P\left(\widetilde{T}^*_{fe,n}(M) \leq t \middle| X_1^n\right) - P\left(\widetilde{T}_{fe,n}(M) \leq t\right) \right| = o_p(1).$$

The restrictions on $k$ and $M$ in Theorems 2.1-2.2 are essentially due to the use of the Gaussian approximation theory [see e.g. Chernozhukov et al. (2015)]. Nevertheless, our numerical results suggest that the simulation-based approach performs reasonably well even when $M$ takes some larger values such as $n/4$ or $n/2$. Below we study the power property of the proposed testing procedure. To proceed, we impose the following conditions.

ASSUMPTION **2.3**. *Assume that* $\max_{1 \leq j_1 < j_2 < \cdots < j_k \leq p} \sum_{l=1}^{k} \gamma_{j_l,j_l} \theta_{j_l}^2 \geq (2k+\epsilon)\log(p)/n$ *for some* $\epsilon > 0$.

ASSUMPTION **2.4**. *Suppose* $||\theta||_0 = \lceil p^r \rceil$ *for some* $0 \leq r < 1/4$*, and the non-zero locations are randomly uniformly drawn from* $\{1, 2, \ldots, p\}$*. And the scheme is independent of* $\{X_i - \theta\}_{i=1}^n$.

ASSUMPTION **2.5**. *Let* $diag^{-1/2}(\Gamma)\Gamma diag^{-1/2}(\Gamma) = (\nu_{ij})_{i,j=1}^p$*. Assume* $\max_{1 \leq i < j \leq p} |\nu_{ij}| \leq c_0 < 1$ *for some constant* $0 < c_0 < 1$*. Further assume that* $\lambda_{\max}(\Sigma) \leq C_0$ *for some constant* $C_0 > 0$.

Define $c^*_\alpha(k) = \inf\{t > 0 : P(T^*_{fe,n}(k) \leq t | X_1^n) \geq 1-\alpha\}$ and $\widetilde{c}^*_\alpha(M) = \inf\{t > 0 : P(\widetilde{T}^*_{fe,n}(M) \leq t | X_1^n) \geq 1-\alpha\}$ the simulation-based critical values. The consistency of the testing procedure is established in the following theorem.

THEOREM **2.3**. *Suppose* $k^2 d(\log(np))^{5/2}/\sqrt{n} = o(1)$*. Under Assumptions 2.1-2.5, we have*

$$P(T_{fe,n}(k) > c^*_\alpha(k)) \to 1. \tag{8}$$

*Moreover, suppose Assumption 2.3 holds with $k = M$. Then for $M$ satisfying that $M^4 d(\log(np))^{5/2}/\sqrt{n} = o(1)$,*

$$P(\widetilde{T}_{fe,n}(M) > \widetilde{c}^*_\alpha(M)) \to 1.$$

Remark **2.2**. Assumption 2.3 implies that

$$\max_{1\leq j\leq p} |\theta_j|/\sqrt{\sigma_{jj}} \geq \sqrt{2\{1/(\sigma_{j_0,j_0}\gamma_{j_0,j_0}) + \epsilon_0/k\}\log(p)/n}, \tag{9}$$

where $\epsilon_0 > 0$ and $j_0 = \arg\max_{1\leq j\leq p}\sqrt{\gamma_{jj}}|\theta_j|$. In this case, the maximum type test in Cai et al. (2014) is expected to be consistent as well. However, when the true sparsity level is greater than one, the asymptotic power of the maximum type test can be lower than that of $T_n(k)$ with $k > 1$, see Section 4.1.

Remark **2.3**. Define $\mathcal{A}(r;k;c) = \{\theta \in \mathbb{R}^p : ||\theta||_0 = \lceil p^r \rceil, \max_{1\leq j_1<\cdots<j_k\leq p}\sum_{l=1}^k |\theta_{j_l}|^2 \geq ck\log(p)/n\}$, and the class of $\alpha$-level tests $\mathcal{D}_\alpha = \{\Phi_a \in \{0,1\} : P_{\theta=0_{p\times 1}}(\Phi_a = 1) \leq \alpha\}$. The arguments in the proof of Theorem 3 of Cai et al. (2014) indeed imply that for small enough $c$, and $r < 1/4$,

$$\inf_{\Phi_a\in\mathcal{D}_\alpha}\ \sup_{\theta\in\mathcal{A}(r;k;c)} P_\theta(\Phi_a = 0) \geq c' > 0.$$

Therefore, the separation rate $k\log(p)/n$ in Assumption 2.3 is minimax optimal in the above sense.

Finally, we point out that the Gaussian assumption can be relaxed by employing the recently developed Central Limit Theorem in high dimension [Chernozhukov et al. (2015)]. For a random variable $X$, we define the sub-Gaussian norm [see Definition 5.7 of Vershynin (2012)] as

$$||X||_\psi = \sup_{q\geq 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q}.$$

Assumption **2.6**. *Assume that $\sup_{v\in\mathbb{S}^{p-1}}||v'X_i||_\psi < c_3$ and $\sup_{v\in\mathbb{S}^{p-1}}||v'\Gamma X_i||_\psi < c_4$ for some constants $c_3, c_4 > 0$, where $\mathbb{S}^{p-1} = \{b \in \mathbb{R}^p : |b| = 1\}$.*

Let $W = (w_1, w_2, \ldots, w_p) \sim N_p(0, \text{diag}^{-1/2}(\Gamma)\Gamma\text{diag}^{-1/2}(\Gamma))$ and define $T^W(k) = \max_{1\leq j_1<j_2<\cdots j_k\leq p}\sum_{l=1}^k w_{j_l}^2$. We especially have the following result, which indicates that under the sub-Gaussian assumption, the distribution of $T_{fe,n}(k)$ can be approximated by its Gaussian counterpart $T^W(k)$.

Proposition **2.1**. *Assume that $d(k\log(np))^{7/2}/\sqrt{n} = o(1)$ and $\phi(\Gamma;k) > c$ for some positive constant $c > 0$. Under Assumptions 2.1, 2.2, 2.6 and $H_0$, we have*

$$\sup_{t\geq 0}\left|P\left(T_{fe,n}(k) \leq t\right) - P\left(T^W(k) \leq t\right)\right| = o(1).$$

Using Proposition 2.1, we can establish the consistency of $T_{fe,n}(k)$ under the sub-Gaussian assumption. The details are omitted here to conserve space.

# 3 Extension to the two-sample problem

## 3.1 Likelihood ratio test

The maximum likelihood viewpoint allows a direct extension of the above procedure to the two-sample problem. Consider two samples $\{X_i\}_{i=1}^{n_1} \sim^{i.i.d} N_p(\theta_1, \Sigma_1)$ and $\{Y_i\}_{i=1}^{n_2} \sim^{i.i.d} N_p(\theta_2, \Sigma_2)$, where the two samples are independent of each other. A canonical problem in multivariate analysis is the hypothesis testing of

$$H_0' : \theta_1 - \theta_2 \in \Theta_0 \quad \text{versus} \quad H_a' : \theta_1 - \theta_2 \in \Theta_a \subseteq \Theta_0^c.$$

Given the priori $\theta_1 - \theta_2 \in \Theta_{a,k}$, we consider

$$H_0' : \Delta \in \Theta_0 \quad \text{versus} \quad H_{a,k}' : \Delta \in \Theta_{a,k},$$

where $\Delta = \theta_1 - \theta_2$.

Notation-wise, let $\Gamma_j = \Sigma_j^{-1}$ for $j = 1, 2$. Define $C_1 = (n_1\Gamma_1 + n_2\Gamma_2)^{-1} n_1\Gamma_1$ and $C_2 = (n_1\Gamma_1 + n_2\Gamma_2)^{-1} n_2\Gamma_2$. Further let $\Psi = n_1\Omega^{21} + n_2\Omega^{12}$, where $\Omega^{21} = C_2'\Gamma_1 C_2$ and $\Omega^{12} = C_1'\Gamma_2 C_1$. The following proposition naturally extends the result in Section 2.1 to the two-sample case.

PROPOSITION **3.1**. *The LR test for testing $H_0'$ against $H_{a,k}'$ is given by*

$$LR_n(k) = \max_{S:card(S)=k} (\Psi\bar{X} - \Psi\bar{Y})_S' (\Psi_{S,S})^{-1} (\Psi\bar{X} - \Psi\bar{Y})_S,$$

*where $\bar{X} = \sum_{i=1}^{n_1} X_i/n_1$ and $\bar{Y} = \sum_{i=1}^{n_2} Y_i/n_2$.*

## 3.2 Equal covariance structure

We first consider the case of equal covariance, i.e., $\Gamma := \Gamma_1 = \Gamma_2$. Simple calculation yields that $\Psi = n_1\Omega^{21} + n_2\Omega^{12} = n_1 n_2 \Gamma/(n_1 + n_2)$. Thus the LR test can be simplified as,

$$LR_n(k) = \max_{S:\text{card}(S)=k} \frac{n_1 n_2}{n_1 + n_2} (\Gamma\bar{X} - \Gamma\bar{Y})_S' (\Gamma_{S,S})^{-1} (\Gamma\bar{X} - \Gamma\bar{Y})_S. \tag{10}$$

We note that $LR_n(1)$ reduces to the two-sample test proposed in Cai et al. (2014). By replacing $\Gamma_{S,S}$ with $\text{diag}(\Gamma_{S,S})$ in (10), we obtain the (infeasible) statistic

$$T_n(k) = \max_{S:\text{card}(S)=k} \frac{n_1 n_2}{n_1 + n_2} (\Gamma\bar{X} - \Gamma\bar{Y})_S' \text{diag}^{-1}(\Gamma_{S,S}) (\Gamma\bar{X} - \Gamma\bar{Y})_S,$$

which is computationally efficient.

Let $\widehat{\Gamma} = (\widehat{\gamma}_{i,j})_{i,j=1}^p$ be a suitable estimator for $\Gamma$ based on the pooled sample. The feasible test is given by

$$T_{fe,n}(k) = \max_{S:\text{card}(S)=k} \frac{n_1 n_2}{n_1 + n_2} (\widehat{\Gamma}\bar{X} - \widehat{\Gamma}\bar{Y})_S' \text{diag}^{-1}(\widehat{\Gamma}_{S,S}) (\widehat{\Gamma}\bar{X} - \widehat{\Gamma}\bar{Y})_S.$$

To approximate the sampling distribution of the above test, one can employ the simulation-based approach described below:

1. Estimate $\widehat{\Gamma}$ using suitable regularization method based on the pooled sample.

2. Let $X^* = \sum_{i=1}^{n_1}(X_i - \bar{X})e_i/n_1$ and $Y^* = \sum_{i=1}^{n_2}(Y_i - \bar{Y})\tilde{e}_i/n_1$, where $\{e_i\}$ and $\{\tilde{e}_i\}$ are two independent sequences of i.i.d $N(0,1)$ random variables that are independent of the sample.

3. Compute the simulation-based statistic $T^*_{fe,n}(k)$ by replacing $\bar{X}$ and $\bar{Y}$ with $X^*$ and $Y^*$.

4. Repeat Steps 2-3 a large number of times to get the $1-\alpha$ quantile of $T^*_{fe,n}(k)$, which serves as the simulation-based critical value.

Next, we briefly discuss the choice of $k$. By Theorem 2.1 in Bai and Saranadasa (1996), we know that the asymptotic power function for the two-sample Hotelling's $T^2$ test is given by

$$\Phi\left(-z_{1-\alpha} + \sqrt{\frac{N(1-b)}{2b}}\frac{n_1 n_2}{N^2}(\theta_1 - \theta_2)'\Gamma(\theta_1 - \theta_2)\right),$$

under $p/N \to b \in (0,1)$, where $N = n_1 + n_2 - 2$. Thus for $k < N$, the asymptotic power of $T_n(k)$ is related to

$$\sqrt{\frac{N-k}{2k}} \max_{\text{card}(S)=k} \{\Gamma(\theta_1 - \theta_2)\}'_S \text{diag}^{-1}(\Gamma_{S,S})\{\Gamma(\theta_1 - \theta_2)\}_S.$$

Notice that

$$(\Gamma\bar{X} - \Gamma\bar{Y})'_S \text{diag}^{-1}(\Gamma_{S,S})(\Gamma\bar{X} - \Gamma\bar{Y})_S - \frac{k(n_1+n_2)}{n_1 n_2}$$

is an unbiased estimator for $\{\Gamma(\theta_1 - \theta_2)\}'_S \text{diag}^{-1}(\Gamma_{S,S})\{\Gamma(\theta_1 - \theta_2)\}_S$. Thus we propose to choose $k$ by

$$\widehat{k} = \arg\max_{1 \le k \le M'} \sqrt{\frac{N-k}{2k}}\left(T_{fe,n}(k) - k\right),$$

where $M'$ is a pre-specified upper bound for $k$. Following the same spirit in Section 2.3, a modified test statistic is given by

$$\widetilde{T}_{fe,n}(M') = \max_{1 \le k \le M'} \sqrt{\frac{1-k/N}{2k}}\left(T_{fe,n}(k) - k\right),$$

and the simulation-based procedure can be used to approximate its sampling distribution.

We can justify the validity of the testing procedure under both the null and alternative hypotheses. The arguments are similar to those in the one sample case, see Sections 2.4 and 7.2.

## 3.3 Unequal covariance structures

In the case of unequal covariance structures i.e., $\Gamma_1 \neq \Gamma_2$, we cannot use the pooled sample to estimate the covariance structures. Let $\widehat{\Gamma}_i$ with $i = 1, 2$ be suitable precision matrix estimators based on each sample separately. Denote by $\widehat{C}_i$ the estimator for $C_i$ with $i = 1, 2$. A particular choice here is given by

$$\widehat{C}_1 = (n_1\widehat{\Gamma}_1 + n_2\widehat{\Gamma}_2)^{-1} n_1\widehat{\Gamma}_1, \quad \widehat{C}_2 = (n_1\widehat{\Gamma}_1 + n_2\widehat{\Gamma}_2)^{-1} n_2\widehat{\Gamma}_2.$$

Further define $\widehat{\Omega}^{21} = \widehat{C}_2'\widehat{\Gamma}_1\widehat{C}_2$ and $\widehat{\Omega}^{12} = \widehat{C}_1'\widehat{\Gamma}_2\widehat{C}_1$. Let $\widehat{\Psi} = (\widehat{\psi}_{ij})_{i,j=1}^p = n_1\widehat{\Omega}^{21} + n_2\widehat{\Omega}^{12}$, and $\widehat{G} = (\widehat{g}_1, \ldots, \widehat{g}_p)' = \widehat{\Psi}(\bar{X} - \bar{Y})$. By replacing $\widehat{\Psi}$ with $\text{diag}(\widehat{\Psi})$, we suggest the following computationally feasible test,

$$T_{fe,n}(k) = \max_{S:\text{card}(S)=k} (\widehat{\Psi}\bar{X} - \widehat{\Psi}\bar{Y})_S' \text{diag}^{-1}(\widehat{\Psi}_{S,S})(\widehat{\Psi}\bar{X} - \widehat{\Psi}\bar{Y})_S. \tag{11}$$

When $k = 1$, we have

$$T_{fe,n}(1) = \max_{1\leq j\leq p} \frac{|\widehat{g}_j|^2}{\widehat{\psi}_{jj}}, \tag{12}$$

which can be viewed as an extension of Cal et al. (2014)'s test statistic to the case of unequal covariances. Again one can employ the simulation-based approach to obtain the critical values for $T_{fe,n}(k)$. In this case, a modified test can be defined in a similar manner as

$$\widetilde{T}_{fe,n}(M'') = \max_{1\leq k\leq M''} \sqrt{\frac{1 - k/N}{2k}} (T_{fe,n}(k) - k)$$

for some upper bound $M''$.

# 4 Numerical studies

## 4.1 Power analysis

To better understand the power performance of $T_n(k)$ (for known $\Gamma$), we present below a small numerical study. Let $W = (w_1, w_2, \ldots, w_p) \sim N_p(0, \text{diag}^{-1/2}(\Gamma)\Gamma\text{diag}^{-1/2}(\Gamma))$ and recall that $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_p)'$ with $\widetilde{\theta}_j = (\Gamma\theta)_j/\sqrt{\gamma_{jj}}$ from the introduction. Define $T^W(k; \widetilde{\theta}) = \max_{1\leq j_1<j_2<\cdots j_k\leq p} \sum_{l=1}^k (w_{j_l} + \sqrt{n}\widetilde{\theta}_{j_l})^2$. It is obvious that $T_n(k) =^d T^W(k; \widetilde{\theta})$, where "$=^d$" means equal in distribution. Denote by $C_k(\alpha)$ the $100(1-\alpha)$th quantile of $T^W(k; 0)$, which can be obtained via simulation (in our study, $C_k(\alpha)$ is estimated via 100000 simulation runs). Define the power function

$$\mathcal{P}(k, \alpha, \widetilde{\theta}, \Gamma) = P(T^W(k; \widetilde{\theta}) > C_k(\alpha)).$$

We focus on the AR(1) covariance structure $\Sigma = (\sigma_{i,j})_{i,j=1}^p$ with $\sigma_{i,j} = 0.6^{|i-j|}$ and $\Gamma = \Sigma^{-1}$. The mean vector $\theta$ is assumed to contain $k_0$ nonzero components with the same magnitude $\sqrt{2r\log(p)/n}$, where the locations of the nonzero components are drawn without replacement from $\{1, 2, \ldots, p\}$. Figure 2 presents the power function $\mathcal{P}(k, 0.05, \widetilde{\theta}, \Gamma)$ as a curve of $r$ which determines the signal strength, where $k_0 = 1, 5, 10, 20$, and $p = 200, 1000$. These results are consistent with our statistical intuition. In particular, for $k_0 > 1$, the power of $T_n(k)$ with $k > 1$ dominates the power of $T_n(1)$. In this case, it is worth noting that even when $k$ is greater than the true sparsity level, the power of $T_n(k)$ does not seem to decrease much. This is potentially due to the presence of additional signals after the transformation based on $\Gamma$, see Figure 1. Therefore from the power consideration, when $\theta$ contains more than one nonzero component, it seems beneficial to use $T_n(k)$ with $k > 1$. This is further confirmed by the simulation results below.

## 4.2 Empirical size and power

In this subsection, we report the numerical results for comparing the proposed testing procedure with some existing alternatives. Specially we focus on the two-sample problem for testing $H_0' : \Delta \in \Theta_0$ against the alternatives $H_{a,k}' : \Delta \in \Theta_{a,k}$. Without loss of generality, we set $\theta_2 = 0$. Note that under $H_{a,k}'$, $\theta_1$ has $k$ non-zero elements. Denote by $\lfloor x \rfloor$ the largest integer not greater than $x$. We consider the settings below.

(1) Case 1: $k = \lfloor 0.05p \rfloor$ and the non-zero entries are equal to $\varphi_j \sqrt{\log(p)/n}$, where $\varphi_j$ are i.i.d random variables with $P(\varphi_j = \pm 1) = 1/2$.

(2) Case 2: $k = \lfloor \sqrt{p} \rfloor$ and the strength of the signals is the same as (1).

(3) Case 3: $k = \lfloor p^{0.3} \rfloor$ and the nonzero entries are all equal to $\sqrt{4r \log p/n}$ with $r = 0.1, 0.2, 0.3, 0.4,$ and $0.5$.

Here the locations of the nonzero entries are drawn without replacement from $\{1, 2, \ldots, p\}$. Following Cai et al. (2014), the following four covariance structures are considered.

(a) (block diagonal $\Sigma$): $\Sigma = (\sigma_{j,k})$ where $\sigma_{j,j} = 1$ and $\sigma_{j,k} = 0.8$ for $2(r-1)+1 \le j \ne k \le 2r$, where $r = 1, \ldots, \lfloor p/2 \rfloor$ and $\sigma_{j,k} = 0$ otherwise.

(b) ('bandable' $\Sigma$): $\Sigma = (\sigma_{j,k})$ where $\sigma_{j,k} = 0.6^{|j-k|}$ for $1 \le j, k \le p$.

(c) (banded $\Gamma$): $\Gamma = (\gamma_{j,k})$ where $\gamma_{j,j} = 2$ for $j = 1, \ldots, p$, $\gamma_{j,(j+1)} = 0.8$ for $j = 1, \ldots, p-1$, $\gamma_{j,(j+2)} = 0.4$ for $j = 1, \ldots, p-2$, $\gamma_{j,(j+3)} = 0.4$ for $j = 1, \ldots, p-3$, $\gamma_{j,(j+4)} = 0.2$ for $j = 1, \ldots, p-4$, $\gamma_{j,k} = \gamma_{k,j}$ for $j, k = 1, \ldots, p$, and $\gamma_{j,k} = 0$ otherwise.

(d) (block diagonal $\Gamma$): Denote by $D$ a diagonal matrix with diagonal elements generated independently from the uniform distribution on $(1, 3)$. Let $\Sigma_0$ be generated according to (a). Define $\Gamma = D^{1/2}\Sigma_0^2 D^{1/2}$ and $\Sigma = \Gamma^{-1}$.

For each covariance structure, two independent random samples are generated with the same sample size $n_1 = n_2 = 80$ from the following multivariate models,

$$X = \theta_1 + \Sigma^{1/2} U_1, \quad Y = \theta_2 + \Sigma^{1/2} U_2, \tag{13}$$

where $U_1$ and $U_2$ are two independent $p$-dimensional random vectors with independent components such that $\mathbb{E}(U_j) = 0$ and $\mathrm{var}(U_j) = I_p$ for $j = 1, 2$. We consider two cases: $U_j \sim N(0, I_p)$, and the component of $U_j$ is standardized Gamma(4,1) random variable such that it has zero mean and unit variance. The dimension $p$ is equal to $50, 100$ or $200$. Throughout the simulations, the empirical sizes and powers are calculated based on 1000 Monte Carlo replications.

To estimate the precision matrix, we use the nodewise square root Lasso [Belloni et al. (2012)] proposed in Liu and Wang (2012), which is essentially equivalent to the scaled-Lasso from Sun and Zhang (2013). To select the tuning parameter $\lambda$ in the nodewise square root Lasso, we consider the following criteria,

$$\lambda^* = \text{argmin}_{\lambda \in \Lambda_n} ||\widehat{\Gamma}(\lambda)\widehat{\Sigma}\widehat{\Gamma}(\lambda)' - \widehat{\Gamma}(\lambda)||_\infty, \tag{14}$$

where $\widehat{\Sigma}$ is the pooled sample covariance matrix and the minimization is taken over a prespecified finite set $\Lambda_n$. Moreover, we employ the data dependent method in Section 3.2 to select $k$ with the upper bound $M' = 40$ (we also tried $M' = 20, 80$ and found that the results are basically the same as those with $M' = 40$). For the purpose of comparison, we also implemented the Hotelling's $T^2$ test and the two-sample tests proposed in Bai and Saranadasa (1996), Chen and Qin (2010), and Cai et al. (2014). As the results under the Gamma model are qualitatively similar to those under the Gaussian model, we only present the results from the Gaussian model. Table 2 summarizes the sizes and powers in cases 1 and 2. The empirical powers in case 3 with $r$ ranging from 0.1 to 0.5 are presented in Figure 3. Some remarks are in order regarding the simulation results: (i) the empirical sizes are reasonably close to the nominal level 5% for all the tests; (ii) the proposed tests and the maximum type test in Cai et al. (2014) significantly outperform the sum-of-squares type testing procedures in terms of power under Models (a), (b) and (d); Under Model (c), the proposed method is quite competitive to Chen and Qin (2010)'s test which delivers more power than Cai et al. (2014)'s test in some cases; (iii) $T_{fe,n}(k)$ is consistently more powerful than Cai et al. (2014)'s test in almost all the cases; (iv) the modified test $\widetilde{T}_{fe,n}(M')$ is insensitive to the upper bound $M'$ (as shown in our unreported results). And its power is very competitive to $T_{fe,n}(k)$ with a suitably chosen $k$.

## 4.3 Power comparison under different signal allocations

We conduct additional simulations to compare the power of the proposed method with alternative approaches under different signal allocations. The data are generated from (13) with Gaussian distribution and bandable covariance structure (b). Let $k = \lfloor 0.1p \rfloor$ and consider the following four patterns of allocation, where the locations of the nonzero entries are drawn without replacement from $\{1, 2, \ldots, p\}$.

i (Square root): the nonzero entries are equal to $\sqrt{4r\log(p)/n}\sqrt{j/k}$ for $1 \le j \le k$.

ii (Linear): the nonzero entries are equal to $\sqrt{4r\log(p)/n}(j/k)$ for $1 \le j \le k$.

iii (Rational): the nonzero entries are equal to $\sqrt{4r\log(p)/n}(1/j)$ for $1 \le j \le k$.

iv (Random): the nonzero entries are drawn uniformly from $(-\sqrt{4r\log(p)/n}, \sqrt{4r\log(p)/n})$.

Figure 4 reports the empirical rejection probabilities for $p = 100, 200$, and $r$ ranging from 0.1 to 0.5. We observe that the slower the strength of the signals decays, the higher power the tests can generate. The proposed method generally outperforms the two-sample tests in Chen and Qin (2010) and Cai et al. (2014) especially when the magnitudes of signals decay slowly. This result makes intuitive sense as when the magnitudes of signals are close, the top few signals together provide a stronger indication for the violation from the null as compared to the indication using only the largest signal. To sum up, the numerical results demonstrate the advantages of the proposed method over some competitors in the literature.

# 5 Data illustration

The proposed method was applied to a microarray expression data set of patients suffering from both T-cell and B-cell acute lymphoblastic leukemia (ALL) from the study in Chiaretti et al. (2004). The ALL data set is publicly available in the R package `ALL` which can be downloaded from https://bioconductor.org. Following Chen and Qin (2010), our analysis focuses on two sub-classes within the B-cell type leukemia representing two molecular classes: the BCR/ABL class ($n_1 = 37$) and NEG class ($n_2 = 42$).

It is known that to achieve certain biological functions, genes tend to work in groups which are known as gene-sets. Analysis based on gene-sets can derive more power than focusing on individual gene in extracting biological insights. The gene-sets are technically defined in gene ontology (GO) system which provides structured vocabularies producing the names of the sets or known as GO terms. Three categories of gene ontologies of interest have been identified: biological processes (BP), cellular components (CC) and molecular functions (MF). A particular gene-set will belong to one of these three groups of GO terms.

Suppose $\mathcal{S}_1, \ldots, \mathcal{S}_G$ are $G$ gene-sets. For the $i$th gene set, let $\theta_{i,1}$ and $\theta_{i,2}$ be the mean vectors of the expression levels of the BCR/ABL class and NEG class respectively. Consider the hypothesis,

$$H_0^{(i)} : \theta_{i,1} = \theta_{i,2} \quad \text{versus} \quad H_a^{(i)} : \theta_{i,1} \neq \theta_{i,2},$$

for $i = 1, 2, \ldots, G$. Because a gene can belong to several functional groups, the gene-sets $\{\mathcal{S}_i\}_{i=1}^G$ can overlap. Using the method in Gentleman et al. (2005), we carried out preliminary screening for gene-filtering, which left 2391 genes for our analysis. Each gene was mapped to gene-sets using the R object `hgu95av2GO2PROBE` in the package `hgu95av2`, which does not associate the child terms from the GO ontology with the gene. To focus on the high-dimensional scenario, only gene-sets with dimension greater than 10 were retained. There are 510 unique GO terms in BP, 142 in MF and 153 in CC for the ALL data set. The largest gene-sets contain 314, 1591, and 946 genes in the BP, MF and CC categories respectively. We are interested in testing the differences in the expression levels of gene-sets between the two sub-classes (BCR/ABL and NEG) for the three categories (BP, MP and CC).

For each category, we applied the proposed two-sample test as well as the two-sample tests in Chen and Qin (2010) and Cai et al. (2014) to the corresponding GO terms with Bonferroni correction to control the family-wise error rate (FWER) at the 5% level. We use the method in Cai et al. (2013) to check the equal covariance assumption. When the assumption was rejected, we employed the test developed in Section 3.3 to perform the analysis. To make a fair comparison, we use the square root Lasso to estimate precision matrices for both the proposed test and Cai et al. (2014)'s test. The numbers of discoveries and the histograms of p-values are presented in Table 1 and Figure 5 respectively. The proposed test is seen to identify most differentially expressed gene-sets, while Chen and Qin (2010)'s test discovers more gene-sets that are not considered as significantly different by the other two tests. This is not surprising because Chen and Qin (2010)'s test is more suitable for detecting weak and dense signals, while the proposed test and Cai et al. (2014)'s test are designed for sparse and strong signals. Our finding suggests that the signals (i.e., the difference in the expression levels between the BCR/ABL and NEG classes) can have quite different shapes across different gene-sets. The test that is most suitable for detecting the signals among the gene-sets may vary from case to case.

Table 1: Number of rejections after controlling the FWER at the 5% level produced by the three tests. The number in the round bracket denotes the number of findings not shared by the other two tests.

| Class | # of gene sets | CQ | CLX | $\widetilde{T}_{fe,n}(N_0)$ |
|---|---|---|---|---|
| BP | 510 | 138 (48) | 148(21) | 162 (31) |
| MF | 142 | 43 (12) | 49 (8) | 53 (9) |
| CC | 153 | 44 (16) | 56 (6) | 63 (10) |

Note: $N_0 = \min\{p, \lfloor (n_1 + n_2 - 2)/2 \rfloor\}$, where $n_1$ and $n_2$ denote the sample sizes for each class respectively.

# 6 Concluding remark

In this paper, we developed a new class of tests named maximum sum-of-squares tests for conducting inference on high dimensional mean under sparsity assumption. The maximum type test has been shown to be optimal under very strong sparsity [Arias-Castro et al. (2011)]. Our result suggests that even for very sparse signals (e.g. $k$ grows slowly with $n$), the maximum type test may be improved to some extent. It is worth mentioning that our method can be extended to more general settings. For example, consider a parametric model with the negative log-likelihood (or more generally loss function) $\mathcal{L}(Y, X'\beta)$, where $\beta \in \mathbb{R}^p$ is the parameter of interest, $X$ is the $p$-dimensional covariate and $Y$ is the response variable. We are interested in testing $H_0 : \beta = 0_{p\times 1}$ versus $H_{a,k} : \beta \in \Theta_{a,k}$. Given $n$ observations $\{Y_i, X_i\}_{i=1}^n$, the LR test for testing $H_0$ against $H_{a,k}$ is then defined as $LR_n(\beta) = 2\sum_{i=1}^n \mathcal{L}(Y_i, 0) - 2\min_{\beta\in\Theta_{a,k}} \sum_{i=1}^n \mathcal{L}(Y_i, X_i'\beta)$. In the case of linear model, it is related with the maximum spurious correlations recently considered in Fan et al. (2015)

under the null. It is of interest to study the asymptotic properties of $LR_n(\beta)$ in this more general context.

# 7 Technical appendix

## 7.1 Preliminaries

We provide proofs of the main results in the paper. Throughout the appendix, let $C$ be a generic constant which is different from line to line. Define the unit sphere $\mathbb{S}^{p-1} = \{b \in \mathbb{R}^p : |b| = 1\}$.

For any $1 \le k \le p$, define

$$\mathcal{A}(t;k) = \bigcap_{j=1}^{\binom{p}{k}} \mathcal{A}_j(t), \quad \mathcal{A}_j(t) = \{w \in \mathbb{R}^p : w'_{S_j} w_{S_j} \le t\}.$$

Here $S_j$ is the $j$th subset of $[p] := \{1, 2, \ldots, p\}$ with cardinality $k$ for $1 \le j \le \binom{p}{k}$. It is straightforward to verify that $\mathcal{A}_j(t)$ is convex and it only depends on $w_{S_j}$, i.e. the components in $S_j$. The dual representation [see Rockafellar (1970)] for the convex set $\mathcal{A}_j(t)$ with $1 \le j \le \binom{p}{k}$ is given by

$$\mathcal{A}_j(t) = \bigcap_{v \in \mathbb{S}^{p-1}, v_{S_j} \in \mathbb{S}^{k-1}} \{w \in \mathbb{R}^p : w'v \le \sqrt{t}\},$$

where we have used the fact that $\sup_{v \in \mathbb{S}^{p-1}, v_{S_j} \in \mathbb{S}^{k-1}} w'v = |w_{S_j}|$ by the Cauchy-Schwartz inequality. Define $\mathcal{F} = \{v \in \mathbb{S}^{p-1}, ||v||_0 \le k\}$. It is not hard to see that

$$\mathcal{A}(t;k) = \bigcap_{v \in \mathcal{F}} \{w \in \mathbb{R}^p : w'v \le \sqrt{t}\}.$$

Let $\mathcal{X}$ be a subset of a Euclidean space and let $\epsilon > 0$. A subset $N_\epsilon$ of $\mathcal{X}$ is called an $\epsilon$-net of $\mathcal{X}$ if every point $x \in \mathcal{X}$ can be approximated to within $\epsilon$ by some point $y \in N_\epsilon$, i.e. $|x - y| \le \epsilon$. The minimal cardinality of an $\epsilon$-net of $\mathcal{X}$, if finite, is denoted by $N(\mathcal{X}, \epsilon)$ and is called the covering number of $\mathcal{X}$.

LEMMA **7.1**. *For $\epsilon > 0$, there exists an $\epsilon$-net of $\mathcal{F}$, denoted by $\mathcal{F}_\epsilon$, such that card($\mathcal{F}_\epsilon$) $\le \left\{\frac{(2+\epsilon)ep}{\epsilon k}\right\}^k$ and*

$$\bigcap_{v \in \mathcal{F}_\epsilon} \{w \in \mathbb{R}^p : w'v \le (1-\epsilon)\sqrt{t}\} \subseteq \mathcal{A}(t;k) \subseteq \bigcap_{v \in \mathcal{F}_\epsilon} \{w \in \mathbb{R}^p : w'v \le \sqrt{t}\}. \tag{15}$$

*Proof of Lemma 7.1.* For the unit sphere $\mathbb{S}^{k-1}$ equipped with the Euclidean metric, it is well-known that the $\epsilon$-covering number $N(\mathbb{S}^{k-1}, \epsilon) \le (1+2/\epsilon)^k$, see e.g. Lemma 5.2 of Vershynin (2012). Notice that

$$\mathcal{F} = \{v \in \mathbb{S}^{p-1}, ||v||_0 \le k\} = \bigcup_{S \subseteq [p]: ||S||_0 = k} \{v \in \mathbb{S}^{p-1} : v_S \in \mathbb{S}^{k-1}\},$$

where $[p] = \{1, 2, \ldots, p\}$. Because $\binom{p}{k} \le (ep/k)^k$, we have

$$N(\mathcal{F}, \epsilon) \le \binom{p}{k}\left(1 + \frac{2}{\epsilon}\right)^k \le \left\{\frac{(2+\epsilon)ep}{\epsilon k}\right\}^k.$$

Recall that

$$\mathcal{A}(t; k) = \bigcap_{v \in \mathcal{F}} \{w \in \mathbb{R}^p : w'v \le \sqrt{t}\}.$$

Let $\mathcal{F}_\epsilon$ be an $\epsilon$-net of $\mathcal{F}$ with cardinality $N(\mathcal{F}, \epsilon)$, and $A_1(t) := A_1(t; \epsilon) = \bigcap_{v \in \mathcal{F}_\epsilon} \{w \in \mathbb{R}^p : w'v \le (1-\epsilon)\sqrt{t}\}$. It is easy to see that

$$\mathcal{A}(t; k) \subseteq \bigcap_{v \in \mathcal{F}_\epsilon} \{w \in \mathbb{R}^p : w'v \le \sqrt{t}\}.$$

For any $v \in \mathcal{F}$, we can find $v_0 \in \mathcal{F}_\epsilon$ such that $|v - v_0| \le \epsilon$. Thus for $w \in A_1(t)$, we have

$$w'v = w'v_0 + |v - v_0| \frac{w'(v - v_0)}{|v - v_0|} \le (1-\epsilon)\sqrt{t} + \epsilon \max_{v_1 \in \mathcal{F}} w'v_1.$$

Taking maximum over $v \in \mathcal{F}$, we obtain $\max_{v \in \mathcal{F}} w'v \le (1-\epsilon)\sqrt{t} + \epsilon \max_{v_1 \in \mathcal{F}} w'v_1$, which implies that $\max_{v \in \mathcal{F}} w'v \le \sqrt{t}$ and thus $w \in \mathcal{A}(t; k)$. ♢

Consider the following optimization problem,

$$\max_{\theta \in \Theta_{a,k}} \sum_{i=1}^n \left\{X_i'\Gamma X_i - (X_i - \theta)'\Gamma(X_i - \theta)\right\} = 2n \max_{S:\mathrm{card}(S)=k} \max_{\theta:\mathrm{supp}(\theta)=S} \left(\theta'\Gamma\bar{X} - \frac{1}{2}\theta'\Gamma\theta\right), \tag{16}$$

where $\mathrm{supp}(\theta)$ denotes the support set of $\theta$. Some simple algebra gives the following result.

Lemma **7.2**. *The solution to the maximization over $\theta$ with supp$(\theta) = S$ in (16) is $\theta_S = \Gamma_{S,S}^{-1}(\Gamma_{S,\cdot}\bar{X})$ and the corresponding maximized value is equal to $(\Gamma_{S,\cdot}\bar{X})'\Gamma_{S,S}^{-1}(\Gamma_{S,\cdot}\bar{X})/2$.*

## 7.2 Proofs of the main results

Let $\widehat{\Sigma} = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'/n$. Denote by $\widehat{\Gamma}_j$ and $\Gamma_j$ the $j$th rows of $\widehat{\Gamma}$ and $\Gamma$ respectively. We first state the following result due to van de Geer et al. (2014).

Proposition **7.1** (van de Geer et al. (2014)). *Under Assumptions 2.1-2.2, we have*

$$\max_{1 \le j \le p} |\widehat{\gamma}_{jj} - \gamma_{jj}| = O_p\left(\sqrt{\frac{d\log(p)}{n}}\right), \tag{17}$$

$$\max_{1 \le j \le p} |\widehat{\Gamma}_j - \Gamma_j|_1 = O_p\left(d\sqrt{\frac{\log(p)}{n}}\right), \tag{18}$$

$$||\widehat{\Gamma}\widehat{\Sigma}\widehat{\Gamma}' - \widehat{\Gamma}'||_\infty = O_p\left(\sqrt{\frac{d\log(p)}{n}}\right). \tag{19}$$

By the arguments in the proofs of Lemma 5.3 and Lemma 5.4 in van de Geer et al. (2014), we have (17), (18) and (19) hold if $\widetilde{X}_i = X_i - \bar{X}$ is replaced by $X_i - \theta$ in the nodewise Lasso regression and $\widehat{\Sigma}$ is replaced by $\sum_{i=1}^n (X_i - \theta)(X_i - \theta)'/n$. A careful inspection of their proofs shows that the conclusion remains valid when $\theta$ is replaced with $\bar{X}$. We omit the technical details here to conserve space.

*Proof of Theorem 2.1.* The triangle inequality yields that

$$\begin{aligned}
&\sup_{t\geq 0}\left|P\left(T^*_{fe,n}(k)\leq t\middle|X_1^n\right)-P\left(T_{fe,n}(k)\leq t\right)\right| \\
\leq &\sup_{t\geq 0}\left|P\left(T_n(k)\leq t\right)-P\left(T_{fe,n}(k)\leq t\right)\right| \\
&+\sup_{t\geq 0}\left|P\left(T_n(k)\leq t\right)-P\left(T^*_{fe,n}(k)\leq t\middle|X_1^n\right)\right| := \rho_{1,n}+\rho_{2,n}.
\end{aligned}$$

We bound $\rho_{1,n}$ and $\rho_{2,n}$ in Step 1 and Step 2 respectively.

**Step 1 (bounding $\rho_{1,n}$)**: Let $\widehat{\xi}_j = \frac{|\widehat{z}_j|}{\sqrt{\widehat{\gamma}_{jj}}}$ and $\widehat{\xi}_{(j)}$ be the order statistic such that

$$\widehat{\xi}_{(1)} \geq \widehat{\xi}_{(2)} \geq \cdots \geq \widehat{\xi}_{(p)}.$$

Similarly we can define $\xi_j$ and $\xi_{(j)}$ in the same way as $\widehat{\xi}_j$ and $\widehat{\xi}_{(j)}$ by replacing $\widehat{\Gamma}$ with the precision matrix $\Gamma$. We have

$$|\widehat{\xi}_j-\xi_j| \leq \left|\frac{|\widehat{z}_j|}{\sqrt{\widehat{\gamma}_{jj}}}-\frac{|z_j|}{\sqrt{\widehat{\gamma}_{jj}}}\right|+\left|\frac{|z_j|}{\sqrt{\widehat{\gamma}_{jj}}}-\frac{|z_j|}{\sqrt{\gamma_{jj}}}\right| \leq \frac{|z_j-\widehat{z}_j|}{\sqrt{\widehat{\gamma}_{jj}}}+\left|\frac{\sqrt{\gamma_{jj}}-\sqrt{\widehat{\gamma}_{jj}}}{\sqrt{\gamma_{jj}}\sqrt{\widehat{\gamma}_{jj}}}\right||z_j| := I_1+I_2.$$

By Proposition 7.1, we have $\max_{1\leq j\leq p}|\widehat{\Gamma}_j-\Gamma_j|_1 = O_p(d\sqrt{\log(p)/n})$ and $\sup_{1\leq j\leq p}|\gamma_{jj}-\widehat{\gamma}_{jj}| = O_p(\sqrt{d\log(p)/n})$. Also note that $c_1 < \min_{1\leq j\leq p}\gamma_{jj} \leq \max_{1\leq j\leq p}\gamma_{jj} < c_2$ for some constants $0 < c_1 \leq c_2 < \infty$. Together with the fact that $|\bar{X}|_\infty = O_p(\sqrt{\log(p)/n})$, we deduce

$$\sup_{1\leq j\leq p}|z_j-\widehat{z}_j| \leq \max_{1\leq j\leq p}|\widehat{\Gamma}_j-\Gamma_j|_1|\bar{X}|_\infty = O_p\left(d\log(p)/n\right),$$

and

$$\sup_{1\leq j\leq p}|\sqrt{\gamma_{jj}}-\sqrt{\widehat{\gamma}_{jj}}| = \sup_{1\leq j\leq p}\left|\frac{\gamma_{jj}-\widehat{\gamma}_{jj}}{\sqrt{\gamma_{jj}}+\sqrt{\widehat{\gamma}_{jj}}}\right| = O_p\left(\sqrt{d\log(p)/n}\right). \tag{20}$$

As $\sup_{1\leq j\leq p}|z_j| = O_p(\sqrt{\log(p)/n})$, we obtain

$$\sup_{1\leq j\leq p}|\widehat{\xi}_j-\xi_j| = O_p\left(d\log(p)/n\right),$$

and

$$\sup_{1\leq j\leq p}|\widehat{\xi}_j^2-\xi_j^2| = O_p\left(d\log(p)/n\right)\sup_{1\leq j\leq p}|\widehat{\xi}_j+\xi_j| = O_p\left(d(\log(p)/n)^{3/2}\right).$$

Thus we deduce that

$$\begin{aligned}|T_n(k)-T_{fe,n}(k)| \leq & \max_{1\leq j_1<j_2<\cdots<j_k\leq p} n\sum_{l=1}^{k}\left|\xi_{j_l}^2-\widehat{\xi}_{j_l}^2\right| \\ \leq & nk\max_{1\leq j\leq p}|\widehat{\xi}_j^2-\xi_j^2| = O_p\left(kd(\log(p))^{3/2}/\sqrt{n}\right).\end{aligned}$$

By the assumption $k^2d(\log(np))^{5/2}/\sqrt{n}=o(1)$, we can pick $\zeta_1$ and $\zeta_2$ such that

$$P(|T_n(k)-T_{fe,n}(k)|\geq\zeta_1)\leq\zeta_2,$$

where $\zeta_1 k\log(np)=o(1)$ and $\zeta_2=o(1)$. Define the event $\mathcal{B}=\{|T_n(k)-T_{fe,n}(k)|<\zeta_1\}$. Then we have

$$\begin{aligned}& |P(T_n(k)\leq t)-P(T_{fe,n}(k)\leq t)| \\ \leq & P(T_n(k)\leq t, T_{fe,n}(k)>t)+P(T_n(k)>t, T_{fe,n}(k)\leq t) \\ \leq & P(T_n(k)\leq t, T_{fe,n}(k)>t,\mathcal{B})+P(T_n(k)>t, T_{fe,n}(k)\leq t,\mathcal{B})+2\zeta_2 \\ \leq & P(t-\zeta_1<T_n(k)\leq t)+P(t+\zeta_1\geq T_n(k)>t)+2\zeta_2.\end{aligned}$$

Let $V_i=\text{diag}^{-1/2}(\Gamma)\Gamma X_i$ and $V=\sum_{i=1}^n V_i/\sqrt{n}$. Notice that $\{T_n(k)\leq t\}=\{V\in\mathcal{A}(t;k)\}=\{\max_{v\in\mathcal{F}}v'V\leq\sqrt{t}\}$. By Lemma 7.1, we can find an $\epsilon$-net $\mathcal{F}_\epsilon$ of $\mathcal{F}$ such that $||\mathcal{F}_\epsilon||_0\leq\{(2+\epsilon)ep/(\epsilon k)\}^k$ and

$$A_1(t):=\bigcap_{v\in\mathcal{F}_\epsilon}\{v'V\leq(1-\epsilon)\sqrt{t}\}\subseteq\mathcal{A}(t;k)\subseteq A_2(t):=\bigcap_{v\in\mathcal{F}_\epsilon}\{v'V\leq\sqrt{t}\}.$$

We set $\epsilon=1/n$ throughout the following arguments. Notice that

$$\begin{aligned}& P(t-\zeta_1<T_n(k)\leq t) \\ = & P(T_n(k)\leq t)-P(T_n(k)\leq t-\zeta_1) \\ \leq & P(\max_{v\in\mathcal{F}_\epsilon}v'V\leq\sqrt{t})-P(\max_{v\in\mathcal{F}_\epsilon}v'V\leq(1-\epsilon)\sqrt{t-\zeta_1}) \\ \leq & P((1-\epsilon)(\sqrt{t}-\sqrt{\zeta_1})\leq\max_{v\in\mathcal{F}_\epsilon}v'V\leq\sqrt{t}) \\ \leq & P((1-\epsilon)(\sqrt{t}-\sqrt{\zeta_1})\leq\max_{v\in\mathcal{F}_\epsilon}v'V\leq(1-\epsilon)\sqrt{t})+P((1-\epsilon)\sqrt{t}<\max_{v\in\mathcal{F}_\epsilon}v'V\leq\sqrt{t}) \\ := & I_1+I_2.\end{aligned}$$

Because $\phi(\Gamma;k)>c>0$, we have $\text{var}(\sum_{i=1}^n v'V_i/\sqrt{n})>c'$ for all $v\in\mathcal{F}$ and some constant $c'>0$. By the Nazarov inequality [see Lemma A.1 in Chernozhukov et al. (2015) and Nazarov (2003)], we have

$$I_1\leq C\sqrt{\zeta_1 k\log(np/k)}=o(1). \tag{21}$$

To deal with $I_2$, we note when $t \le k^3\{\log(np/k)\}^2$, $\epsilon\sqrt{t} \le k^{3/2}\log(np/k)/n$. Again by the Nazarov's inequality, we have

$$I_2 \le P(\sqrt{t} - k^{3/2}\log(np/k)/n < \max_{v\in\mathcal{F}_\epsilon} v'V \le \sqrt{t}) \le k^2\log(np/k)\sqrt{\log(np/k)}/n = o(1).$$

When $t > k^3\{\log(np/k)\}^2$, we have

$$I_2 \le P((1-\epsilon)\sqrt{t} \le \max_{v\in\mathcal{F}_\epsilon} v'V) \le \frac{\mathbb{E}\max_{v\in\mathcal{F}_\epsilon} v'V}{(1-\epsilon)k^{3/2}\log(np/k)}.$$

By Lemma 7.4 in Fan et al. (2015), we have $\mathbb{E}\max_{v\in\mathcal{F}_\epsilon} v'V \le C\sqrt{k\log(np/k)}$. It thus implies that

$$I_2 \le \frac{C\sqrt{k\log(np/k)}}{(1-\epsilon)k^{3/2}\log(np/k)} = o(1).$$

Summarizing the above derivations, we have $\rho_{1,n} = o(1)$.

**Step 2 (bounding $\rho_{2,n}$):** Define $\widehat{V}^* = \text{diag}^{-1/2}(\widehat{\Gamma})\widehat{\Gamma}\sum_{i=1}^n (X_i - \bar{X})e_i/\sqrt{n}$ with $e_i \sim^{i.i.d} N(0,1)$, where $e_i$'s are independent of $X_1^n$. Further define

$$\bar{\rho} = \max\{|P(V\in A_1(t)) - P(\widehat{V}^*\in A_1(t)|X_1^n)|, |P(V\in A_2(t)) - P(\widehat{V}^*\in A_2(t)|X_1^n)|\}.$$

Using similar arguments in Step 1, we have

$$\begin{aligned} P(\widehat{V}^* \in \mathcal{A}(t;k)|X_1^n) \le & P(\widehat{V}^*\in A_2(t)|X_1^n) \le P(V\in A_2(t)) + \bar{\rho} \\ \le & P(V\in A_1(t)) + \bar{\rho} + o(1) \\ \le & P(V\in\mathcal{A}(t;k)) + \bar{\rho} + o(1). \end{aligned}$$

Similarly we have $P(\widehat{V}^*\in\mathcal{A}(t;k)|X_1^n) \ge P(V\in\mathcal{A}(t;k)) - \bar{\rho} - o(1)$. Together, we obtain

$$|P(\widehat{V}^*\in\mathcal{A}(t;k)|X_1^n) - P(V\in\mathcal{A}(t;k))| \le \bar{\rho} + o(1).$$

Let $D = \text{diag}^{-1/2}(\Gamma)\Gamma\text{diag}^{-1/2}(\Gamma)$ and $\widehat{D} = \text{diag}^{-1/2}(\widehat{\Gamma})\widehat{\Gamma}\widehat{\Sigma}\widehat{\Gamma}'\text{diag}^{-1/2}(\widehat{\Gamma})$, where $\widehat{\Sigma} = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'/n$. Define $\Delta_n = \max_{u,v\in\mathcal{F}} |u(\widehat{D} - D)v|$. Notice that $V \sim N(0, D)$ and $\widehat{V}^*|X_1^n \sim N(0, \widehat{D})$. To bound $\bar{\rho}$, we note that by equation (49) in Chernozhukov et al. (2015),

$$\begin{aligned} & \sup_{t\ge 0} |P(V\in A_1(t)) - P(\widehat{V}^*\in A_1(t)|X_1^n)| \\ = & \sup_{t\ge 0} |P(\max_{v\in\mathcal{F}_\epsilon} u'V \le \sqrt{t}) - P(\max_{u\in\mathcal{F}_\epsilon} u'\widehat{V}^* \le \sqrt{t}|X_1^n)| \\ \le & C\Delta_n^{1/3}(k\log(np/k))^{2/3}. \end{aligned}$$

and similarly $|P(V\in A_2(t)) - P(\widehat{V}^*\in A_2(t)|X_1^n)| \le C\Delta_n^{1/3}(k\log(np/k))^{2/3}$. Therefore we get

$$\rho_{2,n} \le C\Delta_n^{1/3}(k\log(np/k))^{2/3} + o(1).$$

**Step 3**: Finally we bound $\Delta_n$. Note that for any $u, v \in \mathcal{F}$,

$$
\begin{aligned}
&|u'(\widehat{D}-D)v| \\
\leq &|u'(\widehat{D}-\operatorname{diag}^{-1/2}(\widehat{\Gamma})\widehat{\Gamma}'\operatorname{diag}^{-1/2}(\widehat{\Gamma}))v| + |u'(\operatorname{diag}^{-1/2}(\widehat{\Gamma})\widehat{\Gamma}'\operatorname{diag}^{-1/2}(\widehat{\Gamma}) - \operatorname{diag}^{-1/2}(\widehat{\Gamma})\Gamma\operatorname{diag}^{-1/2}(\widehat{\Gamma}))v| \\
&+ |u'(\operatorname{diag}^{-1/2}(\widehat{\Gamma})\Gamma\operatorname{diag}^{-1/2}(\widehat{\Gamma}) - D)v| := J_1 + J_2 + J_3.
\end{aligned}
$$

For the first term, we have

$$
\begin{aligned}
J_1 =&|u'\operatorname{diag}^{-1/2}(\widehat{\Gamma})(\widehat{\Gamma}\widehat{\Sigma}'\widehat{\Gamma} - \widehat{\Gamma}')\operatorname{diag}^{-1/2}(\widehat{\Gamma})v| \\
\leq&|\operatorname{diag}^{-1/2}(\widehat{\Gamma})u|_1|(\widehat{\Gamma}\widehat{\Sigma}'\widehat{\Gamma} - \widehat{\Gamma}')\operatorname{diag}^{-1/2}(\widehat{\Gamma})v|_\infty \\
\leq&|\operatorname{diag}^{-1/2}(\widehat{\Gamma})u|_1|\operatorname{diag}^{-1/2}(\widehat{\Gamma})v|_1||\widehat{\Gamma}\widehat{\Sigma}\widehat{\Gamma}' - \widehat{\Gamma}'||_\infty \\
\leq&k|\operatorname{diag}^{-1/2}(\widehat{\Gamma})u|_2|\operatorname{diag}^{-1/2}(\widehat{\Gamma})v|_2||\widehat{\Gamma}\widehat{\Sigma}\widehat{\Gamma}' - \widehat{\Gamma}'||_\infty = O_p(k\sqrt{d\log(p)/n}),
\end{aligned}
$$

where we have used Proposition 7.1. To handle the second term, note that

$$
\begin{aligned}
J_2 =&|v'\operatorname{diag}^{-1/2}(\widehat{\Gamma})(\widehat{\Gamma}-\Gamma)\operatorname{diag}^{-1/2}(\widehat{\Gamma})u| \\
\leq&|\operatorname{diag}^{-1/2}(\widehat{\Gamma})v|_1|(\widehat{\Gamma}-\Gamma)\operatorname{diag}^{-1/2}(\widehat{\Gamma})u|_\infty \\
\leq&\sqrt{k}|\operatorname{diag}^{-1/2}(\widehat{\Gamma})v|_2|\operatorname{diag}^{-1/2}(\widehat{\Gamma})u|_\infty \max_{1\leq j\leq p}|\widehat{\Gamma}_j - \Gamma_j|_1 = O_p(d\sqrt{k\log(p)/n}).
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
J_3 \leq&|u'(\operatorname{diag}^{-1/2}(\widehat{\Gamma})\Gamma\operatorname{diag}^{-1/2}(\widehat{\Gamma}) - \operatorname{diag}^{-1/2}(\Gamma)\Gamma\operatorname{diag}^{-1/2}(\widehat{\Gamma}))v| + |u'(\operatorname{diag}^{-1/2}(\Gamma)\Gamma\operatorname{diag}^{-1/2}(\widehat{\Gamma}) - D)v| \\
\leq&||\Gamma||_2||\operatorname{diag}^{-1/2}(\widehat{\Gamma})||_2 \max_{1\leq j\leq p}|1/\sqrt{\gamma_{jj}} - 1/\sqrt{\widehat{\gamma}_{jj}}| + ||\Gamma||_2||\operatorname{diag}^{-1/2}(\Gamma)||_2 \max_{1\leq j\leq p}|1/\sqrt{\gamma_{jj}} - 1/\sqrt{\widehat{\gamma}_{jj}}| \\
=&\sqrt{d\log(p)/n}.
\end{aligned}
$$

Under the assumption that $k^2d(\log(np))^{5/2}/\sqrt{n} = o(1)$, we have $(\log(np))^2 J_i = o_p(1)$ for $1 \leq i \leq 3$. Therefore we get $(\log(np))^2\Delta_n = o_p(1)$, which implies that $\rho_{2,n} = o_p(1)$. The proof is thus completed by combining Steps 1-3. ♢

*Proof of Theorem 2.2.* We first note that by the triangle inequality,

$$
\begin{aligned}
&\sup_{t_M\geq t_{M-1}\geq\cdots\geq t_1\geq 0}\left|P\left(\bigcap_{j=1}^M\{T^*_{fe,n}(j)\leq t_j\}\Big|X_1^n\right) - P\left(\bigcap_{j=1}^M\{T_{fe,n}(j)\leq t_j\}\right)\right| \\
\leq&\sup_{t_M\geq t_{M-1}\geq\cdots\geq t_1\geq 0}\left|P\left(\bigcap_{j=1}^M\{T_n(j)\leq t_j\}\right) - P\left(\bigcap_{j=1}^M\{T_{fe,n}(j)\leq t_j\}\right)\right| \\
&+\sup_{t_M\geq t_{M-1}\geq\cdots\geq t_1\geq 0}\left|P\left(\bigcap_{j=1}^M\{T_n(j)\leq t_j\}\right) - P\left(\bigcap_{j=1}^M\{T^*_{fe,n}(j)\leq t_j\}\Big|X_1^n\right)\right| := \varrho_{1,n} + \varrho_{2,n}.
\end{aligned}
$$

**Step 1 (bounding $\varrho_{1,n}$)**: Following the proof of Theorem 2.1, we have for any $1 \leq j \leq M$,

$$\max_{1\leq j\leq M} |T_n(j) - T_{fe,n}(j)| \leq nM \max_{1\leq j\leq p} |\widehat{\xi}_j^2 - \xi_j^2| = O_p\left(Md(\log(p))^{3/2}/\sqrt{n}\right).$$

Under the assumption that $M^4 d(\log(np))^{5/2}/\sqrt{n} = o(1)$, one can pick $\zeta$ such that $P(\max_{1\leq j\leq M} |T_n(j) - T_{fe,n}(j)| > \zeta) = o(1)$ and $\zeta M^3 \log(np) = o(1)$. Define $\mathcal{B} = \{\max_{1\leq j\leq M} |T_n(j) - T_{fe,n}(j)| \leq \zeta\}$. We note that

$$\begin{aligned}
&\left| P\left(\bigcap_{j=1}^M \{T_n(j) \leq t_j\}\right) - P\left(\bigcap_{j=1}^M \{T_{fe,n}(j) \leq t_j\}\right)\right| \\
\leq & P\left(\bigcap_{j=1}^M \{T_n(j) \leq t_j\}, \bigcup_{j=1}^M \{T_{fe,n}(j) > t_j\}, \mathcal{B}\right) \\
& + P\left(\bigcup_{j=1}^M \{T_n(j) > t_j\}, \bigcap_{j=1}^M \{T_{fe,n}(j) \leq t_j\}, \mathcal{B}\right) + o(1) \\
\leq & P\left(\bigcup_{j=1}^M \{t_j - \zeta < T_n(j) \leq t_j\}\right) + P\left(\bigcup_{j=1}^M \{t_j < T_n(j) \leq t_j + \zeta\}\right) + o(1).
\end{aligned}$$

By the arguments in the proof of Theorem 2.1, we have $P\left(t_j - \zeta < T_n(j) \leq t_j\right) = o(1)$. A careful inspection of the proof shows that

$$\begin{aligned}
&\max_{1\leq j\leq M} \max\{P\left(t_j - \zeta < T_n(j) \leq t_j\right), P\left(t_j < T_n(j) \leq t_j + \zeta\right)\} \\
\leq & C\left\{\sqrt{\zeta M \log(np)} + M^2(\log(np))^{3/2}/n + 1/(M\sqrt{\log(np/M)})\right\},
\end{aligned}$$

where the uniformity over $1 \leq j \leq M$ is due to the fact that the constant $C$ in (21) is independent of $t$. By the union bound, we deduce that

$$\begin{aligned}
&\left| P\left(\bigcap_{j=1}^M \{T_n(j) \leq t_j\}\right) - P\left(\bigcap_{j=1}^M \{T_{fe,n}(j) \leq t_j\}\right)\right| \\
\leq & \sum_{j=1}^M \left(P\left(t_j - \zeta < T_n(j) \leq t_j\right) + P\left(t_j < T_n(j) \leq t_j + \zeta\right)\right) + o(1) \\
\leq & M \max_{1\leq j\leq M} \left(P\left(t_j - \zeta < T_n(j) \leq t_j\right) + P\left(t_j < T_n(j) \leq t_j + \zeta\right)\right) + o(1) \\
\leq & C\left(\sqrt{\zeta M^3 \log(np)} + M^3(\log(np))^{3/2}/n + 1/\sqrt{\log(np/M)}\right) + o(1) = o(1).
\end{aligned}$$

**Step 2 (bounding $\varrho_{2,n}$)**: For $\mathbf{t} = (t_1, \ldots, t_M)$, define

$$\mathcal{A}(\mathbf{t}) = \bigcap_{j=1}^M \mathcal{A}(t_j; j) = \bigcap_{j=1}^M \bigcap_{S\subseteq[p],\mathrm{card}(S)=j} \{w \in \mathbb{R}^p : w_S' w_S \leq t_j\}.$$

It is easy to see that

$$\bigcap_{j=1}^{M}\{T_n(j)\le t_j\}=\{V\in\mathcal{A}(\mathbf{t})\},\quad \bigcap_{j=1}^{M}\{T^*_{fe,n}(j)\le t_j\}=\{\widehat{V}^*\in\mathcal{A}(\mathbf{t})\}.$$

By Lemma 7.1, we know for any fixed $\mathbf{t}$,

$$\mathbf{A}_1(\mathbf{t}):=\bigcap_{j=1}^{M}A_1(t_j)\subseteq\mathcal{A}(\mathbf{t})\subseteq\mathbf{A}_2(\mathbf{t}):=\bigcap_{j=1}^{M}A_2(t_j),$$

where $A_1(t_j)=\bigcap_{v\in\mathcal{F}_\epsilon(j)}\{w\in\mathbb{R}^p:w'v\le(1-\epsilon)\sqrt{t_j}\}$ and $A_2(t_j)=\bigcap_{v\in\mathcal{F}_\epsilon(j)}\{w\in\mathbb{R}^p:w'v\le\sqrt{t_j}\}$ with $\epsilon=1/n$ and $\mathcal{F}_\epsilon(j)$ being an $\epsilon$-net for $\mathcal{F}(j):=\{v\in\mathbb{S}^{p-1}:||v||_0\le j\}$. Note that $\mathbf{A}_1(\mathbf{t})$ and $\mathbf{A}_2(\mathbf{t})$ are both intersections of no more than $M\{(2+1/n)epn\}^M$ half spaces. Thus following the arguments in Steps 2 and 3 of the proof of Theorem 2.1, we can show that $\varrho_{2,n}=o_p(1)$, which completes our proof. $\diamondsuit$

*Proof of Theorem 2.3.* The proof contains two steps. In the first step, we establish the consistency of the infeasible test $T_n(k)$, while in the second step we further show that the estimation effect caused by replacing $\Gamma$ with $\widehat{\Gamma}$ is asymptotically negligible.

**Step 1:** Consider the infeasible test $T_n(k)=\max_{1\le j_1<j_2<\cdots<j_k\le n}\sum_{l=1}^{k}\frac{nz^2_{j_l}}{\gamma_{j_l,j_l}}$. Define $\widetilde{\theta}=(\widetilde{\theta}_1,\dots,\widetilde{\theta}_p)'$ with $\widetilde{\theta}_j=(\Gamma\theta)_j/\sqrt{\gamma_{jj}}$, and $q_j=\frac{z_j-(\Gamma\theta)_j}{\sqrt{\gamma_{jj}}}$. Note that $\frac{z_j^2}{\gamma_{jj}}=\left(q_j+\widetilde{\theta}_j\right)^2$. Also by Lemma 3 of Cai et al. (2014), we have for any $2r<a<1-2r$,

$$P\left(\max_{j\in H}|\widetilde{\theta}_j-\sqrt{\gamma_{jj}}\theta_j|=O(p^{r-a/2})\max_{j\in H}|\theta_j|\right)\to 1, \tag{22}$$

where $H$ denotes the support of $\theta$. Suppose $\sqrt{\gamma_{j_1^*,j_1^*}}\theta_{j_1^*}\ge\sqrt{\gamma_{j_2^*,j_2^*}}\theta_{j_2^*}\ge\cdots\ge\sqrt{\gamma_{j_p^*,j_p^*}}\theta_{j_p^*}$. We deduce that with probability tending to one,

$$\begin{aligned}
T_n(k)&=\max_{1\le j_1<j_2<\cdots<j_k\le n}\sum_{l=1}^{k}n\left(q^2_{j_l}+\widetilde{\theta}^2_{j_l}+2q_{j_l}\widetilde{\theta}_{j_l}\right)\\
&\ge\left(n\sum_{l=1}^{k}\gamma_{j_l^*,j_l^*}\theta^2_{j_l^*}+n\sum_{l=1}^{k}q^2_{j_l^*}+2n\sum_{l=1}^{k}q_{j_l^*}\sqrt{\gamma_{j_l^*j_l^*}}\theta_{j_l^*}\right)(1+o(1))\\
&\ge\left\{n\sum_{l=1}^{k}\gamma_{j_l^*,j_l^*}\theta^2_{j_l^*}+n\sum_{l=1}^{k}q^2_{j_l^*}-2n\left(\sum_{l=1}^{k}q^2_{j_l^*}\right)^{1/2}\left(\sum_{l=1}^{k}\gamma_{j_l^*j_l^*}\theta^2_{j_l^*}\right)^{1/2}\right\}(1+o(1)).
\end{aligned}$$

We claim that $n\sum_{l=1}^{k}q^2_{j_l}=O_p(k)$ for any $\widetilde{S}:=\{j_1,\dots,j_k\}\subseteq[p]$. Note that

$$\mathrm{var}\left(\sum_{l=1}^{k}(nq^2_{j_l}-1)/\sqrt{k}\right)=\frac{2}{k}\sum_{i,l=1}^{k}\gamma^2_{j_i,j_l}/(\gamma_{j_i,j_i}\gamma_{j_l,j_l})\le\frac{C}{k}\sum_{i,l=1}^{k}\gamma^2_{j_i,j_l}=\frac{C}{k}\mathrm{Tr}(\Gamma^2_{\widetilde{S},\widetilde{S}})\le C/\lambda^2_{\min}\le C,$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. Therefore conditional on $\widetilde{S}$, $n\sum_{l=1}^{k}q^2_{j_l}=O_p(k)$. As $q_j$

is independent of the non-zero locations of $\theta$, $n\sum_{l=1}^{k} q_{j_l^*}^2 = O_p(k)$.

By the assumption that $\sum_{l=1}^{k} \gamma_{j_l^*,j_l^*}\theta_{j_l^*}^2 \geq (2k+\epsilon)\log(p)/n$ and $n\sum_{l=1}^{k} q_{j_l^*}^2 = O_p(k)$, we obtain

$$T_n(k) \geq \left\{(2k+\epsilon)\log(p) + O_p(k) - O_p(\sqrt{(2k+\epsilon)k\log(p)})\right\}(1+o(1)). \tag{23}$$

Under Assumption 2.5, we have by Lemma 6 of Cai et al. (2014),

$$\max_{1\leq j_1<j_2<\cdots<j_k\leq p} n\sum_{l=1}^{k} q_{j_l}^2 \leq kn\max_{1\leq j\leq p} q_j^2 = \{2k\log(p) - k\log\log(p)\} + O_p(1).$$

As shown in Theorem 2.1, the bootstrap statistic $T_{fe,n}^*(k)$ imitates the sampling distribution of $\max_{1\leq j_1<j_2\cdots<j_k\leq n}\sum_{l=1}^{k} nq_{j_l}^2$. By Step 2 in the proof of Theorem 2.1, we have

$$c_\alpha^*(k) \leq \{2k\log p - k\log\log(p)\} + O_p(1). \tag{24}$$

Combining (23) and (24), we get

$$P(T_n(k) > c_\alpha^*(k)) \to 1.$$

**Step 2:** Next we quantify the difference between $T_n(k)$ and $T_{fe,n}(k)$. Note that

$$|T_n(k) - T_{fe,n}(k)| \leq \max_{1\leq j_1<j_2<\cdots<j_k\leq p} n\sum_{l=1}^{k}\left|\xi_{j_l}^2 - \widehat{\xi}_{j_l}^2\right| \leq nk\max_{1\leq j\leq p}|\widehat{\xi}_j^2 - \xi_j^2|.$$

We define $\widehat{\theta}_j$ and $\widehat{q}_j$ by replacing $\Gamma$ with $\widehat{\Gamma}$ in $\widetilde{\theta}_j$ and $q_j$. Simple algebra yields that

$$\widehat{\xi}_j^2 - \xi_j^2 = (\widehat{q}_j + \widehat{\theta}_j)^2 - (q_j + \widetilde{\theta}_j)^2 = (\widehat{q}_j - q_j + \widehat{\theta}_j - \widetilde{\theta}_j)(q_j + \widetilde{\theta}_j + \widehat{q}_j + \widehat{\theta}_j). \tag{25}$$

Using similar argument in Step 1 of the proof of Theorem 2.1, we obtain

$$\max_{1\leq j\leq p}|q_j - \widehat{q}_j| = O_p\left(d\log(p)/n\right),$$

and

$$\max_{1\leq j\leq p}|\widetilde{\theta}_j - \widehat{\theta}_j| = O_p(d\sqrt{\log(p)/n})\max_{1\leq j\leq p}|\theta_j| + O_p(\sqrt{d\log(p)/n})\max_{1\leq j\leq p}|\widetilde{\theta}_j|,$$

where we have used (18), (20), the triangle inequality and the fact that $\max_{1\leq j\leq p}|(\Gamma\theta)_j| \leq C\max_{1\leq j\leq p}|\widetilde{\theta}_j|$ for some $C>0$. By (22), we have with probability tending to one, $\max_{j\in H}|\widetilde{\theta}_j| \leq (C'+o(1))\max_{j\in H}|\theta_j|$ for $C'>0$. Define the event

$$A = \left\{\max_{1\leq j\leq p}|\theta_j| < C_0 k\sqrt{\log(p)/n}\right\},$$

for some large enough constant $C_0 > 0$. On $A$, we have $\max_{1\leq j\leq p}|\widetilde{\theta}_j - \widehat{\theta}_j| = O_p(kd\log(p)/n)$. In view of (25), we have on the event $A$,

$$|\widehat{\xi}_j^2 - \xi_j^2| = O_p(kd(\log(p)/n)^{3/2})$$

which implies that

$$|T_n(k) - T_{fe,n}(k)| \leq O_p(k^2 d(\log(p))^{3/2}/\sqrt{n}).$$

For any $\epsilon > 0$, pick $C''$ such that

$$P(|T_n(k) - T_{fe,n}(k)| \leq C'' k^2 d(\log(p))^{3/2}/\sqrt{n}|A) \geq 1-\epsilon.$$

Thus we have

$$\begin{aligned} P(T_{fe,n}(k) > c_\alpha^*(k)|A) \geq & P(T_n(k) > c_\alpha^*(k) + |T_n(k) - T_{fe,n}(k)||A) \\ \geq & P(T_n(k) > c_\alpha^*(k) + C'' k^2 d(\log(p))^{3/2}/\sqrt{n}|A) - \epsilon. \end{aligned} \tag{26}$$

Recall for $\zeta > 0$ with $\zeta k\log(np) = o(1)$, we have

$$P(t \leq T_n(k) \leq t+\zeta) = o(1).$$

See Step 1 in the proof of Theorem 2.1. Under the assumption that $k^2 d(\log(np))^{5/2}/\sqrt{n} = o(1)$, we have

$$\begin{aligned} & P(T_n(k) > c_\alpha^*(k)|A) - P(T_n(k) > c_\alpha^*(k) + C'' k^2 d(\log(p))^{3/2}/\sqrt{n}|A) \\ = & P(c_\alpha^*(k) < T_n(k) \leq c_\alpha^*(k) + C'' k^2 d(\log(p))^{3/2}/\sqrt{n}|A) = o(1). \end{aligned}$$

Together with (26) and the result in Step 1, we obtain

$$P(T_{fe,n}(k) > c_\alpha^*(k)|A) \geq P(T_n(k) > c_\alpha^*(k)|A) - o(1) - \epsilon \to 1-\epsilon.$$

Suppose $\max_{1\leq j\leq p}|\theta_j| = |\theta_{k_1^*}|$. On $A^c$, we have for large enough $C_0$,

$$T_n(k) \geq n\left(q_{k_1^*}^2 + \widetilde{\theta}_{k_1^*}^2 + 2q_{k_1^*}\widetilde{\theta}_{k_1^*}\right) \geq C_1 k\log(p) > 2k\log(p) - k\log\log(p),$$

which holds with probability tending to one. It implies that $P(T_n(k) > 2k\log(p) - k\log\log(p)|A^c) \to 1$. Similar argument indicates that

$$P(T_{fe,n}(k) > c_\alpha^*(k)|A^c) \to 1. \tag{27}$$

By (26) and (27), we deduce that

$$P(T_{fe,n}(k) > c_\alpha^*(k)) = P(A)P(T_{fe,n}(k) > c_\alpha^*|A) + P(A^c)P(T_{fe,n}(k) > c_\alpha^*|A^c) \to 1 - \epsilon P(A).$$

The conclusion follows as $\epsilon$ is arbitrary.

Finally, we show the consistency of $\widetilde{T}_{fe,n}(M)$. As $\widetilde{T}^*_{fe,n}(M)$ imitates the sampling distribution of $\widetilde{T}_{fe,n}(M)$ under the null, we know

$$\widetilde{c}^*_\alpha(M) = \sqrt{2M}\log(p)(1 + o_p(1)).$$

Therefore we have

$$P(\widetilde{T}_{fe,n}(M) > \widetilde{c}^*_\alpha(M)) \geq P\left(\sqrt{\frac{1 - M/n}{2M}}(T_{fe,n}(M) - M) > \widetilde{c}^*_\alpha(M)\right) \to 1.$$

◇

*Proof of Proposition 2.1.* Under Assumption 2.6, we have $|\bar{X}|_\infty = O_p(\sqrt{\log(p)/n})$ and $|Z|_\infty = O_p(\sqrt{\log(p)/n})$. By the arguments in van de Geer et al. (2014), (17), (18) and (19) still hold under the sub-gaussian assumption. Therefore using the same arguments in Step 1 of the proof of Theorem 2.1, we can pick $\zeta_1$ and $\zeta_2$ such that

$$P(|T_n(k) - T_{fe,n}(k)| > \zeta_1) \leq \zeta_2,$$

where $\zeta_1 k \log(np) = o(1)$ and $\zeta_2 = o(1)$. Thus we have

$$|P(T_n(k) \leq t) - P(T_{fe,n}(k) \leq t)| \leq P(t - \zeta_1 < T_n(k) \leq t) + P(t + \zeta_1 \geq T_n(k) > t) + 2\zeta_2.$$

By Lemma 7.1, we have

$$\begin{aligned}&P(t - \zeta_1 < T_n(k) \leq t)\\ \leq &P((1-\epsilon)(\sqrt{t} - \sqrt{\zeta_1}) \leq \max_{v \in \mathcal{F}_\epsilon} v'V \leq (1-\epsilon)\sqrt{t}) + P((1-\epsilon)\sqrt{t} \leq \max_{v \in \mathcal{F}_\epsilon} v'V \leq \sqrt{t}).\end{aligned}$$

Corollary 2.1 in Chernozukov et al. (2015) yields that

$$\begin{aligned}&P(t - \zeta_1 < T_n(k) \leq t)\\ \leq &P((1-\epsilon)(\sqrt{t} - \sqrt{\zeta_1}) \leq \max_{v \in \mathcal{F}_\epsilon} v'W \leq (1-\epsilon)\sqrt{t}) + P((1-\epsilon)\sqrt{t} \leq \max_{v \in \mathcal{F}_\epsilon} v'W \leq \sqrt{t}) + c_{n,p,k},\end{aligned}$$

where $\epsilon = 1/n$ and $c_{n,p,k} = C\{k\log(pn/k)\}^{7/6}/n^{1/6} = o(1)$ under the assumption in Proposition 2.1. Thus following the arguments in the proof of Theorem 2.1, we can show that

$$\sup_{t \geq 0} |P(T_{fe,n}(k) \leq t) - P(T_n(k) \leq t)| = o_p(1), \tag{28}$$

as we only need to deal with the Gaussian vector $W$ and the arguments are analogous as above.

On the other hand, by Lemma 7.1 and Corollary 2.1 in Chernozukov et al. (2015), we have

$$\begin{aligned} P\left(T_n(k) \leq t\right) - P\left(T^W(k) \leq t\right) \leq & P(\max_{v\in\mathcal{F}_\epsilon} v'V \leq \sqrt{t}) - P(\max_{v\in\mathcal{F}_\epsilon} v'W \leq (1-\epsilon)\sqrt{t}) \\ \leq & P((1-\epsilon)\sqrt{t} \leq \max_{v\in\mathcal{F}_\epsilon} v'W \leq \sqrt{t}) + c_{n,p,k}. \end{aligned} \tag{29}$$

Similarly $P\left(T^W(k) \leq t\right) - P\left(T_n(k) \leq t\right)$ can be bounded above by the same quantity on the RHS of (29). Thus we have

$$\sup_{t\geq 0} |P\left(T_n(k) \leq t\right) - P\left(T^W(k) \leq t\right)| \leq \sup_{t\geq 0} |P((1-\epsilon)\sqrt{t} \leq \max_{v\in\mathcal{F}_\epsilon} v'W \leq \sqrt{t})| + c_{n,p,k} = o(1), \tag{30}$$

where we have used the Nazarov inequality and Lemma 7.4 in Fan et al. (2015) to control the term $\sup_{t\geq 0} |P((1-\epsilon)\sqrt{t} \leq \max_{v\in\mathcal{F}_\epsilon} v'W \leq \sqrt{t})|$. The conclusion thus follows from (28) and (30). ♢

*Proof of Proposition 3.1.* The negative log-likelihood (up to a constant) is given by

$$l_n(\theta_1,\theta_2) = \frac{1}{2}\sum_{i=1}^{n_1}(X_i-\theta_1)'\Gamma_1(X_i-\theta_1) + \frac{1}{2}\sum_{i=1}^{n_2}(Y_i-\theta_2)'\Gamma_2(Y_i-\theta_2).$$

Under the null, we have $\theta := \theta_1 = \theta_2$. The MLE for $\theta$ is given by

$$\widetilde{\theta} = (n_1\Gamma_1 + n_2\Gamma_2)^{-1}\left(\Gamma_1\sum_{i=1}^{n_1}X_i + \Gamma_2\sum_{i=1}^{n_2}Y_i\right).$$

Define

$$(\widetilde{\Delta},\widetilde{\theta}_2) = \underset{\theta_2\in\mathbb{R}^p,\Delta\in\Theta_{a,k}}{\arg\min}\; l_n(\theta_2+\Delta,\theta_2)$$

and $\widetilde{\theta}_1 = \widetilde{\theta}_2 + \widetilde{\Delta}$. Taking the derivative of $l_n(\theta_2+\Delta,\theta_2)$ with respect to $\theta_2$ and setting it to be zero, we obtain

$$\widetilde{\theta}_2 = \widetilde{\theta} - C_1\widetilde{\Delta}.$$

Thus by direct calculation, we have

$$\begin{aligned} \min_{\theta_1-\theta_2\in\Theta_{a,k}} l_n(\theta_1,\theta_2) = & \min_{\Delta\in\Theta_{a,k}}\left[\frac{n_1}{2}\left\{\Delta'C_2'\Gamma_1C_2\Delta - 2\Delta'C_2'\Gamma_1(\bar{X}-\widetilde{\theta})\right\}\right. \\ & \left. + \frac{n_2}{2}\left\{\Delta'C_1'\Gamma_2C_1\Delta + 2\Delta'C_1'\Gamma_2(\bar{Y}-\widetilde{\theta})\right\}\right] + l_n(\widetilde{\theta},\widetilde{\theta}). \end{aligned}$$

The log-likelihood ratio test for testing $H_0'$ against $H_{a,k}'$ is given by

$$\begin{aligned} LR_n(k) =& 2l_n(\widetilde{\theta},\widetilde{\theta}) - 2 \min_{\theta_1-\theta_2\in\Theta_{a,k}} l_n(\theta_1,\theta_2) \\ =& \max_{\Delta\in\Theta_{a,k}} \Big[ n_1 \Big\{ 2\Delta' C_2'\Gamma_1(\bar{X}-\widetilde{\theta}) - \Delta' C_2'\Gamma_1 C_2 \Delta \Big\} \\ & - n_2 \Big\{ \Delta' C_1'\Gamma_2 C_1 \Delta + 2\Delta' C_1'\Gamma_2(\bar{Y}-\widetilde{\theta}) \Big\} \Big]. \end{aligned}$$

Define $\widetilde{X} = C_2'\Gamma_1(\bar{X}-\widetilde{\theta})$ and $\widetilde{Y} = C_1'\Gamma_2(\bar{Y}-\widetilde{\theta})$. Therefore we have

$$\begin{aligned} LR_n(k) &= \max_{S:\text{card}(S)=k} \max_{\Delta_S\in\mathbb{R}^k} \Big[ n_1 \Big\{ 2\Delta_S'\widetilde{X}_S - \Delta_S'\Omega_{S,S}^{21}\Delta_S \Big\} - n_2 \Big\{ \Delta_S'\Omega_{S,S}^{12}\Delta_S + 2\Delta_S'\widetilde{Y}_S \Big\} \Big] \qquad (31) \\ &= \max_{S:\text{card}(S)=k} (n_1\widetilde{X}_S - n_2\widetilde{Y}_S)'(n_1\Omega_{S,S}^{21} + n_2\Omega_{S,S}^{12})^{-1}(n_1\widetilde{X}_S - n_2\widetilde{Y}_S) \\ &= \max_{S:\text{card}(S)=k} (\Psi\bar{X} - \Psi\bar{Y})_S'(\Psi_{S,S})^{-1}(\Psi\bar{X} - \Psi\bar{Y}), \end{aligned}$$

where the maximizer in (31) is equal to $\hat{\Delta}_S = (n_1\Omega_{S,S}^{21} + n_2\Omega_{S,S}^{12})^{-1}(n_1\widetilde{X}_S - n_2\widetilde{Y}_S)$. ♢

# References


[1] Arias-Castro, E., Candès, E. J. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533-2556.

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

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

[4] Belloni, A., Chernozhukov, V. and Wang, L. (2012). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791-806.

[5] Bhatia, R., (1997). *Matrix Analysis*. Springer.

[6] Bickel, P. and Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical nonregular problem. In: Ghosh, et al. (Eds.), *Statistics and Probability*: A Raghu Raj Bahadur Festschrift, Wiley Eastern Limited, New Delhi, pp. 8396.

[7] Bickel, P. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199-227.

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

[9] Cai, T. and Liu, W. D. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Ass.* **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. Am. Statist. Ass.* **106** 594-607.

[11] Cai, T. T., Liu, W. D. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76** 349-372.

[12] Chen, S. X., Li, J. and Zhong, P.-S. (2014). Two-sample tests for high dimensional means with thresholding and data transformation. arXiv:1410.2848.

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

[14] Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Central limit theorems and bootstrap in high dimensions. arXiv:1412.3661.

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

[16] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962-994.

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

[18] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Am. Statist. Ass.* **91** 674-688.

[19] Fan, J., Shao, Q. and Zhou, W. (2015). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. arXiv:1502.04237.

[20] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432-441.

[21] Gentleman, R., Irizarry, R. A., Carey, V. J., Dudoit, S. and Huber, W. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer.

[22] Gregory, K. B., Carroll, R. J., Baladandayuthapani, V. and Lahiri, S. N. (2015). A two-sample test for equality of means in high dimension. *J. Am. Statist. Ass.* **110** 837-849.

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

[24] 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. Imgng* **13** 876-881.

[25] Jin, J., Zhang, C.-H. and Zhang, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723-2772.

[26] Ke, Z., Jin, J. and Fan, J. (2014). Covariance assisted screening and estimation. *Ann. Statist.* **42** 2202-2242.

[27] Li, J. and Zhong, P. S. (2015). A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence. *Ann. Statist.* To appear.

[28] Liu, H. and Wang, L. (2012). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. arXiv:1209.2437.

[29] Liu, X. and Shao, Y. Z. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Statist. Plann. Inference* **123** 61-81.

[30] Lopes, M., Jacob, L. and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems* 1206-1214.

[31] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

[32] Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In: *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics Volume 1807, Springer, pp. 169-187.

[33] Rockafellar, T. R. (1970). *Convex Analysis.* Princeton Univ. Press.

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

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

[36] Sun, T. and Zhang, C. H. (2013). Sparse matrix inversion with scaled lasso. *J. Machine Learning Research* **14** 3385-3418.

[37] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions for high dimensional models. *Ann. Statist.* **42** 1166-1202.

[38] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

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

[40] 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.

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

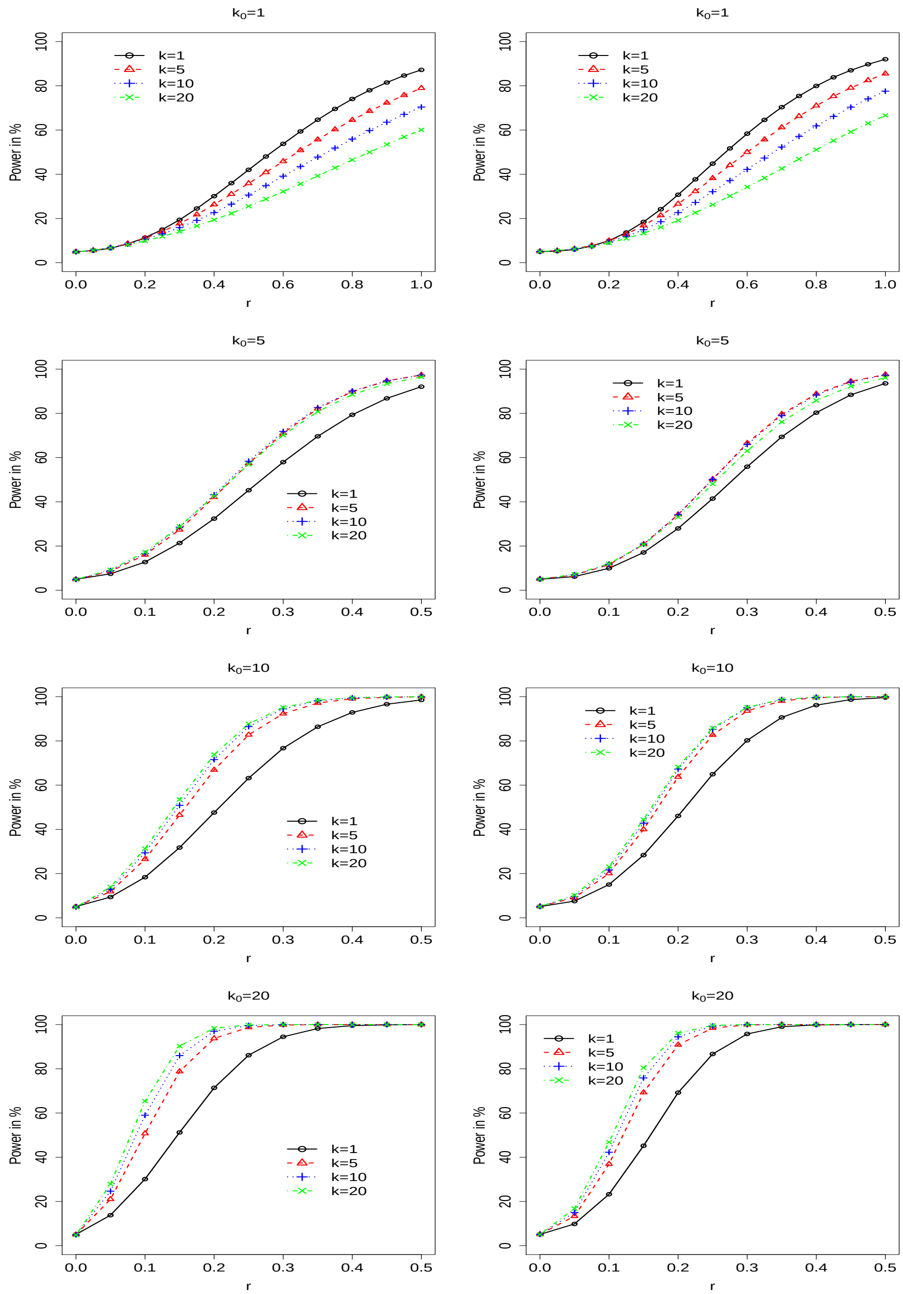



Figure 2: Power curves for $T_n(k)$, where $k = 1, 5, 10, 20$, and $\Sigma = (\sigma_{i,j})_{i,j=1}^{p}$ with $\sigma_{i,j} = 0.6^{|i-j|}$. Here $p = 200$ for the left panels and $p = 1000$ for the right panels. The number of Monte Carlo replications is 100000.

Table 2: Rejection probabilities in % for Models (a), (b), (c), and (d), where $p = 50, 100, 200$, and $n_1 = n_2 = 80$. The results are obtained based on 1000 Monte Carlo replications.

| Model | | $p$ | $T^2$ | BS | CQ | CLX | $T_{fe,n}(4)$ | $T_{fe,n}(8)$ | $T_{fe,n}(12)$ | $T_{fe,n}(24)$ | $\widetilde{T}_{fe,n}(40)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) | $H_0$ | 50 | 5.5 | 6.8 | 6.8 | 4.1 | 5.5 | 5.5 | 5.6 | 5.4 | 6.0 |
| | | 100 | 6.6 | 6.3 | 6.3 | 6.2 | 6.9 | 6.1 | 5.8 | 4.8 | 6.7 |
| | | 200 | NA | 5.1 | 5.1 | 5.7 | 6.8 | 5.3 | 5.5 | 4.3 | 5.1 |
| | Case 1 | 50 | 22.9 | 10.5 | 10.5 | 32.4 | 40.7 | 40.2 | 38.9 | 35.5 | 42.7 |
| | | 100 | 34.8 | 19.6 | 19.6 | 56.8 | 80.4 | 81.7 | 81.6 | 79.4 | 82.1 |
| | | 200 | NA | 28.9 | 28.9 | 81.7 | 96.5 | 97.7 | 98.4 | 98.6 | 98.5 |
| | Case 2 | 50 | 88.3 | 32.0 | 32.0 | 70.3 | 91.1 | 94.3 | 96.1 | 96.1 | 92.6 |
| | | 100 | 55.8 | 37.6 | 37.6 | 77.4 | 93.5 | 96.1 | 96.3 | 97.0 | 95.9 |
| | | 200 | NA | 42.0 | 42.0 | 97.9 | 99.9 | 100.0 | 100.0 | 100.0 | 100.0 |
| (b) | $H_0$ | 50 | 5.5 | 6.6 | 6.6 | 5.2 | 5.9 | 6.7 | 5.9 | 6.2 | 6.4 |
| | | 100 | 6.6 | 8.2 | 8.2 | 5.8 | 8.7 | 6.6 | 5.9 | 5.6 | 6.9 |
| | | 200 | NA | 5.8 | 5.8 | 6.5 | 8.1 | 6.9 | 6.0 | 4.9 | 6.0 |
| | Case 1 | 50 | 16.5 | 9.2 | 9.2 | 23.0 | 28.2 | 28.7 | 27.4 | 24.6 | 29.1 |
| | | 100 | 30.9 | 16.8 | 16.8 | 35.3 | 53.0 | 53.2 | 52.2 | 49.5 | 53.4 |
| | | 200 | NA | 22.8 | 22.8 | 57.7 | 80.3 | 84.0 | 84.9 | 83.3 | 84.1 |
| | Case 2 | 50 | 71.5 | 24.6 | 24.5 | 62.6 | 83.2 | 87.9 | 88.4 | 86.7 | 83.6 |
| | | 100 | 47.0 | 31.3 | 31.3 | 50.1 | 74.8 | 77.9 | 78.7 | 79.7 | 77.7 |
| | | 200 | NA | 33.8 | 33.8 | 78.5 | 93.3 | 95.3 | 95.6 | 95.9 | 95.9 |
| (c) | $H_0$ | 50 | 5.5 | 6.6 | 6.6 | 4.4 | 7.3 | 7.2 | 6.8 | 6.2 | 7.1 |
| | | 100 | 6.6 | 7.0 | 7.0 | 6.9 | 8.1 | 7.4 | 7.2 | 6.4 | 7.1 |
| | | 200 | NA | 5.9 | 5.9 | 7.5 | 8.2 | 7.4 | 6.3 | 5.8 | 6.5 |
| | Case 1 | 50 | 15.6 | 14.4 | 14.4 | 16.5 | 22.0 | 21.2 | 19.5 | 17.6 | 21.8 |
| | | 100 | 19.7 | 28.0 | 28.0 | 30.1 | 43.0 | 42.5 | 40.6 | 37.2 | 43.7 |
| | | 200 | NA | 47.9 | 47.9 | 45.9 | 69.4 | 71.9 | 71.4 | 68.4 | 71.1 |
| | Case 2 | 50 | 47.1 | 52.3 | 52.3 | 37.1 | 56.6 | 60.3 | 59.1 | 57.8 | 56.8 |
| | | 100 | 52.4 | 63.2 | 63.3 | 53.5 | 78.3 | 80.4 | 81.4 | 81.6 | 80.1 |
| | | 200 | NA | 70.0 | 70.0 | 60.0 | 82.9 | 86.6 | 87.5 | 87.9 | 87.7 |
| (d) | $H_0$ | 50 | 5.5 | 6.6 | 6.5 | 3.5 | 5.6 | 5.9 | 5.9 | 5.7 | 5.6 |
| | | 100 | 6.6 | 6.3 | 6.3 | 5.1 | 7.0 | 6.2 | 5.7 | 5.5 | 6.7 |
| | | 200 | NA | 5.6 | 5.6 | 5.3 | 6.5 | 6.1 | 5.2 | 4.6 | 5.5 |
| | Case 1 | 50 | 32.3 | 7.7 | 7.7 | 26.2 | 35.7 | 35.2 | 33.2 | 30.2 | 35.5 |
| | | 100 | 36.8 | 7.8 | 7.8 | 77.5 | 91.9 | 93.8 | 93.8 | 93.1 | 93.6 |
| | | 200 | NA | 7.4 | 7.4 | 96.1 | 99.6 | 100.0 | 100.0 | 100.0 | 100.0 |
| | Case 2 | 50 | 79.9 | 8.2 | 8.2 | 82.7 | 95.9 | 98.0 | 98.4 | 99.0 | 97.5 |
| | | 100 | 86.3 | 7.7 | 7.7 | 99.9 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| | | 200 | NA | 8.4 | 8.4 | 97.5 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Note: $T^2$, BS, CQ and CLX denote the Hotelling's $T^2$ test and the two-sample tests in Bai and Saranadasa (1996), Chen and Qin (2010), and Cai et al. (2014) respectively.

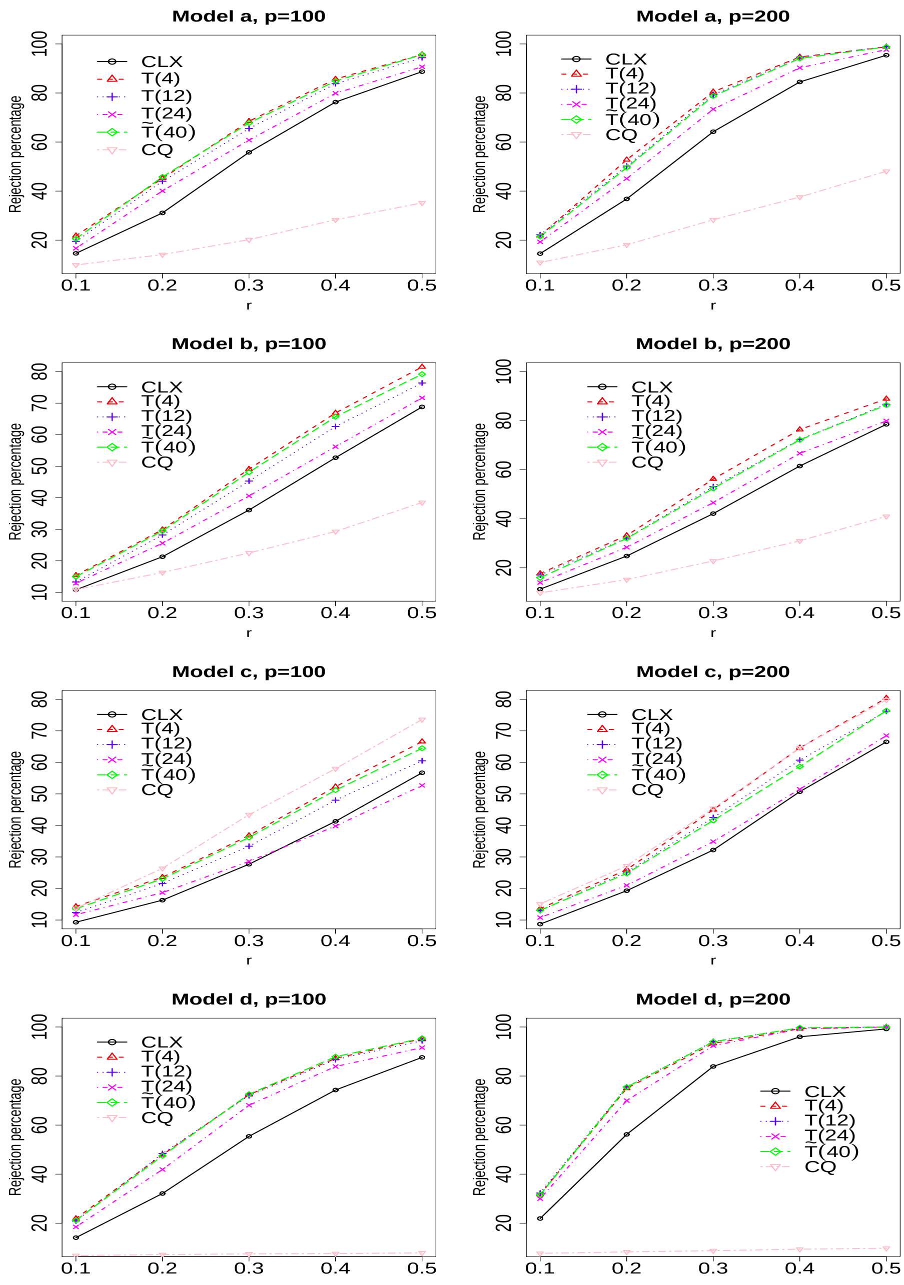



Figure 3: Empirical powers for CQ, CLX and the proposed tests under Models (a), (b), (c), and (d), and case 3, where $n_1 = n_2 = 80$ and $p = 100, 200$. The results are obtained based on 1000 Monte Carlo replications.

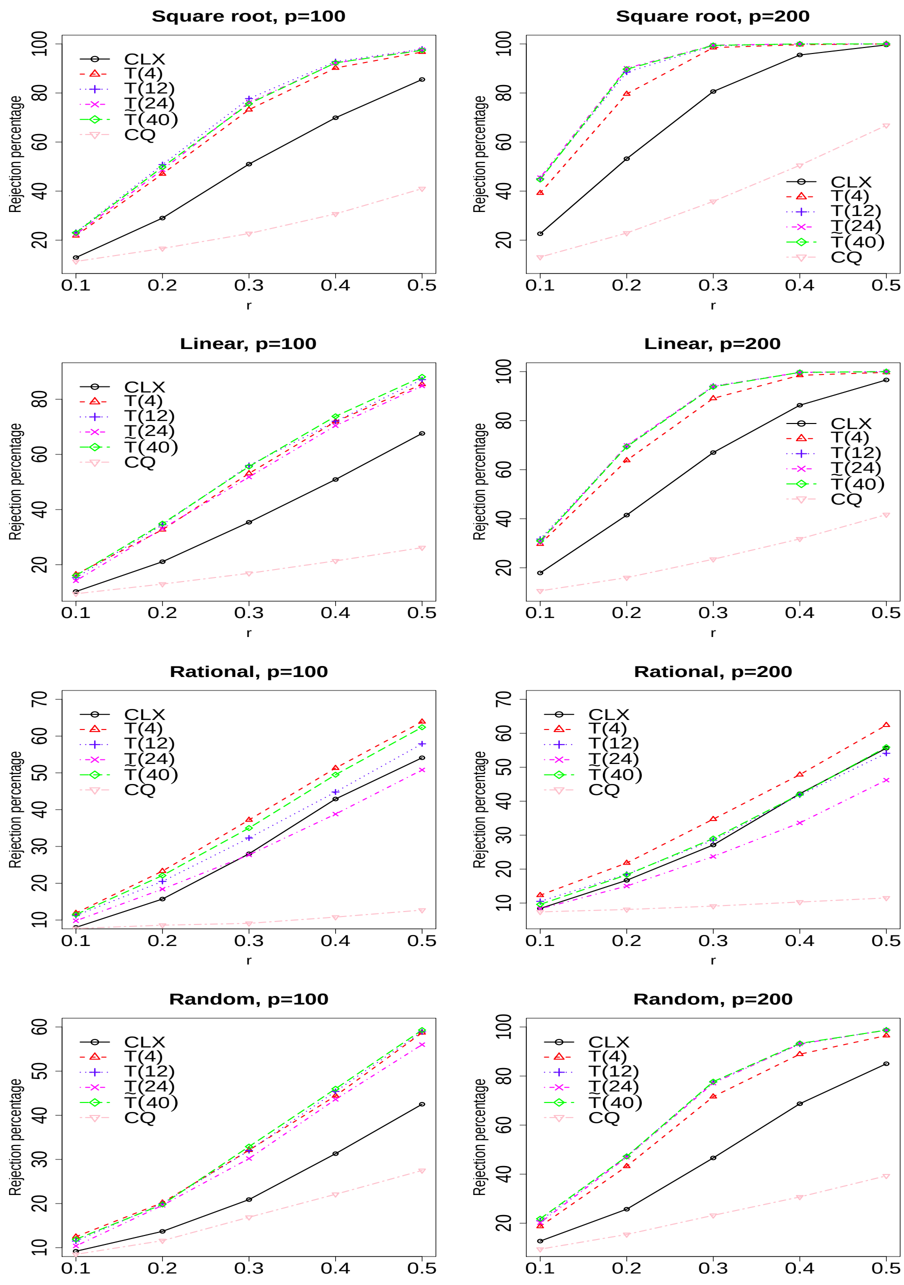
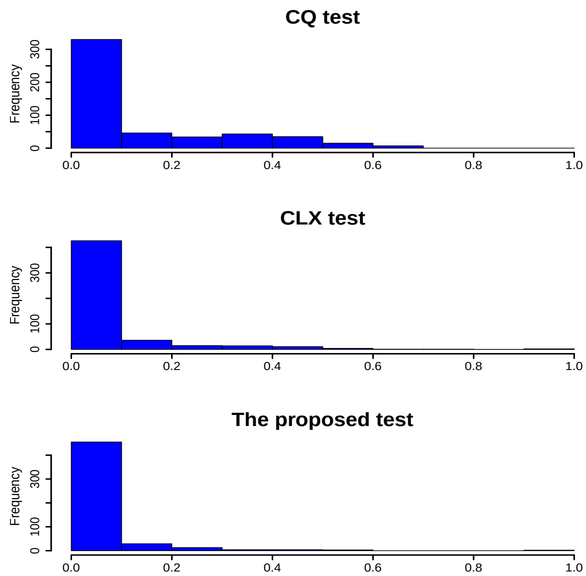
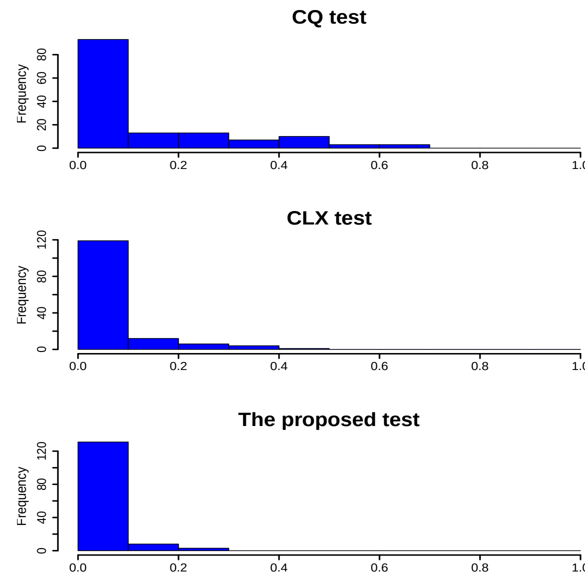


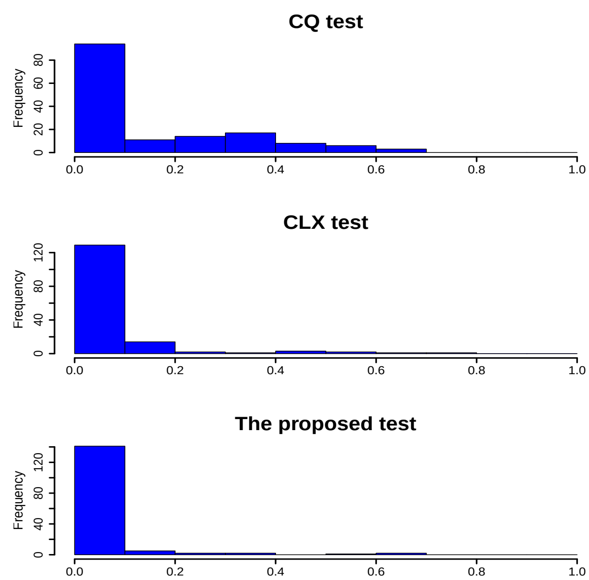

Figure 4: Empirical powers for CQ, CLX and the proposed tests under different signal allocations, where $n_1 = n_2 = 80$ and $p = 100, 200$. The results are obtained based on 1000 Monte Carlo replications.




(a) BP




(b) MF




(c) CC

Figure 5: Histograms of the p-values produced by the three tests.