Journal of Econometrics 235 (2023) 1337–1354

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Testing and signal identification for two-sample high-dimensional covariances via multi-level thresholding

Song Xi Chen [a], Bin Guo [b], Yumou Qiu [c,*]

[a] Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, 100871, China
[b] Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China
[c] Department of Statistics, Iowa State University, Ames, IA 50010, USA

ARTICLE INFO

Article history:
Received 23 October 2021
Received in revised form 15 September 2022
Accepted 15 October 2022
Available online 19 November 2022

JEL classification:
C12
C13
C15

Keywords:
Detection boundary
High dimensionality
Multiple testing
Rare and faint signal
Thresholding

ABSTRACT

The paper considers testing and signal identification for covariance matrices from two populations of marginally sub-Gaussian distributed. A multi-level thresholding procedure is proposed for testing the equality of two high-dimensional covariance matrices, which is designed to detect sparse and faint differences between the covariances. A novel U-statistic composition is developed to establish the asymptotic distribution of the thresholding statistics in conjunction with the matrix blocking and the coupling techniques. It is shown that the proposed test is more powerful than the existing tests in detecting sparse and weak signals in covariances. Multiple testing procedures are constructed to discover different covariances and the sub-groups of variables with different covariance structures between the two populations. The proposed procedures are based on the multi-level thresholding test, which are able to control the false discovery proportion (FDP) with high power. Simulation experiments and a case study on the returns of the S&P 500 stocks before and after the COVID-19 pandemic are conducted to demonstrate and compare the utilities of the proposed methods.

1. Introduction

Understanding how variables interact with each other is one of the key goals in scientific, economic and social research. Covariance which shows the linear association among variables is a commonly used measure for studying dependence. It is widely used in a variety of applications. For example, in portfolio allocation, the aim is to minimize a quadratic form of the covariance among many financial assets. The optimal allocation is determined by the inverse of the covariance (Markowitz, 1952). Understanding the covariance structure is also a basic step in data analysis as different dependence structures lead to different inference procedures. For instance, in the Hotelling's test for means and Fisher's linear discriminant analysis, the pooled covariance estimate is used under the assumption of the same covariance matrix between the two samples. The inverse covariance matrix is used in the Gaussian graphical model (Liu, 2013; Ren et al., 2015), and is utilized to enhance the signal strength in the innovated Higher Criticism test for high-dimensional means (Hall and Jin, 2010; Chen et al., 2019).

Detecting differences in covariance matrices among different experimental and treatment regimes are commonly pursued in econometrics and statistics. In fact, treatment effects may be reflected in the dependence structures in

* Corresponding author.
E-mail addresses: songxichen@pku.edu.cn (S.X. Chen), guobin@swufe.edu.cn (B. Guo), yumouqiu@iastate.edu (Y. Qiu).

https://doi.org/10.1016/j.jeconom.2022.10.008
0304-4076/

additional to means although the latter has been the main focus of treatment effect studies. In financial problems, it is of practice importance to check whether covariances among stocks change over time. If the covariance matrices between two time periods are different, the risk of a portfolio and the optimal allocation would be different (Fan et al., 2015a). The Gaussian graphical model (GGM) is commonly used in network analysis, where network connectivity is determined by the nonzero values in the inverse of a covariance matrix. Thus, different covariance matrices under different treatments or time periods imply different network dependence structures. In recent works, Liu et al. (2021) and Qiu et al. (2016) considered change point detection in GGM estimated at multiple time points and time-varying estimation of GGM, respectively. Two-sample covariance testing can be used as a preliminary step to check whether the network structure has changed between different periods. Therefore, testing for the equality of covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ from two populations is an important problem.

Nagao (1973) and Perlman (1980) presented studies under the conventional fixed dimensional setting; see Anderson (2003) for a comprehensive review. The modern high-dimensional data have generated a renewed interest under the so-called "large p, small n" paradigm. For Gaussian data with the dimension p and the sample size n being the same order, Schott (2007) and Srivastava and Yanagihara (2010) proposed two sample tests based on the distance measure $\|\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\|_F^2$, the squared Frobenius matrix norm between the two covariances. Bai et al. (2009) considered a corrected likelihood ratio test via the large dimensional random matrix theory. For nonparametric settings without explicitly restricting p and the sample sizes, Li and Chen (2012) proposed an ℓ_2-test based on a linear combination of U-statistics which is an unbiased estimator of $\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_F^2$. Qiu and Chen (2012) studied an ℓ_2-test for the bandedness of a covariance. Cai et al. (2013) proposed a test based on the maximal standardized differences (an $\ell_{\max}$-type formulation) between the entries of two sample covariance matrices. Chang et al. (2017) constructed a simulation based approach to approximate the distribution of the maximal statistics. Studies have shown that the ℓ_2-tests are powerful for detecting dense and weak differences (Chen et al., 2019), while the $\ell_{\max}$-formulation is powerful against sparse and strong signal settings. However, both types of tests encounter reduced power when the signals are rare and faint.

There are recent works in statistic and econometric literature that combine multiple tests to enhance power. Fan et al. (2015b) and Yu et al. (2021) proposed to add a thresholding statistic to an ℓ_2-statistic for testing high-dimensional means and covariances. He et al. (2021) constructed a family of U-statistics and combined them with an $\ell_{\max}$-type statistic. Although those tests are powerful to detect dense signals (due to the ℓ_2-component) and sparse and strong signals (due to the thresholding component), they are not powerful to detect signals that are both sparse and weak, which is an open question for covariance testing problems. Detailed discussion of those methods and comparison with the proposed approach are given after Proposition 3.

Detecting rare and faint signals has attracted much attention in high-dimensional statistical inference. Studies have been largely concentrated for the mean problems (Fan, 1996; Donoho and Jin, 2004; Delaigle et al., 2011; Zhong et al., 2013; Qiu et al., 2018), which built various versions of thresholding tests. However, studies of sparse and weak signals for covariance matrices are much less. Although thresholding statistics for means could be extended for covariances, the results of the means tests cannot be readily applied to covariance testing due to the complicated dependence among sample covariances. Meanwhile, Chudik et al. (2018) proposed a thresholding procedure for variable selection in high-dimensional linear regression. Arias-Castro et al. (2012) investigated the near optimal test for detecting nonzero correlations in a one sample setting with Gaussian data and clustered signals.

The first aim of this paper is on enhancing the power performance in testing differences between two covariances when the differences are both sparse and faint, which is the most challenging setting for signal detection and brings about the issue of detection boundary for covariance matrices. We introduce thresholding on the ℓ_2-formulation of Li and Chen (2012) to remove those non-signal bearing entries of the covariances, which reduces the overall noise (variance) level of the test statistic and increases the signal to noise ratio for the testing problem. A multi-level thresholding procedure is proposed to select the threshold level that maximizes a standardized thresholding statistic under the null hypothesis. This makes the proposed global test adaptive to a wider range of signal sparsity and strength. Under the setting of rare and faint differences between the two covariances, the power of the proposed global test is studied and its detection boundary is derived, which shows the benefits of the multi-thresholding over existing two sample covariance tests.

The second aim is to construct a multiple testing procedure based on a step-down procedure of the multi-level thresholding tests to identify the pairs of variables with different covariances between the two populations or treatments. It is shown that the step-down procedure with an augmentation step is able to control the exceedance rate of the false discovery proportion (FDP) with a higher power. A multiple testing procedure to recover the sub-groups of variables with different covariance structures is also proposed.

The paper is organized as follows. We introduce the setting of two-sample covariance testing in Section 2. The multi-level thresholding statistic is presented in Section 3, while Section 4 outlines a proposal for identifying different covariance pairs. Assumptions and asymptotic distributions of the proposed thresholding statistics are reported in Section 5, followed by theoretical power analysis in Section 6, establishing the detection boundary of the proposed global test and the properties of the multiple testing procedure. Simulation studies and a real data analysis on the returns of the S&P 500 stocks before and after the outbreak of the COVID-19 pandemic are presented in Sections 7 and 8, respectively. Further discussion with extensions of the proposed procedure to temporal dependent data and regression models with estimated parameters is made in Section 9. All technical proofs are relegated to the supplementary material (SM).

2. Preliminary

There are two independent samples of p-dimensional random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1} \overset{i.i.d.}{\sim} \mathrm{F}_1$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2} \overset{i.i.d.}{\sim} \mathrm{F}_2$ from two distributions F_1 and F_2, respectively, where $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^{\mathrm{T}}$, $\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kp})^{\mathrm{T}}$, n_1 and n_2 are the sample sizes, and "i.i.d". stands for "independent and identically distributed". Let $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})^{\mathrm{T}}$ and $\boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2p})^{\mathrm{T}}$ be the means of F_1 and F_2, and $\boldsymbol{\Sigma}_1 = (\sigma_{ij1})_{p\times p}$ and $\boldsymbol{\Sigma}_2 = (\sigma_{ij2})_{p\times p}$ be the covariance matrices of F_1 and F_2, respectively. Let $\boldsymbol{\Psi}_1 = (\rho_{ij1})_{p\times p}$ and $\boldsymbol{\Psi}_2 = (\rho_{ij2})_{p\times p}$ be the corresponding correlation matrices. We consider testing

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \quad \text{vs.} \quad H_a : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2 \tag{2.1}$$

under a high-dimensional setting where $p \gg n_1, n_2$.

Let $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 = (\delta_{ij})_{p\times p}$ where $\delta_{ij} = \sigma_{ij1} - \sigma_{ij2}$ are component-wise differences between $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, $q = p(p+1)/2$ be the number of distinct parameters and $n = n_1 n_2/(n_1+n_2)$ be the effective sample size in the testing problem. Let $\{\pi_{\ell,p}\}_{\ell=1}^{p!}$ denote all possible permutations of $\{1, \ldots, p\}$ and $\mathbf{X}_k(\pi_{\ell,p})$ and $\mathbf{Y}_k(\pi_{\ell,p})$ be the reordering of $\mathbf{X}_k$ and $\mathbf{Y}_k$ corresponding to a permutation $\pi_{\ell,p}$. To regulate the dependence among variables, we assume that there is a permutation $\pi_{*,p}$ such that $\mathbf{X}_k(\pi_{*,p})$ and $\mathbf{Y}_k(\pi_{*,p})$ are weakly dependent as defined via β-mixing (Bradley, 2005), or there exists a spatial structure on the variables of $\mathbf{X}_k$ and $\mathbf{Y}_k$ and they satisfy a spatial mixing condition (Jenish and Prucha, 2009) on their dependence. As the proposed statistic in (3.1) is of the ℓ_2-type and is invariant to permutations of $\mathbf{X}_k$ and $\mathbf{Y}_k$, there is no need to know the permutation $\pi_{*,p}$. In the following, we provide two examples for testing covariances in econometrics studies.

Example 1. It is known that the covariance structure is important in portfolio selection (Markowitz, 1952; Fan et al., 2015a). A specific interest is in detecting covariance changes over different time periods. Let $\tilde{\mathbf{X}}_{lk} = (\tilde{X}_{lk,1}, \ldots, \tilde{X}_{lk,p})^{\mathrm{T}}$ be the returns of p stocks on the kth day in period l for $l = 1, 2$. As the returns may be simultaneously affected by market overall performance, we consider the model (Fama and French, 1993)

$$\tilde{X}_{lk,j} = \rho_{lj}\tilde{X}_{lk-1,j} + \tilde{\mathbf{F}}_{lk}^{\mathrm{T}}\boldsymbol{\gamma}_{lj} + \boldsymbol{\epsilon}_{lk,j}$$

for $l = 1, 2$ and $j = 1, \ldots, p$, where $\tilde{\mathbf{F}}_{lk}$ stands for market common factors for all the stock returns, $\boldsymbol{\gamma}_{lj}$ is the coefficient of those common factors on the jth stock, and ρ_{lj} represents the effect of the previous day's return. Let $\boldsymbol{\epsilon}_{lk} = (\boldsymbol{\epsilon}_{lk,1}, \ldots, \boldsymbol{\epsilon}_{lk,p})^{\mathrm{T}}$ be the stock adjusted returns after removing the effects of market factors, and $\boldsymbol{\Sigma}_{\epsilon,l} = \mathrm{Cov}(\boldsymbol{\epsilon}_{lk})$ for $l = 1, 2$. We consider to test $H_0 : \boldsymbol{\Sigma}_{\epsilon,1} = \boldsymbol{\Sigma}_{\epsilon,2}$ over the two periods. Here, we assume $\{\boldsymbol{\epsilon}_{lk}\}$ over time are independent and the dependence among the components of $\boldsymbol{\epsilon}_{lk}$ satisfies a β-mixing condition after certain permutation, which would be reasonable as the stock adjusted returns are more likely to be dependent within the same industry sector, so that the covariance matrix exhibits an approximate block-diagonal structure. Also notice that the time independence assumption of $\{\boldsymbol{\epsilon}_{lk}\}$ can be relaxed to a weakly dependent assumption, which will be discussed in Section 9.

Example 2. Many economic data exhibit spatial features (Arbia and Baltagi, 2008) as space and distance affects economic behavior and regional economic dependence. Studying spatial dependence is a key aspect in spatial data analysis. Our setting covers the problem of testing spatial covariances. Let $\mathbf{X}_{lk} = (X_k(u_1), \ldots, X_k(u_p))^{\mathrm{T}}$ be the kth observation on a spatial domain with p locations in period l for $l = 1, 2$, where u_j denotes the jth location. Let $\boldsymbol{\Sigma}_{s,l} = \mathrm{Cov}(\mathbf{X}_{lk})$ be the spatial covariance of the observed variable. We are interested in testing the spatial covariances being the same over the two periods. Namely, consider the hypotheses $H_0 : \boldsymbol{\Sigma}_{s,1} = \boldsymbol{\Sigma}_{s,2}$ vs. $H_a : \boldsymbol{\Sigma}_{s,1} \neq \boldsymbol{\Sigma}_{s,2}$. Here, the observations are spatially weakly dependent in the sense that the dependence between $X_k(u_i)$ and $X_k(u_j)$ decreases as the distance between u_i and u_j increases.

Let $\bar{\mathbf{X}} = \sum_{k=1}^{n_1}\mathbf{X}_k/n_1$ and $\bar{\mathbf{Y}} = \sum_{k=1}^{n_2}\mathbf{Y}_k/n_2$ be the two sample means where $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_p)^{\mathrm{T}}$ and $\bar{\mathbf{Y}} = (\bar{Y}_1, \ldots, \bar{Y}_p)^{\mathrm{T}}$. Let

$$\widehat{\boldsymbol{\Sigma}}_1 = (\hat{\sigma}_{ij1}) = \frac{1}{n_1}\sum_{k=1}^{n_1}(\mathbf{X}_k - \bar{\mathbf{X}})(\mathbf{X}_k - \bar{\mathbf{X}})^{\mathrm{T}} \text{ and}$$

$$\widehat{\boldsymbol{\Sigma}}_2 = (\hat{\sigma}_{ij2}) = \frac{1}{n_2}\sum_{k=1}^{n_2}(\mathbf{Y}_k - \bar{\mathbf{Y}})(\mathbf{Y}_k - \bar{\mathbf{Y}})^{\mathrm{T}},$$

and $\kappa = \lim_{n_1,n_2\to\infty} n_1/(n_1+n_2)$. Moreover, let $\theta_{ij1} = \mathrm{Var}\{(X_{ki}-\mu_{1i})(X_{kj}-\mu_{1j})\}$, $\theta_{ij2} = \mathrm{Var}\{(Y_{ki}-\mu_{2i})(Y_{kj}-\mu_{2j})\}$; $\rho^{(1)}_{ij,lm} = \mathrm{Cor}\{(X_{ki}-\mu_{1i})(X_{kj}-\mu_{1j}), (X_{kl}-\mu_{1l})(X_{km}-\mu_{1m})\}$, and $\rho^{(2)}_{ij,lm} = \mathrm{Cor}\{(Y_{ki}-\mu_{2i})(Y_{kj}-\mu_{2j}), (Y_{kl}-\mu_{2l})(Y_{km}-\mu_{2m})\}$. Both θ_{ij1} and θ_{ij2} can be estimated by

$$\hat{\theta}_{ij1} = \frac{1}{n_1}\sum_{k=1}^{n_1}\{(X_{ki}-\bar{X}_i)(X_{kj}-\bar{X}_j) - \hat{\sigma}_{ij1}\}^2 \text{ and}$$

$$\hat{\theta}_{ij2} = \frac{1}{n_2}\sum_{k=1}^{n_2}\{(Y_{ki}-\bar{Y}_i)(Y_{kj}-\bar{Y}_j) - \hat{\sigma}_{ij2}\}^2.$$

As $\hat{\theta}_{ij1}/n_1+\hat{\theta}_{ij2}/n_2$ is a ratio-consistent estimator to the variance of $\hat{\sigma}_{ij1}-\hat{\sigma}_{ij2}$, we define a standardized difference between $\hat{\sigma}_{ij1}$ and $\hat{\sigma}_{ij2}$ as

$$M_{ij} = F_{ij}^2 \text{ for } F_{ij} = \frac{\hat{\sigma}_{ij1}-\hat{\sigma}_{ij2}}{(\hat{\theta}_{ij1}/n_1+\hat{\theta}_{ij2}/n_2)^{1/2}},\ 1\le i\le j\le p.$$

Cai et al. (2013) proposed a maximum statistic $M_n = \max_{1\le i\le j\le p} M_{ij}$ that targets at the largest signal between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Li and Chen (2012) proposed an ℓ_2-test that aims at $\|\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\|_F^2$. Donoho and Jin (2015) briefly discussed the possibility of applying the Higher Criticism (HC) statistic for testing $H_0: \mathbf{\Sigma}=\mathbf{I}_p$ with Gaussian data. We are to propose a test by carrying out multi-level thresholding on $\{M_{ij}\}$ to filter out potential signals via an ℓ_2-formulation, and show that such thresholding leads to a more powerful test than both the maximum test and the ℓ_2-type tests when the signals are rare and faint.

3. Multi-level thresholding test for covariances

To remove the non-signal bearing entries in sample covariances, we consider to threshold M_{ij} at $\lambda_p(s)=4s\log(p)$ for $s\in(0,1)$, and construct a thresholding statistic

$$T_n(s) = \sum_{1\le i\le j\le p} M_{ij}\mathbb{I}\{M_{ij}>\lambda_p(s)\}, \tag{3.1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The constrain on the threshold level s less than 1 is due to the large deviation result of F_{ij} that $\mathbb{P}\{\max_{1\le i\le j\le p}|F_{ij}|>2\sqrt{\log(p)}\}\to 0$ as $n,p\to\infty$ under H_0 of (2.1) and Assumption 1A (or 1B), 2, 3 (see Lemma 2 in SM). This implies the thresholding statistic $T_n(s)$ is asymptotically zero for $s\ge 1$ under H_0 of (2.1).

Compared with the ℓ_2-statistic of Li and Chen (2012), the statistic $T_n(s)$ removes those small standardized differences M_{ij} between $\widehat{\mathbf{\Sigma}}_1$ and $\widehat{\mathbf{\Sigma}}_2$, which results in a smaller variance than the ℓ_2-statistic that sums all pairs of M_{ij} for $i\le j$. Compared with the $\ell_{\max}$-test of Cai et al. (2013), thresholding enhances the power to detect the alternative hypotheses by including all relatively large standardized differences M_{ij}. Those comparisons imply two forces of thresholding that increase the signal to noise ratio for hypotheses testing: increase the signals under alternative hypotheses while simultaneous controls the noise variation.

The proposed thresholding statistics for covariances is in the same spirit as the thresholding tests for means (Fan, 1996; Donoho and Jin, 2004; Hall and Jin, 2010) and regression coefficients (Qiu et al., 2018). As the entries of a sample covariance matrix can be stacked together as a long vector, thresholding tests for means could be extended for covariances. However, the results of the means tests cannot be readily applied to covariance testing due to the circular dependence in sample covariances, which makes the analysis of thresholding statistics on sample covariances more challenging. Specifically, the dependence among entries of a sample covariance matrix does not follow a direction as any entries in a row or column of the sample covariance matrix are dependent due to sharing common data component. We propose a novel matrix blocking method in couple with the martingale central limit theorem (Hall and Heyde, 1980) to overcome the challenges for covariance testing and derive the asymptotic distribution of $T_n(s)$, which is detailed in the proof of Theorem 1 in the SM.

Let $\mu_{T_n,0}(s)$ and $\sigma^2_{T_n,0}(s)$ be the mean and variance of the thresholding statistic $T_n(s)$, respectively under H_0. Let $\phi(\cdot)$ and $\bar{\Phi}(\cdot)$ be the density and survival functions of $N(0,1)$, respectively. Recall that $q=p(p+1)/2$. Proposition 1 and Theorem 1 in Section 5 shows that under H_0 of (2.1), Assumptions 2–4, 5A or 5B and either (i) the exponential growth of p (Assumption 1A) with $s>1/2$ or (ii) the polynomial growth of p that $n\sim p^{\xi}$ for $\xi\in(0,2]$ (Assumption 1B) with $s>1/2-\xi/4$,

$$\sigma_{T_n,0}^{-1}(s)\{T_n(s)-\mu_{T_n,0}(s)\}\xrightarrow{d} N(0,1) \quad \text{as } n,p\to\infty,$$

where $\mu_{T_n,0}(s)=\tilde{\mu}_{T_n,0}(s)[1+O\{\lambda_p^{3/2}(s)n^{-1/2}\}]$, $\sigma^2_{T_n,0}(s)=\tilde{\sigma}^2_{T_n,0}(s)\{1+o(1)\}$,

$$\tilde{\mu}_{T_n,0}(s) = q[2\lambda_p^{1/2}(s)\phi\{\lambda_p^{1/2}(s)\}+2\bar{\Phi}\{\lambda_p^{1/2}(s)\}] \quad \textit{and} \tag{3.2}$$

$$\tilde{\sigma}^2_{T_n,0}(s) = q[2\{\lambda_p^{3/2}(s)+3\lambda_p^{1/2}(s)\}\phi\{\lambda_p^{1/2}(s)\}+6\bar{\Phi}\{\lambda_p^{1/2}(s)\}]. \tag{3.3}$$

It is noted from the above that $\tilde{\sigma}^2_{T_n,0}(s)/\sigma^2_{T_n,0}(s)\to 1$. Let $\hat{\mu}_{T_n,0}(s)$ be an estimate of $\mu_{T_n,0}(s)$ that satisfies

$$\hat{\mu}_{T_n,0}(s)-\mu_{T_n,0}(s)=o_p\{\tilde{\sigma}_{T_n,0}(s)\}. \tag{3.4}$$

By Slutsky's theorem, under (3.4), we have

$$\tilde{\sigma}_{T_n,0}^{-1}(s)\{T_n(s)-\hat{\mu}_{T_n,0}(s)\}\xrightarrow{d} N(0,1) \text{ as } n,p\to\infty.$$

A natural choice of $\hat{\mu}_{T_n,0}(s)$ is the main order term $\tilde{\mu}_{T_n,0}(s)$ given in (3.2). According to the expansion of $\mu_{T_n,0}(s)$ in Proposition 1,

$$\frac{\mu_{T_n,0}(s)-\tilde{\mu}_{T_n,0}(s)}{\tilde{\sigma}_{T_n,0}(s)} = O_p\{\lambda_p^{5/4}(s)p^{1-s}n^{-1/2}\}, \tag{3.5}$$

which converges to zero under Assumption 1B and $s > 1 - \xi/2$. Let z_α be the upper α quantile of $N(0, 1)$. We reject the null hypothesis of (2.1) if

$$T_n(s) > \tilde{\mu}_{T_n,0}(s) + z_\alpha \tilde{\sigma}_{T_n,0}(s), \tag{3.6}$$

which is called the single level thresholding test as it depends on a single s.

In the followings, we assume Condition (3.4) is satisfied to simplify the analysis. When estimators satisfying (3.4) are not available, we may choose $\hat{\mu}_{T_n,0}(s) = \tilde{\mu}_{T_n,0}(s)$ while the lower threshold bound has to be chosen as $1 - \xi/2$ to make (3.5) converge to 0.

More accurate estimator of $\mu_{T_n,0}(s)$ can be constructed by establishing expansions for $\mu_{T_n,0}(s)$ and then correcting for the bias empirically. Delaigle et al. (2011) found that more precise moderate deviation results can be derived for the bootstrap calibrated t-statistics, which provides more accurate estimator for the mean.

To enhance the power of the single level thresholding test, we consider thresholding with multiple threshold levels. Specifically, we take the maximum of the standardized statistic of $T_n(s)$ as the test statistic. Let $\mathcal{T}_n(s) = \tilde{\sigma}^{-1}_{T_n,0}(s)\{T_n(s) - \hat{\mu}_{T_n,0}(s)\}$ be the standardization of $T_n(s)$, and

$$\mathcal{S}_n(s_0) = \{s_{ij} : s_{ij} = M_{ij}/\{4\log(p)\} \text{ and } s_0 < s_{ij} \le (1 - \eta)\} \tag{3.7}$$

be a set of threshold levels for a threshold lower bound s_0 and an arbitrarily small positive constant η. The multi-level thresholding statistic is constructed as

$$\mathcal{V}_n(s_0) = \sup_{s \in \mathcal{S}_n(s_0)} \mathcal{T}_n(s), \tag{3.8}$$

which is a similar formulation as the Higher Criticism (HC) statistic (Donoho and Jin, 2004) and an ℓ_2-variant (Zhong et al., 2013) for detecting rare and faint signals in means. From Theorem 1, a choice of s_0 is either $1/2$ or $1/2 - \xi/4$ depending on p having the exponential or polynomial growth with respect to n, where ξ is the polynomial growth rate such that $n \sim p^\xi$.

From Theorem 2 in Section 5, an asymptotic α-level multi-thresholding test (MTT) rejects the null hypothesis of (2.1) if

$$\mathcal{V}_n(s_0) > [q_\alpha + b\{\log(p), s_0, \eta\}]/a\{\log(p)\}, \tag{3.9}$$

where $a(y) = \{2\log(y)\}^{1/2}$, $b(y, s_0, \eta) = 2\log(y) + 2^{-1}\log\log(y) - 2^{-1}\log(\pi) + \log(1 - s_0 - \eta)$, and q_α is the upper α quantile of the Gumbel distribution. This proposed test is powerful against sparse and weak signals as revealed in Section 6.

Since the convergence of $\mathcal{V}_n(s_0)$ to the Gumbel distribution can be slow when the sample size was small, we employed a bootstrap procedure using a consistent covariance estimator, denoted as $\widehat{\mathbf{\Sigma}}$, proposed by Rothman (2012), which ensures the positive definiteness of the estimated covariance. Since $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ under the null hypothesis, the two samples $\{\mathbf{X}_k\}_{k=1}^{n_1}$ and $\{\mathbf{Y}_k\}_{k=1}^{n_2}$ were pooled together to estimate $\mathbf{\Sigma}_1$. For the bth bootstrap resample, we drew n_1 samples of $\mathbf{X}^*$ and n_2 samples of $\mathbf{Y}^*$ independently from $N(0, \widehat{\mathbf{\Sigma}})$. Then, the bootstrap test statistic $\mathcal{V}_n^{*(b)}(s_0)$ was obtained based on $\mathbf{X}^*$ and $\mathbf{Y}^*$. This procedure was repeated $B = 500$ times to obtain $\{\mathcal{V}_n^{*(1)}(s_0), \ldots, \mathcal{V}_n^{*(B)}(s_0)\}$ under the null hypothesis. The bootstrap empirical null distribution of the proposed statistic was $\widehat{F}_0(x) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{\mathcal{V}_n^{*(b)}(s_0) \le x\}$ and the bootstrap p-value was $1 - \widehat{F}_0\{\mathcal{V}_n(s_0)\}$, where $\mathcal{V}_n(s_0)$ was the multi-thresholding statistic from the original samples. We reject the null hypothesis if the p-value is less than the nominal significant level $\alpha = 0.05$. The bootstrap approximation may be justified as follows. First of all, the regularized estimator $\widehat{\mathbf{\Sigma}}$ of the covariance $\mathbf{\Sigma}$ is able to retain the large covariance entries and set all small estimates to zero with probability converging to 1. Second, as the normal random variables are independent if their covariance is zero, the generated parametric bootstrap samples from the normal distribution satisfy Assumptions 2–4 and 5A or 5B in Section 5 for large n. By Theorems 1 and 2, the bootstrap version of the single and multi-thresholding statistics have the same limiting Gaussian distribution and the extreme value distribution, as the proposed statistics using the original data, respectively. The bootstrap approximation would offer more accurate approximation to the distribution of the test statistics of the MTT than the limiting Gumbel distribution.

4. Identify differentially correlated pairs

If the null hypothesis in (2.1) is rejected, we aim to recover the pairs of variables with different covariances between the two populations. This is equivalent to consider the multiple testing hypotheses

$$H_{0,ij} : \sigma_{ij1} = \sigma_{ij2} \quad \text{vs.} \quad H_{a,ij} : \sigma_{ij1} \neq \sigma_{ij2} \tag{4.1}$$

for $1 \le i \le j \le p$. Let m_0 be the number of true null hypotheses, and $m_a = q - m_0$ be the number of covariances with different values between the two populations. For testing the multiple hypotheses (4.1), let Rej be the number of null hypotheses that are rejected, and FP and FN be the numbers of false positives (Type I errors) and false negatives (Type II errors), respectively. The false discovery proportion defined as $\mathsf{FDP} = \mathsf{FP}/\max\{\mathsf{Rej}, 1\}$ is the proportion of falsely rejected null hypotheses among all rejections, and the false discovery rate $\mathsf{FDR} = \mathbb{E}(\mathsf{FDP})$ is the expectation of the FDP.

There are well known multiple testing procedures for FDR control (Benjamini and Hochberg, 1995; Storey, 2002). Particularly, Cai et al. (2013) proposed a test procedure for (4.1) based on the standardized differences $\{F_{ij}\}$ (the signed

root of M_{ij}) that controls the FDR. However, as the FDR is the expected FDP over repeated experiments, controlling FDR at a level α does not imply controlling FDP at α for the observed data of an experiment. As an alternative, Genovese and Wasserman (2006) considered the FDP exceedance rate which takes account of the variation in the FDP and is a more stringent measure for controlling the Type I errors of the multiple testing than the FDR. They proposed a step-down procedure based on a maximum statistic to control $\mathbb{P}(\text{FDP} > c) \le \alpha$ for a given $c \in (0, 1)$. The role of c is analogous to that of the significance level in a testing problem. A smaller c generates a more conservative multiple testing procedure while a larger c leads to more liberal results. Typical candidate values of c could be 0.1 or 0.2 (Genovese and Wasserman, 2006). As shown in Section 6, the MTT in (3.9) is more powerful than the maximum type tests in detecting sparse and weak signals. We develop a step-down procedure based on the MTT for hypotheses (4.1) that is powerful in identifying sparse and small differences between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.

Let $(i_1^*, j_1^*), (i_2^*, j_2^*), \ldots, (i_q^*, j_q^*)$ be the indexes corresponding to the ordered standardized differences $M_{i_1^* j_1^*} \ge M_{i_2^* j_2^*} \ge \cdots \ge M_{i_q^* j_q^*}$ where $q = p(p+1)/2$ denotes the total number of covariances. Let $\mathcal{D}_l = \{(i_l^*, j_l^*), (i_{l+1}^*, j_{l+1}^*), \ldots, (i_q^*, j_q^*)\}$ for $l = 1, 2, \ldots, q$. To detect signals among the pairs in $\mathcal{D}_l$, we device a step-down procedure that applies the proposed MTT for successive hypotheses

$$H_{\mathcal{D}_l,0} : \sigma_{ij1} = \sigma_{ij2} \text{ for all } (i,j) \in \mathcal{D}_l \text{ vs. } H_{\mathcal{D}_l,a} : \sigma_{ij1} \ne \sigma_{ij2} \text{ for some } (i,j) \in \mathcal{D}_l. \tag{4.2}$$

We start with $l = 1$ which corresponds to hypotheses (2.1). If the null hypothesis of (4.2) is rejected, we remove the pair (i_l^*, j_l^*) corresponding to the largest M_{ij} in $\mathcal{D}_l$, and repeat the MTT for $H_{\mathcal{D}_{l+1},0}$. Let $H_{\mathcal{D}_{l_s},0}$ be the first null that is not rejected, and $\mathcal{R}_s = \{(i_1^*, j_1^*), \ldots, (i_{l_s-1}^*, j_{l_s-1}^*)\}$ be the rejection set of pairs identified by the step-down procedure if $l_s > 1$. To enhance the power of the procedure, we expand $\mathcal{R}_s$ by including the next $\lfloor (l_s-1)c/(1-c) \rfloor$ pairs after $(i_{l_s-1}^*, j_{l_s-1}^*)$ in $\mathcal{D}_1$ for a small constant $c \in (0, 1)$ whenever $l_s > 1$, leading to using the first $\lfloor (l_s-1)/(1-c) \rfloor$ pairs in $\mathcal{D}_1$ as the identified signal set over the covariances between the two populations.

The sequence of the MTT for (4.2) serve as a step-down procedure for the multiple testing problem (4.1). From Theorem 2, if there is no signal in $\mathcal{D}_l$, the MTT is able to control the size for testing the hypotheses (4.2) so that the probability of rejecting $H_{\mathcal{D}_l,0}$ is less than α. This implies $1-\alpha$ confidence in controlling the FDP. The rationale for enlarging the rejection set of the step-down procedure is that if we only choose $\mathcal{R}_s$ as the set of identified signals, the FDP would diminish to zero with probability $1-\alpha$. By additionally the top $\lfloor (l_s-1)c/(1-c) \rfloor$ pairs in $\mathcal{D}_{l_s}$, the power of the multiple testing procedure is increased while the rate of FDP exceeding c can still be controlled at α level asymptotically.

The proposed testing procedure can be also applied to identify sub-groups of variables with different covariance structure between the two populations. Let $\tilde{\mathcal{G}}_h = \{j_1, \ldots, j_{p_h}\}$ be the index set of the hth sub-group of variables with size p_h and $\mathcal{G}_h = \tilde{\mathcal{G}}_h \times \tilde{\mathcal{G}}_h$ for $h = 1, \ldots, g_0$. Let $\mathbf{\Sigma}_{1,h}$ and $\mathbf{\Sigma}_{2,h}$ be the covariances of the hth sub-group, which are the sub-matrices of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ with elements in $\mathcal{G}_h$. We are interested in identifying the sub-groups with different covariance matrices, namely,

$$H_{0,h} : \mathbf{\Sigma}_{1,h} = \mathbf{\Sigma}_{2,h} \text{ vs. } H_{a,h} : \mathbf{\Sigma}_{1,h} \ne \mathbf{\Sigma}_{2,h} \tag{4.3}$$

for $h = 1, \ldots, g_0$.

Let $T_{\mathcal{G}_h}(s) = \sum_{(i,j)\in\mathcal{G}_h, i\le j} M_{ij}\mathbb{I}\{M_{ij} > \lambda_p(s)\}$ be the thresholding statistic on $\mathcal{G}_h$, and $\mathcal{V}_{\mathcal{G}_h}(s_0)$ be the corresponding multi-level thresholding statistic on $\mathcal{G}_h$. Let $\{h_1^*, h_2^*, \ldots, h_{g_0}^*\}$ be the indexes for the ordered $\{\mathcal{V}_{\mathcal{G}_h}(s_0)\}$ such that $\mathcal{V}_{\mathcal{G}_{h_1^*}}(s_0) \ge \mathcal{V}_{\mathcal{G}_{h_2^*}}(s_0) \ge \cdots \ge \mathcal{V}_{\mathcal{G}_{h_{g_0}^*}}(s_0)$. Let $\mathcal{H}_l = \mathcal{G}_{h_l^*} \cup \cdots \cup \mathcal{G}_{h_{g_0}^*}$ for $l = 1, \ldots, g_0$. Similar to (4.2), we consider the MTT for the overall hypotheses $H_{\mathcal{H}_l,0} : \sigma_{ij1} = \sigma_{ij2}$ for all $(i,j) \in \mathcal{H}_l$ in a step-down manner. Following the construction of the multiple testing procedure for the hypotheses (4.1), the proposed testing procedure for (4.3) is constituted by a step-down process based on the multi-level thresholding statistic $\mathcal{V}_{\mathcal{H}_l}(s_0)$ and an augmentation step.

5. Assumptions and asymptotic distributions

Let C be a positive constant whose value may change in the context. For two real sequences $\{a_n\}$ and $\{b_n\}$, $a_n \sim b_n$ means that there are two positive constants c_1 and c_2 such that $c_1 \le a_n/b_n \le c_2$ for all n. For two σ-fields $\mathcal{A}$ and $\mathcal{B}$, let

$$\zeta(\mathcal{A}, \mathcal{B}) = \frac{1}{2} \sup \sum_{l_1=1}^{v_1} \sum_{l_2=1}^{v_2} \left| \mathbb{P}(A_{l_1} \cap B_{l_2}) - \mathbb{P}(A_{l_1})\mathbb{P}(B_{l_2}) \right|$$

be the β-mixing coefficient (Bradley, 2005) between $\mathcal{A}$ and $\mathcal{B}$, where the supremum is taken over all finite partitions $\{A_{l_1} \in \mathcal{A}\}_{l_1=1}^{v_1}$ and $\{B_{l_2} \in \mathcal{B}\}_{l_2=1}^{v_2}$ of the sample space, and $v_1, v_2 \in \mathbb{Z}^+$, the set of positive integers. Without loss of generality, we assume $\mathbb{E}(\mathbf{X}_1) = \mathbb{E}(\mathbf{Y}_1) = \mathbf{0}$. We make the following assumptions in our analysis.

Assumption 1A. As $n \to \infty$, $p \to \infty$, $\log p \sim n^{\varpi}$ for a $\varpi \in (0, 1/5)$.

Assumption 1B. As $n \to \infty$, $p \to \infty$, $n \sim p^{\xi}$ for a $\xi \in (0, 2]$.

Assumption 2. There exists a positive constant τ such that

$$\tau < \min_{1\le i\le p}\{\sigma_{ii1},\sigma_{ii2}\} \le \max_{1\le i\le p}\{\sigma_{ii1},\sigma_{ii2}\} < \tau^{-1} \text{ and} \tag{5.1}$$

$$\min_{i,j}\{\theta_{ij1}/(\sigma_{ii1}\sigma_{jj1}),\theta_{ij2}/(\sigma_{ii2}\sigma_{jj2})\} > \tau. \tag{5.2}$$

Assumption 3. There exist positive constants η and C such that for all $|t| < \eta$,

$$\mathbb{E}\{\exp(tX_{ki}^2)\} \le C \quad \text{and} \quad \mathbb{E}\{\exp(tY_{ki}^2)\} \le C \quad \text{for} \quad i = 1,\ldots,p.$$

Assumption 4. There exists a small positive constant ρ_0 such that

$$\max\{|\rho_{ij1}|,|\rho_{ij2}|\} < 1-\rho_0 \quad \text{for any } i \ne j, \tag{5.3}$$

and $\max\{|\rho^{(1)}_{ij,lm}|,|\rho^{(2)}_{ij,lm}|\} < 1-\rho_0$ for any $(i,j) \ne (l,m)$.

Assumption 5A. There is a permutation $\pi_{*,p}$ of the data sequences $\{X_{kj}\}_{j=1}^p$ and $\{Y_{kj}\}_{j=1}^p$ such that the permuted sequences $\mathbf{X}_k(\pi_{*,p})$ and $\mathbf{Y}_k(\pi_{*,p})$ are β-mixing with the mixing coefficients $\zeta_{x,p}(h)$, $\zeta_{y,p}(h) \le C\gamma^h$ for a constant $\gamma \in (0,1)$, any $p \in \mathbb{Z}^+$ and positive integers $h \le p-1$, where $\zeta_{x,p}(h) = \sup_{1\le m\le p-h}\zeta\{\mathcal{F}_1^m(\mathbf{X}_k(\pi_{*,p})),\mathcal{F}_{m+h}^p(\mathbf{X}_k(\pi_{*,p}))\}$, $\mathcal{F}_{m_a}^{m_b}(\mathbf{X}_k(\pi_{*,p})) = \sigma\{X_{kj}(\pi_{*,p}) : m_a \le j \le m_b\}$ for $1 \le m_a \le m_b \le p$, and $\zeta_{y,p}(h)$ is similarly defined.

Assumptions 1A and 1B specify the exponential and polynomial growth rates of p relative to n, respectively. Assumption 2 prescribes that θ_{ij1} and θ_{ij2} are bounded away from zero to ensure the denominators of M_{ij} being bounded away from zero with probability approaching 1. Assumption 3 assumes the distributions of X_{ki} and Y_{ki} are sub-Gaussian. Sub-Gaussianity is commonly assumed in high-dimensional literature (Bickel and Levina, 2008; Xue et al., 2012; Cai et al., 2013), which contains the Gaussian distribution and distributions with bounded support as special cases. This assumption is a sufficient condition for $X_{ki}X_{kj}$ being sub-exponential distributed, which is needed to derive the asymptotic expansion of the large deviation probability $\mathbb{P}(F_{ij} > z)/\bar{\Phi}(z)$, namely, the Cramér type large deviation result (Petrov, 1995). Although concentration inequalities of $\mathbb{P}(F_{ij} > z)$ hold under weaker conditions on Orlicz norm of X_{ki}, those conditions by Orlicz norm are not sufficient to derive the asymptotic expansion of $\mathbb{P}(F_{ij} > z)$ and the ratio-consistence of $\mathbb{P}(F_{ij} > z)$ to the standard normal tail probability $\bar{\Phi}(z)$, which is needed for the asymptotic properties of the proposed statistic. However, this condition can be weakened to the Linnik condition or the Statulevičius condition (Saulis and Statulevičius, 1991) on $X_{ki}X_{kj}$; see the discussion after Proposition 1. Assumption 4 regulates the correlations among variables in $\mathbf{X}_k$ and $\mathbf{Y}_k$, and subsequently the correlations among $\{F_{ij}\}$ where $M_{ij} = F_{ij}^2$. Assumption 5A controls the dependence among sample covariances so that $\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2}$ and $\hat{\sigma}_{lm1} - \hat{\sigma}_{lm2}$ are nearly independent if the four variables are far apart from each other. This condition implies that the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are bandable, which include banded or block diagonal matrices with relatively small block sizes, after certain permutation of variables. The β-mixing coefficients are assumed exponentially decaying for an unknown variable permutation $\pi_{*,p}$. However, there is no need to know this permutation to perform the proposed test in (3.9). The exponential rate is to simplify proofs, while the theoretical results in this section still hold under an arithmetic rate at the expense of more technical details. As special cases, Theorem 3.3 in Bradley (2005) and Mokkadem (1988) showed that both Markov chains and linear processes with i.i.d. innovations are β-mixing under some weak conditions. Meanwhile, normally distributed data with banded or block diagonal covariance after certain variable permutation satisfy this assumption as well. Similar mixing conditions for dependence among variables were also made in Delaigle et al. (2011), Zhong et al. (2013), Xu et al. (2016) and He et al. (2021).

For spatially dependent data, Assumption 5A can be replaced by the following.

Assumption 5B. The jth variables X_{kj} and Y_{kj} are associated with a location index u_j for $j = 1,\ldots,p$. Let $d(u_i,u_j)$ be the distance between u_i and u_j. Assume $d(u_i,u_j) > d_0$ for a positive constant d_0 for all $i,j = 1,\ldots,p$ and $i \ne j$. For two location index sets A_1 and A_2, let $d(A_1,A_2) = \min\{d(u_i,u_j) : u_i \in A_1, u_j \in A_2\}$ be the distance between A_1 and A_2, and $\mathcal{F}(\mathbf{X}_k,A_1) = \sigma\{X_{kj} : u_j \in A_1\}$ be the σ-field generated by the data contained in A_1. The sequences $\{X_{kj}\}_{j=1}^p$ and $\{Y_{kj}\}_{j=1}^p$ are spatially β-mixing with the mixing coefficients $\zeta_{x,p}(h)$, $\zeta_{y,p}(h) \le C\gamma^h$ for a $\gamma \in (0,1)$ and $h > 0$, where $\zeta_{x,p}(h) = \sup_{d(A_1,A_2)>h}\zeta\{\mathcal{F}(\mathbf{X}_k,A_1),\mathcal{F}(\mathbf{X}_k,A_2)\}$, and $\zeta_{y,p}(h)$ is similarly defined.

Assumption 5B regulates the dependence among spatial data. Spatial mixing conditions are often employed in spatial literature (Jenish and Prucha, 2009; Gaetan and Guyon, 2010). The condition $\min_{i,j}\{d(u_i,u_j)\} > d_0$ implies expanding domain asymptotics for the spatial data. Assumptions 5A and 5B are used to control the dependence among sample covariances for different type of data, as motivated in Examples 1 and 2 of Section 2. The theoretical properties of the proposed statistics are derived for both types of mixing assumptions to broaden the application of the proposed tests.

Recall that $\mu_{T_n,0}(s) = \mathbb{E}\{T_n(s)|H_0\}$, $\sigma^2_{T_n,0}(s) = \mathrm{Var}\{T_n(s)|H_0\}$, $q = p(p+1)/2$, and $\phi(\cdot)$ and $\bar{\Phi}(\cdot)$ are the density and survival functions of $N(0,1)$, respectively. The following proposition provides expansions of $\mu_{T_n,0}(s)$ and $\sigma^2_{T_n,0}(s)$.

Proposition 1. *Under Assumptions 1A or 1B, 2 – 4 and 5A or 5B, $\mu_{T_n,0}(s) = \tilde{\mu}_{T_n,0}(s)[1 + O\{\lambda_p^{3/2}(s)n^{-1/2}\}]$ where*

$$\tilde{\mu}_{T_n,0}(s) = q[2\lambda_p^{1/2}(s)\phi\{\lambda_p^{1/2}(s)\} + 2\bar{\Phi}\{\lambda_p^{1/2}(s)\}].$$

In addition, under either (i) Assumption 1A with $s > 1/2$ or (ii) Assumption 1B with $s > 1/2 - \xi/4$, $\sigma^2_{T_n,0}(s) = \tilde{\sigma}^2_{T_n,0}(s)\{1+o(1)\}$, where $\tilde{\sigma}^2_{T_n,0}(s) = q[2\{\lambda_p^{3/2}(s) + 3\lambda_p^{1/2}(s)\}\phi\{\lambda_p^{1/2}(s)\} + 6\bar{\Phi}\{\lambda_p^{1/2}(s)\}]$.

The derivation of $\mu_{T_n,0}(s)$ and $\sigma^2_{T_n,0}(s)$ in Proposition 1 mainly uses the Cramér large deviation result $\mathbb{P}(F_{ij} > z)/\bar{\Phi}(z) = 1 + o(1)$ for $z = o(n^{1/6})$ under sub-exponential distribution of $X_{ki}X_{kj}$ and the null hypothesis of (2.1). Nevertheless, from Lemma 2.3 and Theorem 6.4 in Saulis and Statulevičius (1991), the sub-exponential distribution condition can be relaxed to the Statulevičius condition $\Gamma_m(X_{ki}X_{kj}) \le (m!)^{1+\gamma_1}/\Delta_1^{m-2}$ for constants $\gamma_1 \ge 0$, $\Delta_1 > 0$ and $m = 3, 4, \ldots$, or the Linnik condition $\mathbb{E}\exp(|X_{ki}X_{kj}|^{\gamma_2}) < C$ for a constant $\gamma_2 \in (0, 1)$, where $\Gamma_m(X_{ki}X_{kj})$ denotes the cumulant of $X_{ki}X_{kj}$ of order m. But, the range of the large deviation expansion might be more restrictive comparing to $z = o(n^{1/6})$ under the sub-exponential condition of $X_{ki}X_{kj}$, which would restrict the relationship between $\log p$ and n. Note that the sub-exponential condition of $X_{ki}X_{kj}$ is a special case of the Statulevičius condition with $\gamma_1 = 0$. Therefore, the sub-Gaussian condition of X_{ki} in Assumption 3 can be weakened to $\mathbb{E}\exp(c|X_{ki}|^{\gamma_0}) < \infty$ for positive constants c and $\gamma_0 \in (0, 2]$, which implies $\mathbb{E}\exp(c|X_{ki}X_{kj}|^{\gamma_0/2}) < \infty$. Note that $\gamma_0 = 1$ corresponds to the sub-exponential distribution assumption of X_{ki}. Amosova (2002) also showed that the Statulevičius condition (or the Linnik condition with a different expansion range) is sufficient and necessary for the Cramér large deviation result. To simplify the derivation and better illustrate the theoretical techniques for analyzing the thresholding statistic, we make the sub-Gaussian condition on X_{ki} in Assumption 3 to substitute the Statulevičius and Linnik conditions on $X_{ki}X_{kj}$.

The next theorem derives the asymptotic distribution of $T_n(s)$ at a given s. The main challenge is to deal with the circular dependence among $\{M_{ij}\}$. Although Stein's method (Stein, 1972) can be applied to establish distributional approximation for sum of certain dependent sequences, it cannot be directly applied to our case, as each M_{ij} is dependent with more than p other summands in $T_n(s)$ so that the local dependence structure for Stein's method does not hold; see Theorem 3.6 in Ross (2011). To tackle the challenge that $\{M_{ij}\}$ in the same row and the same column are dependent, we propose a novel matrix blocking technique with Berbee's coupling theorem (Athreya and Lahiri, 2006, page 516) to represent $T_n(s)$ by a U-statistic formulation constructed on independent blocks of variables. The martingale central limit theorem is applied on the U-statistic to show the asymptotic normality of $T_n(s)$. This technique and the proofs are given in the SM.

Theorem 1. *Suppose Assumptions 2–4 and 5A or 5B are satisfied. Then, under the H_0 of (2.1), and either (i) Assumption 1A with $s > 1/2$ or (ii) Assumption 1B with $s > 1/2 - \xi/4$, we have*

$$\sigma^{-1}_{T_n,0}(s)\{T_n(s) - \mu_{T_n,0}(s)\} \xrightarrow{d} N(0, 1) \quad \text{as } n, p \to \infty.$$

It is shown in the SM that the correlation between $M_{ij}\mathbb{I}\{M_{ij} > \lambda_p(s)\}$ and $M_{il}\mathbb{I}\{M_{il} > \lambda_p(s)\}$ with a common variable i decreases with the increase of the threshold level s. Compared with thresholding sample means, the restriction on s in Theorem 1 is to control the dependence among the thresholded sample covariances in $T_n(s)$.

Recall that $a(y) = \{2\log(y)\}^{1/2}$ and $b(y, s_0, \eta) = 2\log(y) + 2^{-1}\log\log(y) - 2^{-1}\log(\pi) + \log(1 - s_0 - \eta)$. The asymptotic distribution of the multi-thresholding statistic $\mathcal{V}_n(s_0)$ is given in the following theorem.

Theorem 2. *Suppose conditions of Theorem 1 and (3.4) hold, under H_0 of (2.1),*

$$\mathbb{P}[a\{\log(p)\}\mathcal{V}_n(s_0) - b\{\log(p), s_0, \eta\} \le x] \to \exp(-\exp(-x)).$$

6. Power analysis

This section evaluates the power of the proposed MTT in (3.9) under the alternative hypotheses of (2.1) with few non-zero δ_{ij} and the values of those nonzero δ_{ij} being faint, where $\delta_{ij} = \sigma_{ij1} - \sigma_{ij2}$. Let m_a denote the number of nonzero δ_{ij} for $i \le j$, and $\lfloor\cdot\rfloor$ be the integer truncation function. We consider the covariance class with sparse and weak differences

$$\begin{aligned}\mathcal{C}(\beta, \{r_{0,ij}\}) = \Big\{ &(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) : m_a = \lfloor q^{1-\beta} \rfloor \text{ nonzero } \delta_{ij} \text{ with signal strength} \\ &\delta_{ij} = \{2r_{0,ij}\log(q)/n\}^{1/2} = \{4r_{0,ij}\log(p)/n\}^{1/2}\{1 + o(1)\} \text{ for } \delta_{ij} \neq 0 \Big\},\end{aligned} \tag{6.1}$$

where $\beta \in (1/2, 1)$ is the signal sparsity parameter and $\{r_{0,ij}\}$ are the signal strength parameters. The covariance class $\mathcal{C}(\beta, \{r_{0,ij}\})$ constitutes the most challenging setting for detecting differences between two covariances. Here, as shown in Cai et al. (2013), $\{\log(p)/n\}^{1/2}$ in (6.1) is the minimum rate for successful signal detection under the sparse setting. For testing $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, Li and Chen (2012) proposed an ℓ_2-type test. They showed that their test is powerful if the Frobenius distance $\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_F$ is larger than $c(p/n)^{1/2}$ for a positive constant c. Note that if all the signals have strength larger than $n^{-1/2}$ and the number of signals is larger than p, the Frobenius distance between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ is larger than $(p/n)^{1/2}$, and the ℓ_2-test is powerful. Therefore, we call the case with more than p signals as dense signal regime, which corresponds to $\beta \in (0, 1/2]$. Meanwhile, $\beta \in (1/2, 1)$ is regarded as the sparse signal regime. See Donoho and Jin (2004), Hall and Jin (2010) for similar settings in the context of testing means.

For $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathcal{C}(\beta, \{r_{0,ij}\})$, define the standardized signal strength

$$r_{ij} = r_{0,ij}/\{(1-\kappa)\theta_{ij1} + \kappa\theta_{ij2}\} \text{ for } \sigma_{ij1} \neq \sigma_{ij2}, \tag{6.2}$$

by recognizing that the denominator is the main order term of the variance of $\sqrt{n}(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2})$. Under Gaussian distributions, $\theta_{ij1} = \sigma_{ii1}\sigma_{jj1} + \sigma_{ij1}^2$ and $\theta_{ij2} = \sigma_{ii2}\sigma_{jj2} + \sigma_{ij2}^2$. Since the difference between σ_{ij1} and σ_{ij2} is at the order $\{\log(p)/n\}^{1/2}$ under $\mathcal{C}(\beta, \{r_{0,ij}\})$, we have $r_{ij} = r_{0,ij}/(\sigma_{ii1}\sigma_{jj1} + \sigma_{ij1}^2)(1 + O[\{\log(p)/n\}^{1/2}])$. Define the maximal and minimal standardized signal strength

$$\bar{r} = \max_{(i,j):\sigma_{ij1}\neq\sigma_{ij2}} r_{ij} \text{ and } \underline{r} = \min_{(i,j):\sigma_{ij1}\neq\sigma_{ij2}} r_{ij}. \tag{6.3}$$

For $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathcal{C}(\beta, \{r_{0,ij}\})$, let $\mu_{T_n,1}(s)$ and $\sigma^2_{T_n,1}(s)$ be the mean and variance of $T_n(s)$ under the alternative hypothesis, and let

$$\mathsf{Power}_n(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = \mathbb{P}\big(\mathcal{V}_n(s_0) > [q_\alpha + b\{\log(p), s_0, \eta\}]/a\{\log(p)\} | \mathbf{\Sigma}_1, \mathbf{\Sigma}_2\big)$$

be the power of the MTT in (3.9). Let $\mathsf{SNR}(s) = \{\mu_{T_n,1}(s) - \mu_{T_n,0}(s)\}/\sigma_{T_n,1}(s)$ be the signal to noise ratio of the single level thresholding test in (3.6). Since

$$\mathcal{V}_n(s_0) = \max_{s\in\mathcal{S}_n(s_0)} \frac{\sigma_{T_n,1}(s)}{\tilde{\sigma}_{T_n,0}(s)} \left\{ \frac{T_n(s) - \mu_{T_n,1}(s)}{\sigma_{T_n,1}(s)} - \frac{\hat{\mu}_{T_n,0}(s) - \mu_{T_n,0}(s)}{\sigma_{T_n,1}(s)} + \mathsf{SNR}(s) \right\},$$

the power of the MTT is determined by the maximum of $\mathsf{SNR}(s)$ over $s \in \mathcal{S}_n(s_0)$. This indicates the power enhancement of the MTT over the single level thresholding test, as it utilizes the maximum signal to noise ratio over a range of threshold levels.

Recall that $\xi \in (0, 2]$ gives the growth rate of p with respect to n as $n \sim p^\xi$. To present the power of the MTT, we first introduce a family of detection boundaries indexed by $\xi \in [0, 2]$:

$$\rho^*(\beta, \xi) = \begin{cases} \{(4-2\xi)^{1/2} - (6-8\beta-\xi)^{1/2}\}^2/8, & 1/2 < \beta \le 5/8 - \xi/16, \\ \beta - 1/2, & 5/8 - \xi/16 < \beta \le 3/4, \\ \{1-(1-\beta)^{1/2}\}^2, & 3/4 < \beta < 1. \end{cases} \tag{6.4}$$

The following proposition shows that the power of MTT is determined by the standardized signal strength and the signal sparsity under polynomial growth rate of p in Assumption 1B.

Proposition 2. *Under Assumptions 1B, 2–4, 5A or 5B, (3.4) and the covariance class $\mathcal{C}(\beta, \{r_{0,ij}\})$ in (6.1), for $s_0 = 1/2 - \xi/4$, if $\underline{r} > \rho^*(\beta, \xi)$, the power of the MTT converges to 1 so that $\mathsf{Power}_n(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \to 1$ for a series of nominal sizes $\alpha_n = \bar{\Phi}\{(\log p)^\epsilon\} \to 0$ as $n, p \to \infty$, where ϵ is an arbitrarily small positive constant.*

Proposition 2 shows that the power of the proposed MTT converges to 1 if $\underline{r}$ is above the boundary $\rho^*(\beta, \xi)$. It can be shown that $\rho^*(\beta, \xi)$ is a non-increasing function of ξ for a given $\beta \in (1/2, 5/8)$. As smaller ξ implies larger p, this means a higher detection boundary of MTT for a larger dimension, and stronger signals are needed to guarantee the power of MTT converging to 1 for a higher growth rate of p. Also notice that $\rho^*(\beta, 2)$ (namely, if $n \sim p^2$) is the optimal detection boundary for testing means with uncorrelated Gaussian data (Donoho and Jin, 2004), which corresponds to the lower limit s_0 of the threshold levels being zero in Theorems 1 and 2. Restricting $s \ge s_0 = 1/2 - \xi/4$ controls the sizes of the thresholding tests asymptotically, which slightly reduces the utility of the MTT in detecting weak signals for $1/2 < \beta \le 5/8 - \xi/16$.

The following proposition shows that $\rho^*(\beta, 0)$ is the detection boundary when dimension p grows exponentially fast with n, which can be viewed as a degenerated polynomial growth case with $\xi = 0$.

Proposition 3. *Under Assumptions 1A, 2–4, 5A or 5B, (3.4) and the covariance class $\mathcal{C}(\beta, \{r_{0,ij}\})$ in (6.1), for $s_0 = 1/2$, if $\underline{r} > \rho^*(\beta, 0)$, the power of the MTT converges to 1 so that $\mathsf{Power}_n(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \to 1$ for a series of nominal sizes $\alpha_n = \bar{\Phi}\{(\log p)^\epsilon\} \to 0$ as $n, p \to \infty$, where ϵ is an arbitrarily small positive constant.*

From Cai et al. (2013), the power of the $\ell_{\max}$-test converges to 1 if

$$\max_{1\le i\le j\le p} \frac{|\sigma_{ij1} - \sigma_{ij2}|}{(\theta_{ij1}/n_1 + \theta_{ij2}/n_2)^{1/2}} > 4(\log p)^{1/2},$$

which is equivalent to the standardized signal strength r_{ij} larger than 4 in our context. From the detection boundary in (6.4), the MTT can have power approaching 1 for $r_{ij} \in (0, 1)$. This means the signal strength required by the $\ell_{\max}$-test is stronger than that required by the proposed test. Also, the ℓ_2-test of Li and Chen (2012) does not have non-trivial power for $\beta > 1/2$. Hence, the proposed MTT is more powerful than both the ℓ_2-tests and $\ell_{\max}$-tests in detecting sparse and weak signals.

Several recent works consider to utilize results from different tests to enhance power. In particular, Fan et al. (2015b) and Yu et al. (2021) combined an ℓ_2 type test and a thresholding component with a high threshold level. He et al. (2021)

combined multiple sum-of-qth-power tests for $q \geq 1$ with an $\ell_{\max}$-type test. However, our proposal is different from those existing methods in the following aspects. A key difference is the way to enhance the signal to noise ratio. The proposed approach uses a thresholding procedure to filter out the non-signal components while keeping those with signals. In this way, the signals in the proposed thresholding statistic are retained, but its variance is much reduced. To achieve this, the proposed procedure uses a sequence of thresholds at $\sqrt{4s\log p}$ for $s \in (0, 1)$, and chooses the threshold level s that gives the most significant result. For sparse and weak signals with standardized strength r_{ij} less than 1, the threshold level s needs to be chosen smaller than 1 to retain most of the signals.

In contrast, Fan et al. (2015b), Yu et al. (2021) chose the threshold at $\log(\log n)\sqrt{\log p}$. Note that $\log(\log n)$ diverges to infinity slowly comparing to $\sqrt{4s}$ in our procedure. Using this threshold level cannot detect sparse and weak signals. Moreover, translating the power results in Fan et al. (2015b) to our setting, the power of their test converges to 1 if the maximal standardized signal strength is larger than $9\{\log(\log n)\}^2/2$, which is much stronger than that required by the proposed test. Similar arguments also apply to the test of He et al. (2021) which combined multiple sum-of-qth-power tests for $q \geq 1$ with an $\ell_{\max}$ type test. From Equation (2.11) and Proposition 2.4 in He et al. (2021), the signal to noise ratio of their U-statistic includes the variances of all variables in the denominator, which would suffer power loss for testing sparse signals, and their maximum test requires strong signals to have power approaching 1. Therefore, the tests in Fan et al. (2015b), Yu et al. (2021) and He et al. (2021) are designed for detecting dense signals (due to the component of ℓ_2 or ℓ_q statistics) and strong signals (due to the thresholding component with a high threshold level or an $\ell_{\max}$ statistic). While, our proposed test is designed for detecting signals that are both sparse and weak, which is the most challenging case in signal detection.

Let $\mathcal{R}_a$ be the set of discovered different covariances between the two populations by the proposed multiple testing procedure for (4.1), which includes the rejections $\mathcal{R}_s$ from the step-down process and those from the augmentation step. Let Rej, FP and TP be the size of $\mathcal{R}_a$, and the numbers of false positives and true positives in $\mathcal{R}_a$, respectively, where $\mathsf{TP} = \mathsf{Rej} - \mathsf{FP}$. The following theorem shows that the proposed procedure is able to recover all the different covariances while maintains the FDP exceedance rate controlled.

Theorem 3. *Suppose Assumptions 2–4, 5A or 5B, and (3.4) are satisfied. Under the covariance class $\mathcal{C}(\beta, \{r_{0,ij}\})$ in (6.1) and either (i) Assumption 1A with $s_0 = 1/2$ or (ii) Assumption 1B with $s_0 = 1/2 - \xi/4$, if $\underline{r} > \beta$, the proposed multiple testing procedure for the hypotheses (4.1) controls the* FDP *exceedance rate asymptotically such that $\mathbb{P}(\mathsf{FDP} > c) \leq \alpha$, as $n, p \to \infty$. And, the power of the proposed test converges to 1 such that $\mathsf{TP}/m_a \to 1$ in probability as $n, p \to \infty$.*

Note that the condition $\underline{r} > \beta$ on the standardized signal strength is required for identifying rare and faint signal. For variable selection in high-dimensional linear regression models, Ji and Jin (2012) showed that consistent variable selection is impossible if $\underline{r} < \beta$.

7. Simulation results

We report results from simulation experiments which were designed to evaluate the performances of the proposed two-sample MTT under high dimensionality with sparse and faint signals. We also compared the proposed test with the tests in Srivastava and Yanagihara (2010) (SY), Li and Chen (2012) (LC) and Cai et al. (2013) (CLX).

The two random samples $\{\mathbf{X}_k\}_{k=1}^{n_1}$ and $\{\mathbf{Y}_k\}_{k=1}^{n_2}$ were respectively generated from

$$\mathbf{X}_k = \boldsymbol{\Sigma}_1^{\frac{1}{2}}\mathbf{Z}_{1k} \quad \text{and} \quad \mathbf{Y}_k = \boldsymbol{\Sigma}_2^{\frac{1}{2}}\mathbf{Z}_{2k}, \tag{7.1}$$

where $\{\mathbf{Z}_{1k}\}$ and $\{\mathbf{Z}_{2k}\}$ are i.i.d. random vectors from a common population. We considered two distributions for the innovation vectors $\mathbf{Z}_{1k}$ and $\mathbf{Z}_{2k}$: (i) $N(0, \mathbf{I}_p)$; (ii) Gamma distribution where components of $\mathbf{Z}_{1k}$ and $\mathbf{Z}_{2k}$ were i.i.d. standardized Gamma(4,2) with mean 0 and variance 1. It is noted that the Gamma distribution is not sub-Gaussian. Our inclusion of the Gamma setting is to check on the robustness of the proposed test with respect to the sub-Gaussian assumption. Moreover, as discussed after Proposition 1, the sub-Gaussian assumption can be relaxed to weaker Linnik condition or Statulevičius condition on $X_{ki}X_{kj}$, which would include this Gamma distribution setting as a special case. To design the covariances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, let $\boldsymbol{\Sigma}_1^{(0)} = \mathbf{D}_0^{\frac{1}{2}}\boldsymbol{\Sigma}^{(*)}\mathbf{D}_0^{\frac{1}{2}}$, where $\mathbf{D}_0 = \mathrm{diag}(d_1, \ldots, d_p)$ with elements generated according to the uniform distribution U(0.1, 1), and $\boldsymbol{\Sigma}^{(*)} = (\sigma_{ij}^*)$ was a positive definite correlation matrix. Once generated, $\mathbf{D}_0$ was held fixed throughout the simulation and created heterogeneity for different dimensions of the data. Two designs of $\boldsymbol{\Sigma}^{(*)}$ were considered in the simulation:

$$\text{Design 1: } \sigma_{ij}^* = 0.4^{|i-j|}; \tag{7.2}$$

$$\text{Design 2: } \sigma_{ij}^* = 0.5\mathbb{I}(i = j) + 0.5\mathbb{I}(i, j \in [4k_0 - 3, 4k_0]); \tag{7.3}$$

for $k_0 = 1, \ldots, \lfloor p/4 \rfloor$. Design 1 has an auto-regressive structure and Design 2 is block diagonal of size 4.

To generate scenarios of sparse and weak signals, we chose

$$\boldsymbol{\Sigma}_1^{(\star)} = \boldsymbol{\Sigma}_1^{(0)} + \epsilon_c\mathbf{I}_p \quad \text{and} \quad \boldsymbol{\Sigma}_2^{(\star)} = \boldsymbol{\Sigma}_1^{(0)} + \mathbf{U} + \epsilon_c\mathbf{I}_p, \tag{7.4}$$

Table 1
Empirical sizes for the tests of Srivastava and Yanagihara (2010) (SY), Li and Chen (2012) (LC), Cai et al. (2013) (CLX) and the proposed multi-level thresholding test based on the limiting distribution calibration in (3.9) (MTT) and the bootstrap calibration (MTT-BT) for Designs 1 and 2 under the Gaussian and Gamma distributions with the nominal level of 5%.

p	(n_1, n_2)	SY	LC	CLX	MTT	MTT-BT
		Gaussian Design 1				
175	(60, 60)	0.048	0.058	0.054	0.088	0.058
277	(80, 80)	0.052	0.052	0.058	0.064	0.056
396	(100, 100)	0.042	0.046	0.058	0.064	0.054
530	(120, 120)	0.056	0.048	0.050	0.056	0.046
		Gaussian Design 2				
175	(60, 60)	0.060	0.048	0.052	0.094	0.048
277	(80, 80)	0.040	0.060	0.040	0.064	0.052
396	(100, 100)	0.052	0.042	0.044	0.090	0.048
530	(120, 120)	0.050	0.046	0.044	0.060	0.054
		Gamma Design 1				
175	(60, 60)	0.046	0.060	0.066	0.068	0.046
277	(80, 80)	0.046	0.050	0.044	0.062	0.042
396	(100, 100)	0.046	0.052	0.046	0.046	0.044
530	(120, 120)	0.060	0.056	0.040	0.038	0.040
		Gamma Design 2				
175	(60, 60)	0.070	0.056	0.048	0.066	0.046
277	(80, 80)	0.038	0.058	0.036	0.062	0.048
396	(100, 100)	0.048	0.050	0.042	0.036	0.044
530	(120, 120)	0.054	0.056	0.048	0.046	0.042

where $\mathbf{U} = (u_{kl})_{p\times p}$ is a banded symmetric matrix and ϵ_c is a positive number to guarantee the positive definiteness of $\mathbf{\Sigma}_2^{(\star)}$. Specifically, let $k_0 = \lfloor m_p/p \rfloor$, where $m_p = \lfloor q^{1-\beta}/2 \rfloor$ is the number of distinct pairs with nonzero u_{kl}, $u_{l+k_0+1\,l} = u_{l\,l+k_0+1} = \{4r\log(p)/n\}^{1/2}$ for $l = 1, \ldots, k_1$, $k_1 = m_p - pk_0 + k_0(k_0+1)/2$, and $u_{kl} = \{4r\log(p)/n\}^{1/2}$ for $|k-l| \le k_0$ and $k \ne l$ if $k_0 \ge 1$. Set $\epsilon_c = |\min\{\lambda_{\min}(\mathbf{\Sigma}_1^{(0)} + \mathbf{U}), 0\}| + 0.05$, where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A. Since $\epsilon_c > 0$ and $\lambda_{\min}(\mathbf{\Sigma}_2^{(\star)}) \ge \lambda_{\min}(\mathbf{\Sigma}_1^{(0)} + \mathbf{U}) + \epsilon_c > 0$, both $\mathbf{\Sigma}_1^{(\star)}$ and $\mathbf{\Sigma}_2^{(\star)}$ were positive definite under Designs 1 and 2. Under the null hypothesis, we chose $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_1^{(0)}$ implied under (7.1), while under the alternative hypothesis $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_1^{(\star)}$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_2^{(\star)}$.

The simulated data were generated from (7.1) according to a randomly selected permutation π_p of $\{1, \ldots, p\}$, which was held fixed once generated. To mimic the regime of sparse and faint signals, we considered a set of β and r. We fixed $\beta = 0.6$ and set $r = 0.1, 0.2, \ldots, 1$ to create different signal strengths, and $r = 0.6$ while β was varied from 0.3 to 0.9 to show the impacts of sparsity levels on the tests. The sample sizes (n_1, n_2) were (60, 60), (80, 80), (100, 100) and (120, 120), respectively, and the corresponding dimensions $p = 175, 277, 396$ and 530 according to $p = \lfloor 0.25 n_1^{1.6} \rfloor$. We set $s_0 = 0.5$ according to Theorem 1 and the discussion following (3.8), and η was chosen as 0.05 in (3.7). We chose $\hat{\mu}_{T_n,0}(s) = \tilde{\mu}_{T_n,0}(s)$. The simulation was replicated 500 times for each simulation setting.

Table 1 reports the empirical sizes of the proposed multi-thresholding test based on the asymptotic Gumbel distribution (denoted as MTT) and the bootstrap calibration (MTT-BT), together with three existing methods, at the nominal level 0.05, for the Gaussian and Gamma distributed random vectors, respectively. We observe that the MTT based on the asymptotic distribution exhibited some size distortion when the sample size was small. However, with the increase of the sample size, the sizes of the MTT became closer to the nominal level. At the meantime, the CLX and SY tests also experienced some size distortion under the Gamma scenario in smaller samples. It is observed that the proposed multi-thresholding test with the bootstrap calibration (MTT-BT) performed consistently well under all the scenarios with accurate empirical sizes. Besides, the empirical sizes with Gamma distribution illustrated that the proposed test were robust to the sub-Gaussian assumption. This shows that the bootstrap calibration offered more accurate approximation to the distribution of the test statistic under the null hypothesis.

Fig. 1 displays the empirical powers with respect to different signal strengths r for covariance matrix Designs 1 and 2 with $n_1 = n_2 = 80$ and $p = 277$, and $n_1 = n_2 = 100$ and $p = 396$ for Gaussian distributed data, respectively. Fig. 2 reports the empirical powers under different sparsity (β) levels when the signal strength r was fixed at 0.6. Simulation results on the powers for the Gamma distribution are available in the SM. It is noted that at $\beta = 0.6$, there were only 68 and 90 unequal entries between the upper triangles of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ among a total of $q = 38503$ and 78606 entries for $p = 277$ and 396, respectively. To make the powers comparable for different methods, we adjusted the critical values of the tests by their respective empirical null distributions so that the actual sizes were approximately equal to the nominal level 5%. Due to the size adjustment, we only reported the numerical power results for the MTT-BT.

Fig. 1 reveals that the power of the proposed MTT-BT was the highest under all the scenarios. Although the powers of the other tests improved as the signal strength r was increased, the MTT-BT maintained a lead over the whole range of $r \in [0.1, 1]$. The power advantage of the MTT-BT over the other three tests got larger as the signal strength r increased.

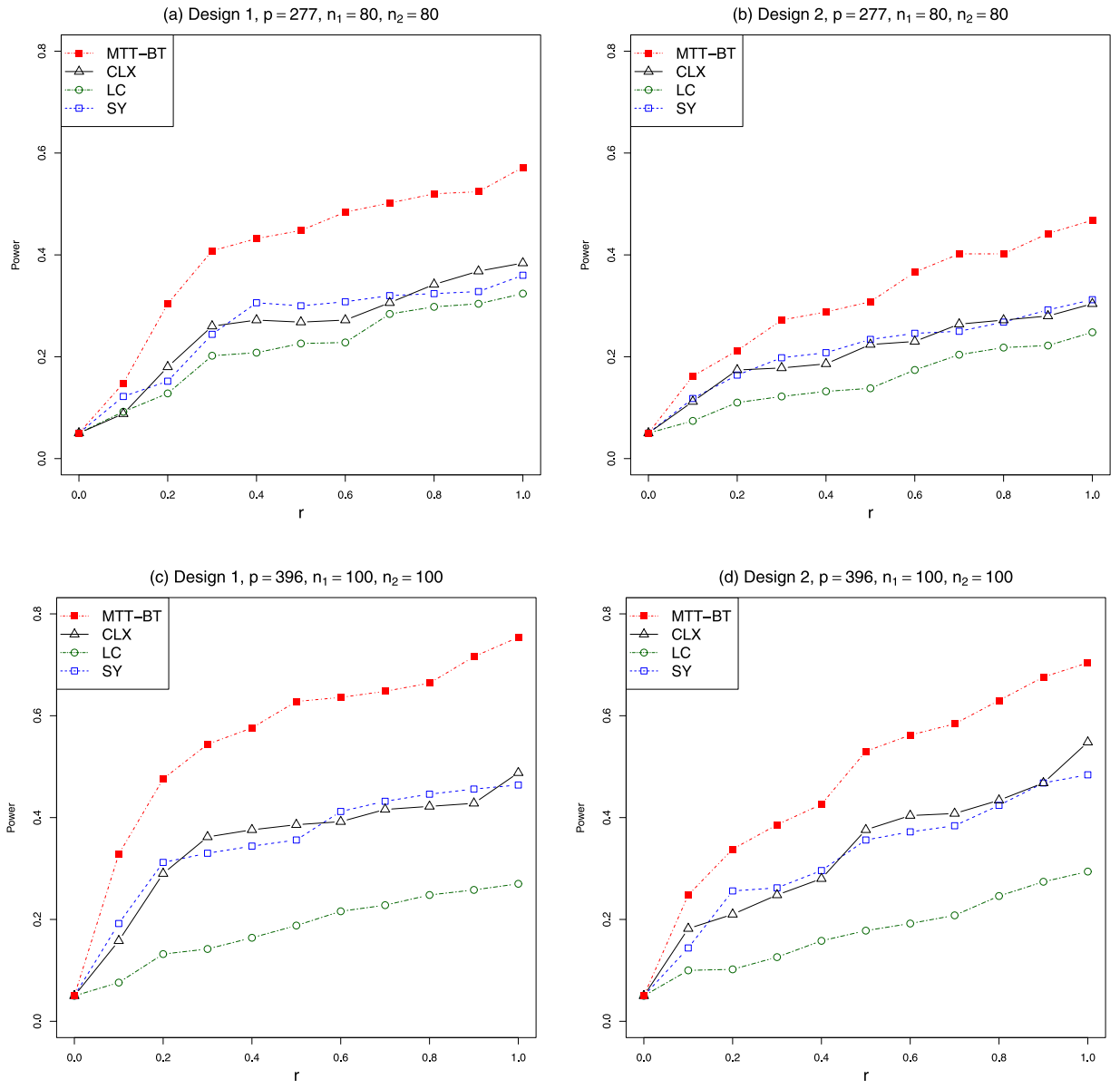

Fig. 1. Empirical powers with respect to the signal strength r for the tests of Srivastava and Yanagihara (2010) (SY), Li and Chen (2012) (LC), Cai et al. (2013) (CLX) and the proposed multi-level thresholding test with the bootstrap calibration (MTT-BT) for Designs 1 and 2 with Gaussian innovations under $\beta = 0.6$ when $p = 277$, $n_1 = n_2 = 80$ and $p = 396$, $n_1 = n_2 = 100$ respectively.

We observe from Fig. 2 that the proposed test also had the highest empirical power across the range of β. The powers of the MTT-BT at the high sparsity level ($\beta \geq 0.7$) were higher than those of the CLX test. The latter test is known for doing well under sparsity. We take this as an empirical confirmation to the attractive detection boundary of the proposed MTT established in the theoretical analysis reported in Section 6. The monotone decrease pattern in the power profile of the four tests reflected the reality of reduction in the number of signals as β was increased. It is noted that the two ℓ_2 norm based tests SY and LC are known to have good powers when the signals are dense ($\beta \leq 0.5$). This was well reflected in Fig. 2 indicating the two had comparable powers to the MTT-BT when $\beta = 0.3$ and 0.4. However, after β was larger than 0.5, both SY and LC's powers started to decline quickly and were surpassed by the CLX and the MTT-BT. This is consistent with the result of Fig. 1 that the ℓ_2-tests without regularization incorporated too many uninformative dimensions and lowered their signal to noise ratios.

To identify the differentially correlated pairs, we evaluated the proposed step-down procedure with augmentation (SDA), denoted as SDA-MTT, for hypotheses (4.1), and compared it with the two support recovery procedures SR and SR-FWER in Cai et al. (2013), where FWER denotes the family-wise error rate that is the probability of making false

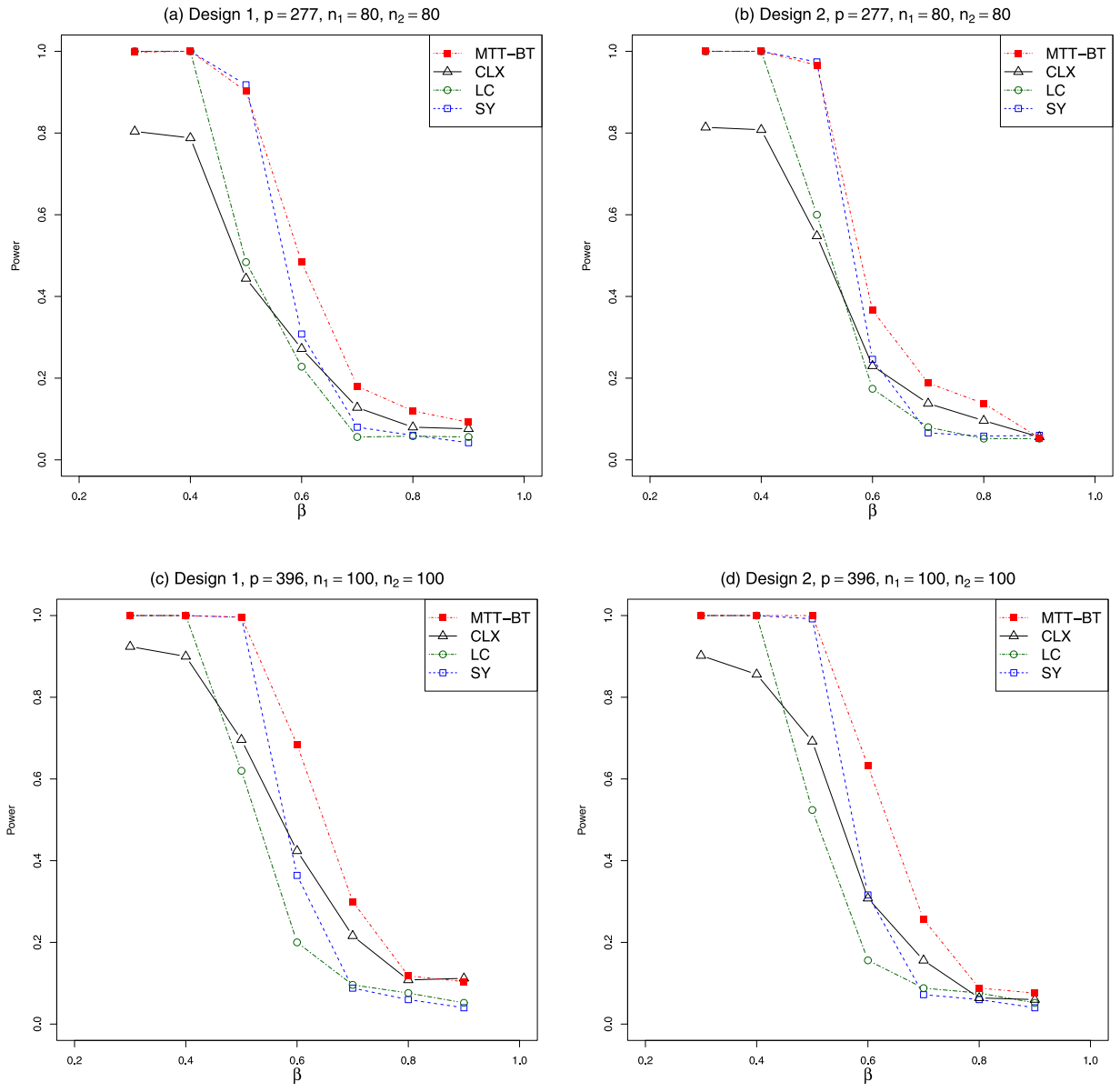

Fig. 2. Empirical powers with respect to the sparsity level β for the tests of Srivastava and Yanagihara (2010) (SY), Li and Chen (2012) (LC), Cai et al. (2013) (CLX) and the proposed multi-level thresholding test with the bootstrap calibration (MTT-BT) for Designs 1 and 2 with Gaussian innovations under $r = 0.6$ when $p = 277$, $n_1 = n_2 = 80$ and $p = 396$, $n_1 = n_2 = 100$ respectively.

positive errors. The SR procedure thresholds M_{ij} by $4 \log p$, and the SR-FWER procedure uses a threshold on M_{ij} that controls the FWER at a given α. We also applied the step-down with augmentation procedure in conjunction with the $\ell_{\max}$-test of Cai et al. (2013), based on the maximum of M_{ij} over a sequence of sets, which is denoted as SDA-Max. Similar bootstrap calibration for the MTT for the global test of (2.1) was used to obtain more accurate quantile of $\mathcal{V}_{\mathcal{D}_l}(s_0)$ under the null hypotheses in the proposed SDA-MTT procedure. The covariances were set as $\mathbf{\Sigma}_1 = 0.6\mathbf{I}_p$ and $\mathbf{\Sigma}_2 = 0.6\mathbf{I}_p + \mathbf{U}$, where $\mathbf{U}$ is symmetric with entries $u_{ll+1} = u_{l+1l} = \{4r_l \log(p)/n\}^{1/2}$ for $l = \lfloor q^{1-\beta}/2 \rfloor$ and $u_{kl} = 0$ otherwise. To mimic a range of signal strength, r_l was set as 0.8, 0.6 and 0.5 with proportions 0.2, 0.4 and 0.4, while the sparsity β was set as 0.5 and 0.6, respectively. Six combinations of sample sizes and dimension were considered as reported in Table 2. The exceedance parameter c and significant level α in the step-down procedures for SDA-MTT and SDA-Max were 0.1 and 0.05, respectively.

Table 2 reports the empirical proportion of FDP larger than 0.1 (FDPEx), the FDR, and the power (number of true discoveries over number of different covariances) of the four signal recovery procedures, SDA-MTT, SR, SR-FWER and SDA-Max. We observe that the proposed SDA-MTT procedure could control the FDP exceedance rate around the nominal

Table 2
The empirical rate of FDP > 0.1 (FDPEx), FDR and power of the proposed SDA-MTT procedure, the support recovery procedures (SR and SR-FWER) in Cai et al. (2013) which are based on thresholding M_{ij}, and SDA-Max procedure that applies the $\ell_{\max}$-test of Cai et al. (2013) in conjunction with the stepdown-augmentation procedure for the multiple hypotheses (4.1) of signal identification at $\alpha = 0.05$ and $c = 0.1$, under the mixture signal setting with strengths 0.8, 0.6 and 0.5 by proportions 0.2, 0.4 and 0.4 respectively, and signal sparsity $\beta = 0.5, 0.6$.

Methods	$n_1 = 150$, $n_2 = 150$, $p = 100$						$n_1 = 200$, $n_2 = 200$, $p = 100$					
	$\beta = 0.5$			$\beta = 0.6$			$\beta = 0.5$			$\beta = 0.6$		
	FDPEx	FDR	Power	FDPEx	FDR	Power	FDPEx	FDR	Power	FDPEx	FDR	Power
SDA-MTT	0.042	0.038	0.956	0.064	0.030	0.910	0.028	0.048	0.966	0.042	0.036	0.918
SR-FWER	0	0.006	0.804	0.006	0.014	0.752	0	0.006	0.826	0.006	0.012	0.772
SR	0	0.008	0.840	0.006	0.016	0.794	0	0.008	0.860	0.006	0.014	0.810
SDA-Max	0	0.006	0.876	0	0.008	0.806	0	0.008	0.898	0	0.006	0.832
	$n_1 = 150$, $n_2 = 150$, $p = 200$						$n_1 = 200$, $n_2 = 200$, $p = 200$					
SDA-MTT	0.048	0.058	0.976	0.058	0.042	0.938	0.014	0.056	0.980	0.042	0.044	0.960
SR-FWER	0	0.002	0.806	0	0.006	0.762	0	0.004	0.836	0	0.006	0.796
SR	0	0.004	0.836	0	0.008	0.788	0	0.004	0.862	0	0.008	0.826
SDA-Max	0	0.006	0.886	0	0.006	0.822	0	0.006	0.918	0	0.006	0.868
	$n_1 = 150$, $n_2 = 150$, $p = 400$						$n_1 = 200$, $n_2 = 200$, $p = 400$					
SDA-MTT	0.040	0.064	0.984	0.054	0.056	0.958	0.024	0.068	0.988	0.026	0.052	0.966
SR-FWER	0	0.002	0.810	0	0	0.768	0	0.002	0.844	0	0.004	0.790
SR	0	0.002	0.836	0	0.006	0.796	0	0.002	0.866	0	0.004	0.820
SDA-Max	0	0.002	0.896	0	0.004	0.844	0	0.004	0.920	0	0.002	0.872

level 0.05 with high powers under all the cases. Comparing with SDA-MTT, the other three approaches were less powerful and more conservative with FDR close to 0. Simply thresholding M_{ij} by a universal level (SR and SR-FWER) were able to recover 70%–80% of the signals. But they were not as powerful as the proposed SDA-MTT whose power reached over 95% for most of cases. The less showing of the SDA-Max confirmed that the MTT was more powerful than the $\ell_{\max}$-test, as conveyed from Figs. 1 and 2.

8. Case study

The S&P 500 index is an important benchmark financial indicator for the U.S. and beyond. The index fell dramatically in March 2020 after the outbreak of the COVID-19 in the U.S. The stock prices of some sectors, like transportation and energy, dropped more than 60%. The impact of COVID-19 on the S&P 500 index have been recently studied via the vector autoregressive model in Yilmazkuday (2021), Lento and Gradojevic (2021). We analyzed the change of the dependence among the S&P 500 stocks before and after the outbreak of the COVID-19 pandemic. Specifically, we focus on the covariance matrix of the pre-whitened daily returns of the S&P 500 stocks, and tested for change in the covariances before and after the outbreak of the pandemic. To reflect the two regimes, we considered the time span from January 1st, 2019 to December 31st, 2020 with 252 trading days as the pre-pandemic period, and the time range from April 1st, 2020 to September 1st, 2021 (342 trading days) as the in-pandemic period. As the COVID-19 virus was detected in January 2020, and the sudden crash of the U.S. stock market happened in March 2021, we did not include the period from January 1st to March 31st, 2020 in the analysis to avoid the most volatile impacts of the pandemic.

After removing the non-overlapping S&P 500 stocks in the two periods, there were 493 common stocks left for the analysis. Let Z_{1t_1j} and Z_{2t_2j} be the closing prices of the jth stock at the t_1th day in the pre-pandemic period and at the t_2th day in the in-pandemic period, respectively, where $t_1 = 1, \ldots, n_1 = 252$ and $t_2 = 1, \ldots, n_2 = 342$. Let $R_{ltj} = \log(Z_{ltj}/Z_{lt-1j})$ be the daily log-returns for $l = 1$ and 2. To remove the time dependence in the returns $\{R_{ltj}\}$, we fit an auto-regressive model with order 1 to each stock for the pre-whitening purpose (Cryer and Chan, 2008). Specifically, we fitted the AR(1) model $R_{ltj} = \phi_{j0} + \phi_{j1}R_{lt-1j} + \epsilon_{ltj}$ for $l = 1, 2$ and $j = 1, \ldots, p = 493$. Let $\boldsymbol{\epsilon}_{lt} = (\epsilon_{lt1}, \ldots, \epsilon_{ltp})^{\mathrm{T}}$, and $\boldsymbol{\Sigma}_l$ be the covariance matrix of the pre-whitened return $\boldsymbol{\epsilon}_{lt}$ for period l.

We first considered testing the overall hypotheses $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ vs. $H_a : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. The proposed multi-thresholding test with bootstrap calibration (MTT-BT) and the maximum test of Cai et al. (2013) (CLX) were applied on the residuals $\{\hat{\boldsymbol{\epsilon}}_{lt}\}$ from the fitted AR(1) model for each stock. Both MTT-BT and CLX rejected the null hypothesis of equal covariance matrices at the significant level 0.05. Furthermore, we applied the proposed multiple testing procedure SDA-MTT to identify the pairs of stocks with different covariances, as well as the other three procedures, SR-FWER, SR and SDA-Max, considered in the simulation studies. Table 3 provides the cross tabulation of the numbers of identified pairs of stocks among the four methods, where the diagonal entries give the numbers of identified pairs by each method, and the off-diagonal values present the number of commonly identified pairs by two test methods.

Table 3 shows that the SDA-MTT procedure detected 717 significant stock pairs with different covariances, which was the highest among the four methods. The SDA-MTT procedure identified all the significantly different covariances found by the other three methods, reflecting higher power of the MTT test. The detected pairs by the other three tests were nested so that all the detected stock pairs by SR-FWER (SDA-Max) were also detected by SDA-Max (SR), which was due to the three tests are all based on the sorted standardized difference M_{ij} for the cut-off values.

Table 3
Cross tabulations of the numbers of stock pairs with significantly different covariances for the S&P 500 data between the pre-pandemic and in-pandemic periods. The numbers on the diagonals show the significant stock pairs detected by each test, while the off-diagonal entries are the numbers of common stock pairs detected by the two methods in the row and column.

Methods	SDA-MTT	SR-FWER	SR	SDA-Max
SDA-MTT	717	247	318	274
SR-FWER		247	247	247
SR			318	274
SDA-Max				274

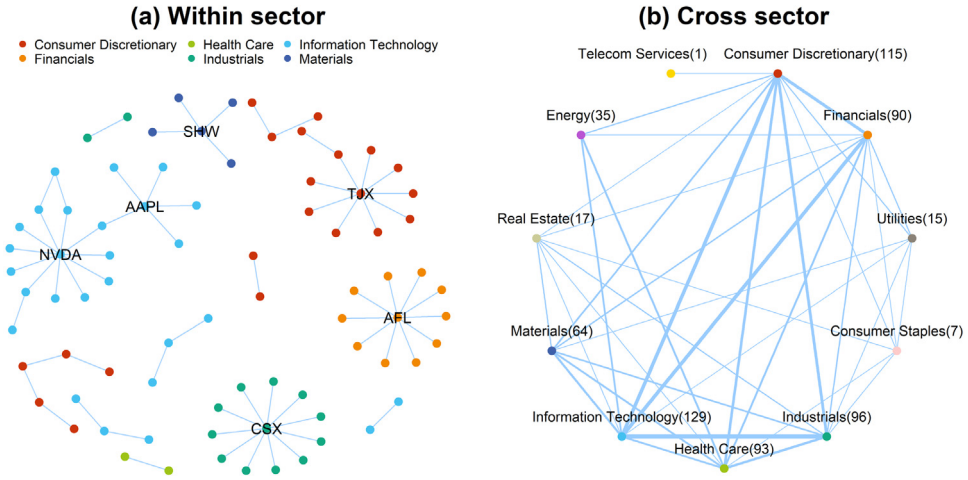

Fig. 3. Network connectivity plots for those significant stock pairs discovered by SDA-MTT but not SDA-MAX. For the within sector plot in Panel (a), two nodes connected by an edge represent the significant pairs additionally detected by SDA-MTT; for the cross sector plot (Panel b), each node represents a sector and the edge width is proportional to the number of the additional significant stock pairs. The numbers inside the parentheses in Panel (b) were the numbers of additional significant cross sector stock pairs found by SDA-MTT.

As both SDA-MTT and SDA-Max procedures are built on the step-down process, we focus on the two methods. Comparing with SDA-Max, the proposed SDA-MTT procedure additionally detected 79 within sector and 364 cross sector stock pairs. Fig. 3 displays network connectivity plots for significant within sector and cross sector covariance pairs detected only by SDA-MTT. We notice from Panel (a) that the proposed procedure identified at least six within sector stock clusters with significant covariances changes, not discovered by SDA-MAX. Apple Inc. (AAPL) and NVIDIA Corporation (NVDA) in the Information Technology sector, Aflac Inc. (AFL) in the Financial sector and CSX Corporation (CSX) in the Industrial sector, TJX Companies Inc. (TJX) in the Consumer Discretionary sector and Sherwin-Williams Company (SHW) in the Materials sector were the hubs of the newly detected covariance clusters. In particularly, the covariances connected to Apple, NVIDIA and Sherwin-Williams decreased after the pandemic as compared to the pre-pandemic period, while those associated with AFL, CSX and TJX increased positively. Changes of covariances among stocks would affect portfolio allocation, as holdings of positively correlated stocks tend to have higher risk profile than holdings of weakly or negatively correlated stocks.

To better understand the changes of covariances before and after the start of the pandemic, we summarize and report three categories of the identified significant pairs by SDA-MTT and SDA-MAX (in parenthesis) in Table 4, which are "positively enhanced", "negatively enhanced" and "decreased in strength". The "positively (negatively) enhanced" category contained stock pairs whose covariances increased (decreased) and their values were positive (negative) in the in-pandemic period. And the "decreased in strength" category includes pairs with unchanged signs of covariances in the two periods but the magnitude of the covariances decreased. From Table 4, it is noticed that most of the significant pairs both within and cross the Financial and Industrial sectors were positively enhanced, while most of those associated with the Information Technology and Health Care sectors were either negatively enhanced or decreased in the covariance. These findings on the dependence patterns within and cross stocks of sectors detected by the proposed SDA-MTT procedure would have immediate implications on the risk management strategies as we know stocks in a portfolio with higher positive (lower negative) covariances would generates higher (lower) risk. Thus, a portfolio mainly allocated in the Financial and Industrial sectors may suffer a higher risk during the in-pandemic period than the pre-pandemic period. Furthermore, portfolios invested in the Information Technology and Health Care sectors may have a better control of risks.

Table 4
Number of significantly different covariance stock pairs detected by SDA-MTT and SDA-Max (in parentheses) within sectors (diagonal entries) and cross sectors (off diagonal entries) of S&P 500 stocks for three categories of significant stock pairs: "Positively enhanced", "Negatively enhanced" and "Decreased in strength". The number of stocks in each sector is given in the parenthesis to the sector name. The numbers of significant different covariance detected by the SDA-MTT (SDA-Max) for the three categories were 219 (105), 312 (96) and 186 (73), respectively.

(a) Positively enhanced

	Cons.Disc.	Utilities	Financials	Industrials	I.T.	Health Care	Energy	Cons.Stap.	Materials	Real Estate	Telecom
Cons.Disc. (77)	18 (7)	9 (3)	27 (14)	10 (4)	4 (2)	0 (0)	0 (0)	2 (1)	7 (3)	5 (4)	1 (0)
Utilities (27)		20 (20)	7 (0)	5 (3)	1 (0)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)
Financials (66)			33 (21)	17 (12)	3 (2)	2 (0)	3 (0)	3 (2)	7 (3)	1 (0)	0 (0)
Industrials (68)				10 (1)	2 (0)	0 (0)	0 (0)	1 (0)	2 (0)	5 (0)	0 (0)
I.T. (83)					8 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Health Care (60)						0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Energy (22)							0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Cons.Stap. (33)								0 (0)	0 (0)	0 (0)	0 (0)
Materials (26)									1 (0)	4 (0)	0 (0)
Real Estate (27)										0 (0)	0 (0)
Telecom (4)											0 (0)

(b) Negatively enhanced

	Cons.Disc.	Utilities	Financials	Industrials	I.T.	Health Care	Energy	Cons.Stap.	Materials	Real Estate	Telecom
Cons.Disc. (77)	15 (7)	0 (0)	13 (2)	14 (3)	42 (18)	26 (6)	10 (3)	2 (1)	8 (1)	0 (0)	0 (0)
Utilities (27)		0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)	0 (0)
Financials (66)			0 (0)	0 (0)	46 (20)	13 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Industrials (68)				1 (0)	40 (14)	10 (2)	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)
I.T. (83)					6 (3)	3 (1)	10 (0)	0 (0)	14 (2)	2 (0)	0 (0)
Health Care (60)						0 (0)	24 (8)	1 (0)	4 (1)	4 (2)	0 (0)
Energy (22)							0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Cons.Stap. (33)								0 (0)	0 (0)	2 (0)	0 (0)
Materials (26)									0 (0)	0 (0)	0 (0)
Real Estate (27)										0 (0)	0 (0)
Telecom (4)											0 (0)

(c) Decreased in strength

	Cons.Disc.	Utilities	Financials	Industrials	I.T.	Health Care	Energy	Cons.Stap.	Materials	Real Estate	Telecom
Cons.Disc. (77)	0 (0)	0 (0)	0 (0)	3 (0)	8 (4)	1 (0)	0 (0)	0 (0)	3 (2)	0 (0)	0 (0)
Utilities (27)		1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Financials (66)			0 (0)	3 (1)	5 (1)	3 (0)	0 (0)	0 (0)	4 (0)	0 (0)	0 (0)
Industrials (68)				8 (4)	21 (10)	12 (3)	0 (0)	0 (0)	26 (14)	0 (0)	0 (0)
I.T. (83)					45 (24)	19 (6)	1 (0)	0 (0)	8 (3)	0 (0)	0 (0)
Health Care (60)						1 (0)	0 (0)	0 (0)	10 (0)	0 (0)	0 (0)
Energy (22)							0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Cons.Stap. (33)								0 (0)	0 (0)	0 (0)	0 (0)
Materials (26)									3 (0)	1 (0)	0 (0)
Real Estate (27)										0 (0)	0 (0)
Telecom (4)											0 (0)

9. Discussion

Testing for covariance of regression errors. Example 1 and the case study consider testing the covariance of regression errors where the estimated residuals are used to formulate the test procedure. Here, we briefly explain that the limiting distribution of the proposed test statistic based on the estimated residuals would be the same as that based on true regression errors. To illustrate the key point, we consider the one-sample case for simplicity. Let $\tilde{X}_{kj} = \tilde{\mathbf{F}}_k^{\mathrm{T}}\boldsymbol{\gamma}_j + \epsilon_{kj}$ and $\boldsymbol{\epsilon}_k = (\epsilon_{k1}, \ldots, \epsilon_{kp})^{\mathrm{T}}$ for $j = 1, \ldots, p$ and $k = 1, \ldots, n$, where $\tilde{\mathbf{F}}_k$ are fixed dimensional covariates, and $\{\boldsymbol{\epsilon}_k\}$ are i.i.d. with $\mathrm{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_\epsilon$. Let $\hat{\boldsymbol{\gamma}}_j$ and $\{\hat{\epsilon}_{kj}\}$ be the estimated coefficient and residuals, respectively. Let $\hat{\sigma}_{\epsilon,ij}$ be the sample covariance between $\{\hat{\epsilon}_{ki}\}_{k=1}^n$ and $\{\hat{\epsilon}_{kj}\}_{k=1}^n$, $\hat{\theta}_{\epsilon,ij} = \sum_{k=1}^n (\hat{\epsilon}_{ki}\hat{\epsilon}_{kj} - \hat{\sigma}_{\epsilon,ij})^2/n$, and $\hat{F}_{\epsilon,ij} = \sqrt{n}\hat{\sigma}_{\epsilon,ij}\hat{\theta}_{\epsilon,ij}^{-1/2}$ be the standardized sample covariance. The proposed thresholding statistic for testing $\boldsymbol{\Sigma}_\epsilon$ being diagonal can be constructed as $T_{\epsilon,n}(s) = \sum_{1\le i<j\le p} \hat{F}_{\epsilon,ij}^2 \mathbb{I}\{\hat{F}_{\epsilon,ij}^2 > \lambda_p(s)\}$.

Let $\tilde{\sigma}_{\epsilon,ij} = \sum_{k=1}^n \epsilon_{ki}\epsilon_{kj}/n - \bar{\epsilon}_i\bar{\epsilon}_j$ and $\theta_{\epsilon,ij} = \mathrm{Var}(\epsilon_{ki}\epsilon_{kj})$, where $\bar{\epsilon}_i = \sum_{k=1}^n \epsilon_{ki}/n$. Let $\tilde{T}_{\epsilon,n}(s)$ be the same thresholding statistic built on the true residuals $\epsilon_{kj} = \tilde{X}_{kj} - \tilde{\mathbf{F}}_k^{\mathrm{T}}\boldsymbol{\gamma}_j$. From the proofs of Proposition 1 and Theorem 1, the key to derive the limiting distribution of $T_{\epsilon,n}(s)$ is the large deviation result for $\hat{F}_{\epsilon,ij}$, similar as Lemma 2 for F_{ij} in the SM. Since $\hat{\sigma}_{\epsilon,ij}$ can be written as $\tilde{\sigma}_{ij}$ plus some small order terms, it can be shown that the tail probability $\mathbb{P}(\sqrt{n}\hat{\sigma}_{\epsilon,ij}\hat{\theta}_{\epsilon,ij}^{-1/2} > c\sqrt{\log p})$ is asymptotically equivalent to $\mathbb{P}(\sqrt{n}\tilde{\sigma}_{\epsilon,ij}\theta_{\epsilon,ij}^{-1/2} > c\sqrt{\log p})$ for any positive constant c. Hence, the limiting distributions of $T_{\epsilon,n}(s)$ and $\tilde{T}_{\epsilon,n}(s)$ would be the same.

Notice that $T_{\epsilon,n}(s)$ can be viewed as a plug-in version of $\tilde{T}_{\epsilon,n}(s)$ with $\boldsymbol{\gamma}_j$ replaced by $\hat{\boldsymbol{\gamma}}_j$. It is well known that plug-in statistics typically have more complicated asymptotic distributions due to the additional variability of the plugged-in estimates. However, the proposed statistic does not suffer from this problem due to the regularization offered by the thresholding, which avoids error accumulation as in ℓ_2-type statistics.

Testing for covariance under temporally dependent data. The proposed test procedures are developed for independent data. The proposed approach can be extended to weakly dependent data. Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ are stationary time series with mean zero and covariance matrices $\boldsymbol{\Sigma}_1 = (\sigma_{ij1})$ and $\boldsymbol{\Sigma}_2 = (\sigma_{ij2})$ respectively. Clearly, we still use $\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2}$ to estimate the difference between σ_{ij1} and σ_{ij2}. However, due to the time dependence, θ_{ij1} and θ_{ij2} are no longer the variances of $\sqrt{n_1}\hat{\sigma}_{ij1}$ and $\sqrt{n_2}\hat{\sigma}_{ij2}$. Long-run variances of the sequences $\{X_{ki}X_{kj}\}_{k=1}^{n_1}$ and $\{Y_{ki}Y_{kj}\}_{k=1}^{n_2}$ need to be estimated, which can be used to construct a standardized estimate of $\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2}$. Let $\widehat{W}_{ij1}$ and $\widehat{W}_{ij2}$ be the long-run covariance estimators based on the kernel smoothing method (Andrews, 1991) for the sequences $\{X_{ki}X_{kj}\}_{k=1}^{n_1}$ and $\{Y_{ki}Y_{kj}\}_{k=1}^{n_2}$, respectively. Similar as F_{ij} for the independent case, the standardized difference between $\hat{\sigma}_{ij1}$ and $\hat{\sigma}_{ij2}$ is defined as $\hat{F}_{ij} = (\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2})(\widehat{W}_{ij1}/n_1 + \widehat{W}_{ij2}/n_2)^{-1/2}$ for the time-series case. Then, similar as (3.1), the single level thresholding statistic is $\hat{T}_n(s) = \sum_{1\le i<j\le p} \hat{F}_{ij}^2 \mathbb{I}\{\hat{F}_{ij}^2 > \lambda_p(s)\}$, and the multi-level thresholding test statistic can be constructed in the same way as (3.8). Similar as the discussion on the regression errors, a theoretical validation of the proposed procedure under time dependent data relies on the large deviation results of $\hat{F}_{ij}$ under weakly dependent data, which can be derived under suitable conditions.

Similarly, the proposed test could also be extended to time series data with heteroskedasticity, for instance the multivariate GARCH model (Bollerslev, 1990) where the conditional variance of each variable is modeled by a univariate GARCH model and the correlations among variables do not change over time. The interest is on testing the correlation matrix among variables. Note that the correlation between X_{ki} and X_{kj} can be estimated by the standardized data $\check{X}_{k,i} = \hat{d}_{k,i}^{-1/2} X_{k,i}$, where $\hat{d}_{k,i}$ is an estimate of the conditional variance of X_{ki} under the GARCH model. Similar as the time series with homogeneous variances discussed above, the proposed test for the correlation matrix can be constructed based on the standardized data $\check{X}_{k,i}$. We would consider these topics in future investigation.

Acknowledgments

We thank the Associate Editor and two referees for constructive comments and suggestions which have improved both the content and presentation of the work. The research was partially supported by National Natural Science Foundation of China Grants 92046021, 12026607, 12071013 and 11971390, Center for Statistical Science at Peking University; LMEQF at Peking University, China; the Office of Science (BER), U.S. Department of Energy, USA, Grant no. DE-SC0020355; and the Plant Sciences Institute, Iowa State University, Scholars Program, USA.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.10.008.

References

Amosova, N.N., 2002. Necessity of the cramér, linnik, and statulevičius conditions for large-deviation probabilities. J. Math. Sci. 109, 2031–2036.
Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third ed. John Wiley & Sons, New York.
Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.
Arbia, G., Baltagi, B.H., 2008. Spatial Econometrics: Methods and Applications. Physica-Verlag.
Arias-Castro, E., Bubeck, S., Lugosi, G., 2012. Detection of correlations. Ann. Statist. 40, 412–435.
Athreya, K., Lahiri, S., 2006. Measure Theory and Probability Theory. Springer, New York.
Bai, Z.D., Jiang, D.D., Yao, J.F., Zheng, S.R., 2009. Corrections to LRT on large-dimensional covariance matrix by RMT. Ann. Statist. 37, 3822–3840.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57, 289–300.
Bickel, P., Levina, E., 2008. Covariance regularization by thresholding. Ann. Statist. 36, 2577–2604.
Bollerslev, T., 1990. Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. Rev. Econ. Stat. 72, 498–505.
Bradley, R., 2005. Basic properties of strong mixing conditions: a survey and some open questions. Prob. Surv. 2, 107–144.
Cai, T., Liu, W.D., Xia, Y., 2013. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. J. Am. Stat. Assoc. 108, 265–277.
Chang, J.Y., Zhou, W., Zhou, W.X., Wang, L., 2017. Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. Biometrics 73, 31–41.
Chen, S.X., Li, J., Zhong, P.S., 2019. Two-sample and ANOVA tests for high dimensional means. Ann. Statist. 47, 1443–1474.
Chudik, A., Kapetanios, G., Pesaran, M.H., 2018. A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. Econometrica 86, 1479–1512.
Cryer, J.D., Chan, K.S., 2008. Time Series Analysis: with Applications in R. Springer, New York.
Delaigle, A., Hall, P., Jin, J., 2011. Robustness and accuracy of methods for high dimensional data analysis based on student's t-statistic. J. R. Stat. Soc. Ser. B Stat. Methodol. 73, 283–301.
Donoho, D., Jin, J., 2004. Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist. 32, 962–994.
Donoho, D., Jin, J., 2015. Higher criticism for large-scale inference, especially for rare and weak effects. Stat. Sci. 30, 1–25.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financial Econ. 33, 3–56.

Fan, J., 1996. Test of significance based on wavelet thresholding and Neyman's truncation. J. Amer. Statist. Assoc. 91, 674–688.
Fan, J., Liao, Y., Shi, X., 2015a. Risks of large portfolios. J. Econometrics 186, 367–387.
Fan, J., Liao, Y., Yao, J., 2015b. Power enhancement in high-dimensional cross-sectional tests. Econometrica 83, 1497–1541.
Gaetan, C., Guyon, X., 2010. Spatial Statistics and Modeling. Springer, New York.
Genovese, C., Wasserman, L., 2006. Exceedance control of the false discovery proportion. J. Amer. Statist. Assoc. 101, 1408–1417.
Hall, P., Heyde, C.C., 1980. Martingale Limit Theory and Its Application. Academic Press.
Hall, P., Jin, J., 2010. Innovated higher criticism for detecting sparse signals in correlated noise. Ann. Statist. 38, 1686–1732.
He, Y., Xu, G., Wu, C., Pan, W., 2021. Asymptotically independent U-statistics in high-dimensional testing. Ann. Statist. 49, 154–181.
Jenish, N., Prucha, I.R., 2009. Central limit theorems and uniform laws of large numbers for arrays of random fields. J. Econometrics 150, 86–98.
Ji, P., Jin, J., 2012. UPS delivers optimal phase diagram in high-dimensional variable selection. Ann. Statist. 40, 73–103.
Lento, C., Gradojevic, N., 2021. S & P 500 index price spillovers around the COVID-19 market meltdown. J. Risk Financial Manag. 14 (330).
Li, J., Chen, S.X., 2012. Two sample tests for high-dimensional covariance matrices. Ann. Statist. 40, 908–940.
Liu, W., 2013. Gaussian graphical model estimation with false discovery rate control. Ann. Statist. 41, 2948–2978.
Liu, B., Zhang, X., Liu, Y., 2021. Simultaneous change point inference and structure recovery for high dimensional Gaussian graphical models. J. Mach. Learn. Res. 22, 1–62.
Markowitz, H.M., 1952. Portfolio selection. J. Finance 7, 77–91.
Mokkadem, A., 1988. Mixing properties of ARMA processes. Stochastic Process. Appl. 29, 309–315.
Nagao, H., 1973. On some test criteria for covariance matrix. Ann. Statist. 1, 700–709.
Perlman, M.D., 1980. Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. Ann. Statist. 8, 247–263.
Petrov, V.V., 1995. Limit Theorems of Probability Theorey: Sequences of Independent Random Variables. Clarendon Press, London.
Qiu, Y., Chen, S.X., 2012. Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. Ann. Statist. 40, 1285–1314.
Qiu, Y., Chen, S.X., Nettleton, D., 2018. Detecting rare and faint signals via thresholding maximum likelihood estimators. Ann. Statist. 46, 895–923.
Qiu, H., Han, F., Liu, H., Caffo, B., 2016. Joint estimation of multiple graphical models from high dimensional time series. J. R. Stat. Soc. Ser. B Stat. Methodol. 78, 487–504.
Ren, Z., Sun, T., Zhang, C.H., Zhou, H., 2015. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. Ann. Statist. 43, 991–1026.
Ross, N., 2011. Fundamentals of Stein's method. Prob. Surv. 8, 210–293.
Rothman, A.J., 2012. Positive definite estimators of large covariance matrices. Biometrika 99, 539–550.
Saulis, L., Statulevičius, V.A., 1991. Limit Theorems for Large Deviations. Kluwer Academic, Dordrecht.
Schott, J.R., 2007. A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. Comput. Statist. Data Anal. 51, 6535–6542.
Srivastava, M.S., Yanagihara, H., 2010. Testing the equality of several covariance matrices with fewer observations than the dimension. J. Multivariate Anal. 101, 1319–1329.
Stein, C., 1972. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory. Vol. 6. pp. 583–603.
Storey, J.D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 64, 479–498.
Xu, G., Lin, L., Wei, P., Pan, W., 2016. An adaptive two-sample test for high-dimensional means. Biometrika 103, 609–624.
Xue, L.Z., Ma, S.Q., Zou, H., 2012. Positive-definite ℓ_1-penalized estimation of large covariance matrices. J. Am. Stat. Assoc. 107, 1480–1491.
Yilmazkuday, H., 2021. COVID-19 effects on the S & P 500 index. Appl. Econ. Lett. 1–7.
Yu, X., Li, D., Xue, L., Li, R., 2021. Power-enhanced simultaneous test of high-dimensional mean vectors and covariance matrices with application to gene-set testing. arXiv:2109.15287.
Zhong, P.S., Chen, S.X., Xu, M.Y., 2013. Tests alternative to higher criticism for high dimensional means under sparsity and column-wise dependence. Ann. Statist. 41, 2820–2851.