Statistica Sinica **28** (2018), 2671-2696
doi:https://doi.org/10.5705/ss.202017.0213

HIGH-DIMENSIONAL TWO-SAMPLE COVARIANCE MATRIX TESTING VIA SUPER-DIAGONALS

Jing He and Song Xi Chen

Southwestern University of Finance and Economics and Peking University

Abstract: This paper considers testing for two-sample covariance matrices of high-dimensional populations. We formulate a multiple test procedure by comparing the super-diagonals of the covariance matrices. The asymptotic distributions of the test statistics are derived and the powers of individual tests are studied. The test statistics, by focusing on the super-diagonals, have smaller variation than the existing tests that target on the entire covariance matrix. The advantage of the proposed test is demonstrated by simulation studies, as well as an empirical study on a prostate cancer dataset.

Key words and phrases: High dimensional test, multiple test, sparse alternative, two-sample test for covariance matrices.

1. Introduction

With the advent of the information technology used for data collection, data whose dimensionality is much larger than the sample sizes are increasingly encountered in statistical analyses. Conventional statistical inference methods proposed under this setting, and their performance under the paradigm of high dimensionality, require investigation and update. Thus, in checking on the equality of two multivariate distributions in the study of certain treatment effects, Hotelling's test for the means (Hotelling (1931)) has shortcomings in a high-dimensional setting, as shown in Bai and Saranadasa (1996) who proposed a modification that removes the inverse of the sample covariance (S_n) from the original test statistic. Srivastava and Du (2008) considered using the diagonal matrix of S_n to replace S_n, and Chen and Qin (2010) suggested using U-statistics. See also Cai, Liu and Xia (2014) for a test based on the maximal norm, and Hall and Jin (2009) that utilizes the dependence to enhance the signal strength of the testing problem.

Testing for the equality of two covariance matrices constitutes another way for checking on the equality of the two distributions by focusing on the dependence structure. For instance, one determines if pooling of the two sample vari-

ances can be exercised in linear discriminant analysis and Hotelling's test for the two-sample means. Let X_1 and X_2 be generic p-dimensional random vectors from two populations with respective covariance matrices Σ_1 and Σ_2. Let $\{\mathbf{X}_{1,j}\}_{j=1}^{n_1}$ be independent and identically distributed (i.i.d.) copies of X_1 and $\{\mathbf{X}_{2,j}\}_{j=1}^{n_2}$ be i.i.d. copies of $\mathbf{X}_2$ that are mutually independent.

The primary interest of this paper is in testing

$$\mathrm{H}_0 : \Sigma_1 = \Sigma_2 \text{ vs. } \mathrm{H}_1 : \Sigma_1 \neq \Sigma_2. \tag{1.1}$$

In the conventional fixed-dimensional setting, the testing problem had been well studied in classical multivariate analysis, as summarized in Anderson (2003) and Muirhead (1982), which include the likelihood ratio tests and related formulations as investigated in John (1972), Nagao (1973), and Gupta and Tang (1984). However, these conventional methods may no longer be valid under the high-dimensional setting. Here, Bai et al. (2009) found that the two-sample likelihood ratio test (LRT) is not consistent due to a bias caused by the inconsistency of the sample covariance estimator (Bai and Yin (1993); Johnstone (2001)).

Two sample covariance testing in high dimensions has attracted much attention in the last decade or so. Schott (2007) proposed a test that targeted the Frobenius norm of $\Sigma_1 - \Sigma_2$ for $p/n_i \to c_i \in [0, \infty)$, $i = 1, 2$, with Gaussian distributed data. Bai et al. (2009) developed a corrected LRT under $p/n_i \to c_i \in (0, 1)$. For much higher dimensionality, Srivastava and Yanagihara (2010) considered a test based on a consistent estimator of $\mathrm{tr}(\Sigma_1^2)/\{\mathrm{tr}(\Sigma_1)\}^2 - \mathrm{tr}(\Sigma_2^2)/\{\mathrm{tr}(\Sigma_2)\}^2$, also for Gaussian data. Li and Chen (2012) suggested using U-statistics as the test statistic, which is an unbiased estimator of the Frobenius norm of $\Sigma_1 - \Sigma_2$. Cai, Liu and Xia (2013) introduced a test statistic defined as the maximum of standardized element-wise differences between the two sample covariance matrices. The last two tests are nonparametric, allowing flexible data distributions of the two populations.

We develop a test for (1.1) by targeting the differences in the super-diagonals between the two covariance matrices. For a square matrix $(a_{i,j})_{p\times p}$, we define the l-th super-diagonal consisting of the matrix elements $\{a_{k,k+\ell}\}_{k=1}^{p-\ell}$ for $\ell = 1, \ldots, p-1$, which is broader than only referring to those elements directly above the main diagonal. The main diagonal corresponds to $\ell = 0$. The purpose of designing the test in such a fashion is to re-distribute the dimensionality of the testing problem for power gains. Most of the existing tests tend to focus on the entire difference between Σ_1 and Σ_2 where the test statistics gather the variation from all the components in the high-dimensional covariance matrices.

This accumulation of variance can be severe due to the high dimensionality, and hence can undermine the power of the tests.

The proposed test is formulated by first conducting two-sample tests on the super-diagonals followed by a simultaneous test over multiple super-diagonals to produce an overall test rule. The simultaneous test is facilitated by using a multiple testing procedure advocated in Benjamini and Hochberg (1995) and Storey, Taylor and Siegmund (2004). The test statistic on each super-diagonal is of a smaller dimension than those based on the whole $\Sigma_1 - \Sigma_2$ matrix, as the size of each super-diagonal is no more than p whereas the size of $\Sigma_1 - \Sigma_2$ is of order p^2. Although the overall dimensionality involved in the test is not necessarily less if we combine the two stages of the formulation, the lower dimensionality in the first stage brings in more power. The power of the proposed test dominates especially when Σ_1 and Σ_2 have a "bandable" structure and the differences between them are sparse in terms of super-diagonals, as compared with the tests of Srivastava and Yanagihara (2010), Li and Chen (2012) and Cai, Liu and Xia (2013).

The rest of this paper is structured as follows. Section 2 provides the notation and technical assumptions, and introduces the test statistics. Section 3 investigates the theoretical properties of the proposed test. Section 4 studies its numerical performance. Section 5 reports an analysis on a prostate cancer dataset. Proofs of the main results are in the supplementary document.

2. Preliminaries

Let $\{\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n_i}\}$ be i.i.d. samples from populations $\mathbf{X}_i$ for $i = 1$ and 2, where $\mathbf{X}_{i,j} = (\mathrm{X}_{i,j,1}, \ldots, \mathrm{X}_{i,j,p})^\top$, $j = 1, \ldots, n_i$, is a p-dimensional random vector with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\Sigma_i = (\sigma_{i,k,\ell})_{p\times p}$.

Let $\mathrm{S}_q = \sum_{s=1}^{p-q}(\sigma_{1,s,s+q} - \sigma_{2,s,s+q})^2$ measure the difference between the q-th super-diagonal of Σ_1 and Σ_2 for $q = 0, \ldots, p-1$, where $q = 0$ stands for the main diagonal. Let $\mathrm{D}_{i,q} = \sum_{s=1}^{p-q}\sigma_{i,s,s+q}^2$ for $i = 1$ and 2, and $\mathrm{D}_{c,q} = \sum_{s=1}^{p-q}\sigma_{1,s,s+q}\sigma_{2,s,s+q}$. It can be shown that $\mathrm{S}_q = \mathrm{D}_{1,q} + \mathrm{D}_{2,q} - 2\mathrm{D}_{c,q}$. We will first develop a test procedure for testing $\mathrm{H}_{0,q} : \mathrm{S}_q = 0$ versus $\mathrm{H}_{1,q} : \mathrm{S}_q > 0$ that facilitates a two-sample test on the covariance matrices via multiple testing of $\mathrm{H}_{0,q}$ over a range of q.

We first propose an unbiased estimator of S_q for the q-th super-diagonal by constructing unbiased estimators of $\mathrm{D}_{1,q}$, $\mathrm{D}_{2,q}$ and $\mathrm{D}_{c,q}$, respectively. Motivated by the unbiasedness and other attractive properties of U-statistics, we propose linear combinations of U-statistics,

$$\hat{\mathrm{D}}_{i,nq} = \sum_{s=1}^{p-q} \Bigg\{ \frac{1}{P_{n_i}^2} \sum_{j,k}^{*} (\mathrm{X}_{i,j,s}\mathrm{X}_{i,j,s+q})(\mathrm{X}_{i,k,s}\mathrm{X}_{i,k,s+q}) - \frac{2}{P_{n_i}^3} \sum_{j,k,\ell}^{*} \mathrm{X}_{i,j,s}\mathrm{X}_{i,k,s+q}(\mathrm{X}_{i,\ell,s}\mathrm{X}_{i,\ell,s+q}) + \frac{1}{P_{n_i}^4} \sum_{j,k,\ell,m}^{*} \mathrm{X}_{i,j,s}\mathrm{X}_{i,k,s+q}\mathrm{X}_{i,\ell,s}\mathrm{X}_{i,m,s+q} \Bigg\} \tag{2.1}$$

for $i = 1$ and 2, and

$$\begin{aligned} \hat{\mathrm{D}}_{c,nq} = \sum_{s=1}^{p-q} \Bigg\{ & \frac{1}{n_1 n_2} \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} (\mathrm{X}_{1,j,s}\mathrm{X}_{1,j,s+q})(\mathrm{X}_{2,k,s}\mathrm{X}_{2,k,s+q}) \\ & - \frac{1}{n_1 n_2 (n_1 - 1)} \sum_{j,k}^{*} \sum_{\ell=1}^{n_2} \mathrm{X}_{1,j,s}\mathrm{X}_{1,k,s+q}(\mathrm{X}_{2,\ell,s}\mathrm{X}_{2,\ell,s+q}) \\ & - \frac{1}{n_1 n_2 (n_2 - 1)} \sum_{\ell=1}^{n_1} \sum_{j,k}^{*} (\mathrm{X}_{1,\ell,s}\mathrm{X}_{1,\ell,s+q})\mathrm{X}_{2,j,s}\mathrm{X}_{2,k,s+q} \\ & + \frac{1}{n_1 n_2 (n_1 - 1)(n_2 - 1)} \sum_{j,k}^{*} \sum_{\ell,m}^{*} \mathrm{X}_{1,j,s}\mathrm{X}_{1,k,s+q}\mathrm{X}_{2,\ell,s}\mathrm{X}_{2,m,s+q} \Bigg\}, \end{aligned} \tag{2.2}$$

where $\sum^{*}$ denotes the summation over mutually distinct subscripts and $P_n^b = n!/(n-b)!$. Thus, $\hat{\mathrm{S}}_{nq}$ is given by

$$\hat{\mathrm{S}}_{nq} = \hat{\mathrm{D}}_{1,nq} + \hat{\mathrm{D}}_{2,nq} - 2\hat{\mathrm{D}}_{c,nq}, \tag{2.3}$$

To quantify the variance of $\hat{\mathrm{S}}_{nq}$, we define $\mathrm{Y}_{i,j}^{s_1,s_2} = \mathrm{X}_{i,j,s_1}\mathrm{X}_{i,j,s_2} - \sigma_{i,s_1,s_2}$ and a $(p-q) \times 1$ random vector

$$\mathbf{Y}_{i,j}(q) = \left(\mathrm{Y}_{i,j}^{1,1+q}, \ldots, \mathrm{Y}_{i,j}^{(p-q),p}\right)^{\top}$$

for each q. Let the covariance matrix of $\mathbf{Y}_{i,j}(q)$ be $W_{i,q} = (\omega_{i,q}^{s_1,s_2})_{(p-q)\times(p-q)}$, where $\omega_{i,q}^{s_1,s_2} = \mathrm{Cov}(\mathrm{Y}_{i,j}^{s_1,s_1+q}, \mathrm{Y}_{i,j}^{s_2,s_2+q})$. For each given i and q, $\{\mathbf{Y}_{i,j}(q)\}_{j=1}^{n_i}$ are i.i.d..

As the components of the data vectors concerned may be dependent, we use the notion of α-mixing (Doukhan (1994)) to measure the degree of the dependence. The α-mixing coefficient for the generic $\mathbf{X}_i = (\mathrm{X}_{i,1}, \ldots, \mathrm{X}_{i,p})^{\top}$, $i = 1, 2$, is

$$\alpha_{\mathrm{X}_i}(k) \equiv \sup_{m} \left\{ |\Pr(A \cap B) - \Pr(A)\Pr(B)| : A \in \mathcal{G}_1^m, B \in \mathcal{G}_{m+k}^p \right\},$$

where $\mathcal{G}_a^b$ denotes the σ-algebra generated by $\{\mathrm{X}_{i,a}, \ldots, \mathrm{X}_{i,b}\}$ for $a \le b$.

Denote the minimum and maximum eigenvalues of a matrix A by $\lambda_{\min}(A)$

and $\lambda_{\max}(A)$, respectively. For two nonrandom sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if there exist constants $0 < c < C$ such that $c|b_n| \le |a_n| \le C|b_n|$. Throughout, c, C and M with or without subscripts denote positive constants whose values do not depend on any parameter. We need some technical assumptions.

(C1) As $n = \min\{n_1, n_2\} \to \infty$, $n_1/(n_1+n_2) \to c$ for a fixed positive constant $c \in (0,1)$ and $p = p(n_1, n_2) \to \infty$.

(C2) For $i = 1, 2$, $\lambda_{\min}(\Sigma_i) \ge c_0 > 0$.

(C3) There are positive constants c and $\beta \in (1, \infty)$ such that the α-mixing coefficient $\alpha_{\mathrm{X}_i}(k) \le ck^{-\beta}$ for $i = 1, 2$.

(C4) There exists a $p \times m$ constant matrix $\Gamma_i = (\Gamma_{i,j,\ell})_{p\times m}$ such that $\mathbf{X}_{i,j} = \Gamma_i \mathbf{Z}_{i,j}$ and $\Gamma_i \Gamma_i^T = \Sigma_i$ for some $m \ge p$, and $\mathbf{Z}_{i,1}, \ldots, \mathbf{Z}_{i,n_i}$ are i.i.d. m-dimensional random vectors with zero mean and identity covariance matrix. If $\mathbf{Z}_{i,j} = (\mathrm{Z}_{i,j,1}, \ldots, \mathrm{Z}_{i,j,m})^T$, the $\mathrm{Z}_{i,j,\ell}$ have uniformly bounded eighth moments, and, for distinct subscripts $j_1, \ldots, j_s$ and any integers $\ell_\nu \ge 0$ with $\sum_{\nu=1}^s \ell_\nu \le 8$,

$$\mathrm{E}(\mathrm{Z}_{i,1,j_1}^{\ell_1} \mathrm{Z}_{i,1,j_2}^{\ell_2} \cdots \mathrm{Z}_{i,1,j_s}^{\ell_s}) = \mathrm{E}(\mathrm{Z}_{i,1,j_1}^{\ell_1}) \mathrm{E}(\mathrm{Z}_{i,1,j_2}^{\ell_2}) \cdots \mathrm{E}(\mathrm{Z}_{i,1,j_s}^{\ell_s}). \tag{2.4}$$

(C5) for each fixed q, $\mathrm{tr}(W_{i,q}^2) \asymp p - q$ and $\mathrm{tr}(W_{1,q} W_{2,q}) \asymp p - q$.

The first part of (C1) is a standard condition in the two-sample asymptotic analysis that prescribes that the two sample sizes are comparable asymptotically. The asymptotic mechanism of p in (C1) allows it to be much larger than the sample sizes.

Under (C3), Σ_1 and Σ_2 are "bandable" in the magnitude of the average squared elements on the super-diagonals. To appreciate this, let $h_q = \mathrm{S}_q/(p-q)$ be the average squares of the qth super-diagonal elements of $\Sigma_1 - \Sigma_2$. According to Lemma 1 in the supplementary document, for each given q, $h_q \le Cq^{-\beta}$, which indicates that h_q is bounded by a polynomial rate of q. Hence, $\sum_{q>k} h(q) \to 0$ as $k \to \infty$ and $p \to \infty$. This condition is similar to the "bandable" condition in Bickel and Levina (2008). Our proposal does not require a banded structure for the Σ_is as does Qiu and Chen (2012). We only assume a gradual decay of dependence among the components of the data vectors. Similar pattern of dependence structure was also considered in Qiu and Chen (2015). In fact, we only need assume that there is a permutation of the data components under which (C3) is satisfied. In practice, we can employ algorithms, for instance that proposed in Friendly (2002) to re-arrange the data vector to make it bandable.

Under (C3) and (C4), the eigenvalues of Σ_i are uniformly bounded from above, as shown in Lemma 2 in the supplementary document. This, along with

(C2), implies that the eigenvalues of Σ_i are uniformly bounded from below and above. Bounded eigenvalues assumptions are common in the literature of high-dimensional covariance inference, for instance in Bickel and Levina (2008) and Cai, Zhang and Zhou (2010). As the proposed test targets the super-diagonals rather than the whole covariance matrices, the bounded eigenvalues assumption can be relaxed at the expense of much more complicated derivations.

Assumption (C4) is in Bai and Saranadasa (1996), Li and Chen (2012) and Qiu and Chen (2012); it gives a general multivariate model for generating high-dimensional data with a wide range of multivariate distributions for $\mathbf{X}_{i,j}$. Condition (2.4) prescribes factorization of the moments of products to products of moments so that it can be viewed as a pseudo-independent condition of $\mathrm{Z}_{i,j}$. Trackable expressions of higher order cross moments of $\mathbf{X}_{i,j}$ can be obtained under this condition. It can be shown that, under the bounded eigenvalues assumption and (C4), the $X_{i,j,\ell}$ have uniformly bounded eighth moment. Assumption (C5) is used in the asymptotic analysis of $\hat{\mathrm{S}}_{nq}$ and is a mild condition. Under (C2) to (C4), Lemma 2 in the supplementary document shows that the eigenvalues of $W_{i,q}$ are also uniformly bounded from above. Assumption (C5) allows a small fraction of eigenvalues of $W_{i,q}$ to be zero.

Assumptions (C2) to (C5) implicitly impose restrictions on the extent of dimensionality and the dependence. Examples satisfying these conditions can be found in time series and spatial data, in which the covariances decay as the time interval or the distance grows.

According to Proposition 1 in Section 3, $\hat{\mathrm{S}}_{nq}$ is an unbiased estimator to S_q and is location-shift invariant. Thus, without loss of generality, we assume $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$ in the rest of the paper.

3. Main Results

In this section, we establish the limiting distribution of $\hat{\mathrm{S}}_{nq}$ that target the super-diagonals of the covariance matrices, followed by an analysis on the power of the test.

We first establish the mean and the variance of $\hat{\mathrm{S}}_{nq}$ defined in (2.3). Let $J_{i,q} = (\sigma_{i,1,(1+q)}, \ldots, \sigma_{i,(p-q),p})^\top$, for $i = 1, 2$, $J_q = J_{1,q} - J_{2,q}$, and

$$V_{nq}^2 = \frac{4}{n_1} J_q^\top W_{1,q} J_q + \frac{4}{n_2} J_q^\top W_{2,q} J_q + \frac{2}{n_1(n_1-1)} \mathrm{tr}(W_{1,q}^2) + \frac{2}{n_2(n_2-1)} \mathrm{tr}(W_{2,q}^2) + \frac{4}{n_1 n_2} \mathrm{tr}(W_{1,q} W_{2,q}). \tag{3.1}$$

Proposition 1. *Under the Assumptions* (C1) - (C5), *for* $q = 0, 1, \ldots, p-1$,

$$\mathrm{E}(\hat{\mathrm{S}}_{nq}) = \mathrm{S}_q \quad \textit{and} \quad \mathrm{Var}(\hat{\mathrm{S}}_{nq}) = V_{nq}^2 + o\left(V_{nq}^2\right).$$

Under $H_{0,q}$, $\mathrm{E}(\hat{\mathrm{S}}_{nq}) = 0$ and $J_q = 0$. Hence, Proposition 1 implies that the leading order variance V_{nq}^2 is

$$V_{0,nq}^2 = \frac{2}{n_1(n_1-1)}\mathrm{tr}(W_{1,q}^2) + \frac{2}{n_2(n_2-1)}\mathrm{tr}(W_{2,q}^2) + \frac{4}{n_1 n_2}\mathrm{tr}(W_{1,q}W_{2,q}).$$

In order to formulate a test procedure, we need to estimate $V_{0,nq}^2$. Motivated by Chen, Zhang and Zhong (2010), we consider the estimator

$$\hat{V}_{0,nq}^2 = \frac{2}{n_1(n_1-1)}\mathrm{R}_{1,nq} + \frac{2}{n_2(n_2-1)}\mathrm{R}_{2,nq} + \frac{4}{n_1 n_2}\mathrm{R}_{c,nq}, \tag{3.2}$$

where

$$\mathrm{R}_{i,nq} = \frac{1}{P_{n_i}^2}\sum_{j,k}^{*}\left\{\hat{\mathbf{Y}}_{i,j}(q)^\top \hat{\mathbf{Y}}_{i,k}(q)\right\}^2 \text{ for } i = 1 \text{ and } 2, \text{ and}$$

$$\mathrm{R}_{c,nq} = \frac{1}{n_1 n_2}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}\sum_{s_1,s_2=1}^{p-q}\left(\hat{\mathrm{Y}}_{1,j}^{s_1,s_1+q}\hat{\mathrm{Y}}_{1,j}^{s_2,s_2+q}\right)\left(\hat{\mathrm{Y}}_{2,k}^{s_1,s_1+q}\hat{\mathrm{Y}}_{2,k}^{s_2,s_2+q}\right)$$

with $\hat{\mathbf{Y}}_{i,j}(q) = \left(\hat{\mathrm{Y}}_{i,j}^{1,1+q}, \ldots, \hat{\mathrm{Y}}_{i,j}^{(p-q),p}\right)^\top$, $\hat{\mathrm{Y}}_{i,j}^{s,s+q} = (\mathrm{X}_{i,j,s} - \bar{\mathrm{X}}_{i,s})(\mathrm{X}_{i,j,s+q} - \bar{\mathrm{X}}_{i,s+q}) - \hat{\sigma}_{i,s,s+q}$, and $\hat{\sigma}_{i,s,s+q}$ is the sample covariance based on the ith population. The subtraction of the sample mean $\bar{\mathrm{X}}_{i,s}$ and $\bar{\mathrm{X}}_{i,s+q}$ in the definition of $\hat{\mathrm{Y}}_{i,j}^{s,s+q}$ maintains the consistency of $\hat{V}_{0,nq}^2$ to $V_{0,nq}^2$ when $\boldsymbol{\mu}_i \neq 0$. The consistency of $\hat{V}_{0,nq}^2$ is established in the following proposition, which is valid beyond the null hypothesis.

Proposition 2. *Under Assumptions* (C1) - (C5), *for* $q = o(p)$, $\hat{V}_{0,nq}^2 / V_{0,nq}^2 \xrightarrow{p} 1$ *as* $n \to \infty$ *and* $p \to \infty$.

Theorem 1. *Under the Assumptions* (C1) - (C5), *for* $q = o(p)$, *as* $n \to \infty$ *and* $p \to \infty$,

$$V_{nq}^{-1}\left(\hat{\mathrm{S}}_{nq} - \mathrm{S}_q\right) \xrightarrow{d} \mathcal{N}(0,1).$$

The theorem only covers the case of $q = o(p)$. For q being a larger order of $o(p)$, for instance $q/p \to c \in (0,1)$, conditions (A.29) and (A.30) used in establishing the martingale central limit theorem (given in the supplementary document) for $\hat{\mathrm{S}}_{nq}$ cannot be guaranteed. This means that it is uncertain whether the asymptotic normality in Theorem 1 is valid. Given this reality, we consider testing over N super-diagonals where $N \to \infty$ gradually so that $N = o(p)$ is maintained.

Theorem 1 indicates that under $\mathrm{H}_{0,q}$, $\hat{\mathrm{S}}_{nq}/\hat{V}_{0,nq} \xrightarrow{d} \mathcal{N}(0,1)$ as $n, p \to \infty$, according to Theorem 1 and Proposition 2. This facilitates a test at a nominal α level of significance which rejects $\mathrm{H}_{0,q}$ if

$$\hat{\mathrm{S}}_{nq} > z_{1-\alpha}\hat{V}_{0,nq}, \quad \text{for} \quad q = o(p), \tag{3.3}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of $\mathcal{N}(0,1)$.

Next, we evaluate the power of the test since the asymptotic normality in Theorem 1 is also valid under $\mathrm{H}_{1,q} : \mathrm{S}_q > 0$. Let $\beta_{np,q}(\alpha)$ denote the power of the test given in (3.3), and $\delta_{np,q} = \mathrm{S}_q/V_{nq}$. Since S_q represents the signal strength on the q-th super-diagonal and V_{nq} is the level of noise of $\hat{\mathrm{S}}_{nq}$, $\delta_{np,q}$ can be viewed as the signal to noise ratio for the individual test problem. Under $\mathrm{H}_{1,q}$, we have

$$\beta_{np,q}(\alpha) = \Pr\left(\frac{\hat{\mathrm{S}}_{nq} - \mathrm{S}_q}{V_{nq}} > z_{1-\alpha}\frac{\hat{V}_{0,nq}}{V_{nq}} - \delta_{np,q}\right).$$

According to (3.1), $V_{nq}^2 \geq V_{0,nq}^2$ for large n since $W_{1,q}$ and $W_{2,q}$ are nonnegative definite. Hence,

$$\beta_{np,q}(\alpha) \geq \Pr\left(\frac{\hat{\mathrm{S}}_{nq} - \mathrm{S}_q}{V_{nq}} > z_{1-\alpha}\frac{\hat{V}_{0,nq}}{V_{0,nq}} - \delta_{np,q}\right). \tag{3.4}$$

Theorem 2. *Under Assumptions* (C1) - (C5) *and* $\mathrm{H}_{1,q} : \mathrm{S}_q > 0$*, for* $q = o(p)$*,*

$$\liminf_{n,p\to\infty} \beta_{np,q}(\alpha) \geq 1 - \Phi\left(z_{1-\alpha} - \liminf_{n,p\to\infty} \delta_{np,q}\right).$$

The signal to noise ratio $\delta_{np,q}$ is a key quantity in determining the power of the individual test in Theorem 2. If $\delta_{np,q}$ diverges to infinity, the power of the test converges to one, implying the consistency of the test. To quantify $\delta_{np,q}$, we first specify the order of V_{nq}^2 under three regimes.

(i) $V_{nq}^2 \asymp (p-q)/n$, for $q = o(n^{1/\beta})$;

(ii) $V_{nq}^2 \asymp (p-q)/n$, for q such that $q^{-\beta} \asymp 1/n$; and

(iii) $V_{nq}^2 \asymp (p-q)/n^2$, for q such that $n^{1/\beta} = o(q)$ and $q = o(p)$.

Detailed analysis for the order of V_{nq}^2 can be found in the proof of Proposition 1 in the supplementary document.

With $h_q = \mathrm{S}_q/(p-q)$ as the average signal strength on the q-th super-diagonal, due to different orders of V_{nq}^2 it can be shown that

$$\delta_{np,q} \asymp \begin{cases} n^{1/2}(p-q)^{1/2}h_q & \text{for } q \text{ in Regimes (i) and (ii);} \\ n(p-q)^{1/2}h_q & \text{for } q \text{ in Regime (iii).} \end{cases} \tag{3.5}$$

According to (3.5) and Theorem 2, the test has nontrivial power if h_q is at least of order $n^{-1/2}(p-q)^{-1/2}$ in Regimes (i) and (ii), and $n^{-1}(p-q)^{-1/2}$ in Regimes (iii). The multipliers $n^{1/2}(p-q)^{1/2}$ and $n(p-q)^{1/2}$ in these two regimes of q compensate for the diminishing signal h_q as q increases, and thus maintain the consistency of the test. They indicate that, for a given signal level h_q, not only larger n but also larger p are beneficial to the power, since they lead to larger $\delta_{np,q}$. The simulation study in Section 4 provides numerical confirmation to this observation.

In the following we use the individual super-diagonal test to facilitate a test for the overall hypothesis $\mathrm{H}_0 : \Sigma_1 = \Sigma_2$. This is done by testing simultaneously the first N super-diagonals, where $N = o(p)$ gradually diverges. The simultaneous testing is carried out by controlling the false discovery rate (FDR). The reason for choosing $N = o(p)$ is that asymptotic normality of $\hat{\mathrm{S}}_{nq}$ is only guaranteed for $q = o(p)$. As N is allowed to diverge, the test covers more and more super-diagonals, asymptotically. Empirically, we can choose $N = \lfloor Cp^\theta \rfloor$ for a $\theta \in (0,1)$, where $\lfloor x \rfloor$ denotes integer truncation and C is a constant such that $Cp^\theta < p$. We employ this approach in selecting N in the simulation and the case study sections later in the paper.

Considering the dependence among $\hat{\mathrm{S}}_{nq}$ at different super-diagonals, we employ the method proposed in Storey, Taylor and Siegmund (2004) to control the FDR that extended the approach of Benjamini and Hochberg (1995) to accommodate more general types of dependence. Let $p_1, \ldots, p_N$ be the p-values corresponding to the hypotheses $\mathrm{H}_{0,1}, \ldots, \mathrm{H}_{0,N}$, respectively. Let $\mathcal{S}$ be the index set of the super-diagonals which are included in the multiple testing, $\mathcal{N}_0 = \{q \in \mathcal{S} : S_q = 0\}$ and $\mathcal{N}_1 = \{q \in \mathcal{S} : S_q \neq 0\}$. For $t \in [0,1]$, take $V(t) = \#\{j \in \mathcal{N}_0 : p_j \leq t\}$ to be the number of false discoveries and $R(t) = \#\{j : p_j \leq t\}$ to be the total number of rejected null hypotheses. Then, given a tuning parameter $\lambda \in [0,1)$, an estimate of the false discovery rate at a significance threshold t proposed in Storey, Taylor and Siegmund (2004) is

$$\widehat{\mathrm{FDR}}_\lambda(t) = \frac{\hat{\pi}_0(\lambda)t}{\{R(t) \vee 1\}/N},$$

where $R(t) \vee 1 = \max\{R(t), 1\}$ and $\hat{\pi}_0(\lambda) = \{N - R(t)\}/\{(1-\lambda)N\}$ is an estimate of the proportion of true nulls and $N = \#\{j : j \in \mathcal{S}\}$.

Given α and λ, $\mathrm{H}_{0,q} : \mathrm{S}_q = 0$ is rejected if its corresponding p-value p_q is less than or equal to $t_\alpha(\widehat{\mathrm{FDR}}_\lambda) = \sup(0 \leq t \leq 1 : \widehat{\mathrm{FDR}}_\lambda(t) \leq \alpha)$. In addition, if the $\hat{\mathrm{S}}_{nq}$ are positively correlated, we can employ the Benjamini and Hochberg (1995) procedure. Specifically, it arranges the p-values in the ascending order

$p_{(1)} \leq \cdots \leq p_{(N)}$ for $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(N)}$ the corresponding null hypotheses, and we reject $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(K)}$ where $K = \max_j\{p_{(j)} \leq j\alpha/N\}$. Storey, Taylor and Siegmund (2004) has shown that their method is equivalent to the Benjamini and Hochberg (1995) procedure by setting $\lambda = 0$. Benjamini and Hochberg's procedure is more conservative under the assumption of independent and uniformly distributed p-values, while the same is not necessarily the case for dependent case. The empirical powers of the proposed test with the Storey, Taylor and Siegmund (2004) multiple test procedure and Benjamini and Hochberg (1995) procedure are reported in the simulation study in Section 4 and the supplementary document, which show that both procedures produced largely similar test performance.

4. Simulation Results

In this section, we study the numerical performance of the proposed test and compare it with the tests proposed by Srivastava and Yanagihara (2010), Li and Chen (2012) and Cai, Liu and Xia (2013). The latter three tests are denoted respectively by SY, LC and CLX.

We first considered banded covariances where $\Sigma_i = \{\sigma_{i,k,\ell}\mathrm{I}(|k-\ell| \leq s)\}_{p\times p}$ for a bandwidth s. Under $\mathrm{H}_0 : \Sigma_1 = \Sigma_2$, we generated i.i.d. random vectors $\{\mathbf{X}_{i,j}\}_{j=1}^{n_i}$, $i = 1, 2$, from a moving average (MA) model of order 5:

$$\mathrm{X}_{i,j,k} = \mathrm{Z}_{i,j,k} + \sum_{\ell=1}^{5} 0.4\mathrm{Z}_{i,j,k-\ell}, \tag{4.1}$$

where $\{\mathrm{Z}_{i,j,k}\}_{k=1}^{p}$, $i = 1, 2$, $j = 1, \ldots, n_i$ were i.i.d. random variables from $\mathcal{N}(0, 1)$ or the standardized Gamma distribution $\mathcal{G}(1, 0.5)$ with zero mean and unit variance. Under (4.1), Σ_1 and Σ_2 are both banded with the bandwidth $s = 5$. To evaluate the power, we generated the first sample from model (4.1) and the second from the MA(4) model

$$\mathrm{X}_{i,j,k} = \mathrm{Z}_{i,j,k} + \sum_{\ell=1}^{4} 0.4\mathrm{Z}_{i,j,k-\ell}. \tag{4.2}$$

Under (4.2), the covariance matrix Σ_2 is banded with the bandwidth $s = 4$. Thus, the $\Sigma_1 - \Sigma_2$ matrix is banded with $\mathrm{S}_q \neq 0$ for only $q = 0, 1, \ldots, 5$. The average signals $h_q = \mathrm{S}_q/(p-q)$ are plotted in the top left panel of Figure 5.

Then we considered covariance matrices whose elements decay at an exponential rate. Under H_0, $\mathbf{X}_{i,j} = \Gamma_i\mathbf{Z}_{i,j}$ with $\Gamma_i\Gamma_i^\top = \Sigma_i = (\sigma_{i,k,\ell})_{p\times p}$, $i = 1, 2$,

where

$$\sigma_{i,k,\ell} = 4\mathrm{I}(k=\ell) + \exp\left(-\frac{|k-\ell|}{20}\right), \text{ for } k,\ell = 1,\ldots,p. \tag{4.3}$$

We generated i.i.d. $\mathbf{Z}_{i,j}$ from the Gaussian and the Gamma distributions. The covariance model (4.3) is commonly used in such spatial statistical models as Cressie (1993), Lee and Yu (2010) and Rodríguez and Bárdossy (2014).

To evaluate the power, we first considered a banded $\Sigma_1 - \Sigma_2$ matrix with the first ten super-diagonals nonzero and decaying at a polynomial rate. Let $U = (u_{k,\ell})_{p\times p}$ with

$$u_{k,\ell} = |k-\ell|^{-0.2}\mathrm{I}(1 \le |k-\ell| \le 10), \text{ for } k,\ell = 1,\ldots,p. \tag{4.4}$$

Define Σ_1^* according to (4.3) and $\delta = |\min\{\lambda_{\min}(\Sigma_1^*), \lambda_{\min}(\Sigma_1^* + U)\}| + 0.05$, needed below to maintain the positive definiteness of covariance matrices. Let $\Sigma_1 = \Sigma_1^* + \delta I$ and $\Sigma_2 \hat{=} \Sigma_1^* + U + \delta I$. We generated $\mathbf{X}_{1,j} = \Gamma_1 \mathbf{Z}_{1,j}$ with $\Gamma_1\Gamma_1^\top = \Sigma_1$ and $\mathbf{X}_{2,j} = \Gamma_2 \mathbf{Z}_{2,j}$ with $\Gamma_2\Gamma_2^\top = \Sigma_2$. Adding δI to both covariance matrices ensures that $\Sigma_2 - \Sigma_1 = U$. The average signals h_q are shown in the top right panel of Figure 5.

Another simulation setting had Σ_1 and Σ_2 different in consecutive super-diagonal blocks exhibiting a "wave-like" super-diagonal structure. In this setup, Let $U = (u_{k,\ell})_{p\times p}$ be symmetric with

$$u_{k,k+q} = q^{-1.2\times\omega_q}\mathrm{I}(q \in \mathcal{K}), \text{ for } q = 0,\ldots,p-1, k = 1,\ldots,p-q, \tag{4.5}$$

with $\mathcal{K}$ the index set of super-diagonals where Σ_1 and Σ_2 are different, and $\omega_q \overset{i.i.d.}{\sim} Unif(0,1)$ for $q \in \mathcal{K}$ is used to allow for different decay rates. For $p = 50$, $\mathcal{K} = \{1,\ldots,5,11,\ldots,15\}$; for $p = 100$, $\mathcal{K} = \{1,\ldots,5,11,\ldots,15,\ 21,\ldots,25\}$; and for p even larger, $\mathcal{K} = \{1,\ldots,5,\ 11,\ldots,15,\ 21,\ldots,25,\ 31,\ldots,\lfloor 0.5p^{0.7}\rfloor\}$. Still, the first sample was generated with $\Sigma_1 = \Sigma_1^* + \delta I$, Σ_1^* defined in (4.3), and the second with $\Sigma_2 = \Sigma_1^* + U + \delta I$. By this definition, $\Sigma_2 - \Sigma_1 = U$ is denser than the models (4.1) to (4.4), in the sense that half of the super-diagonals of $\Sigma_1 - \Sigma_2$ we considered are nonzero. For fair comparison, we only generated ω_q once and used it throughout all (n,p)-settings to maintain the same signal strengths. Figure 5 gives the average signal h_q, which shows that h_q is not necessarily monotone decreasing as q grows.

We took $n_1 = n_2 = n$. To make p and n increase simultaneously, we took p from 50 to 1,000 and considered three sample sizes for each p in all simulation setups: $n = 30, 50, 80$ for $p = 50$; $n = 50, 80, 100$ for $p = 100$; $n = 80, 100, 120$ for $p = 200$; $n = 100, 120, 150$ for $p = 400$; $n = 120, 150, 180$ for $p = 600$; and $n = 150, 180, 200$ for $p = 1{,}000$. Multiple testing procedures were implemented

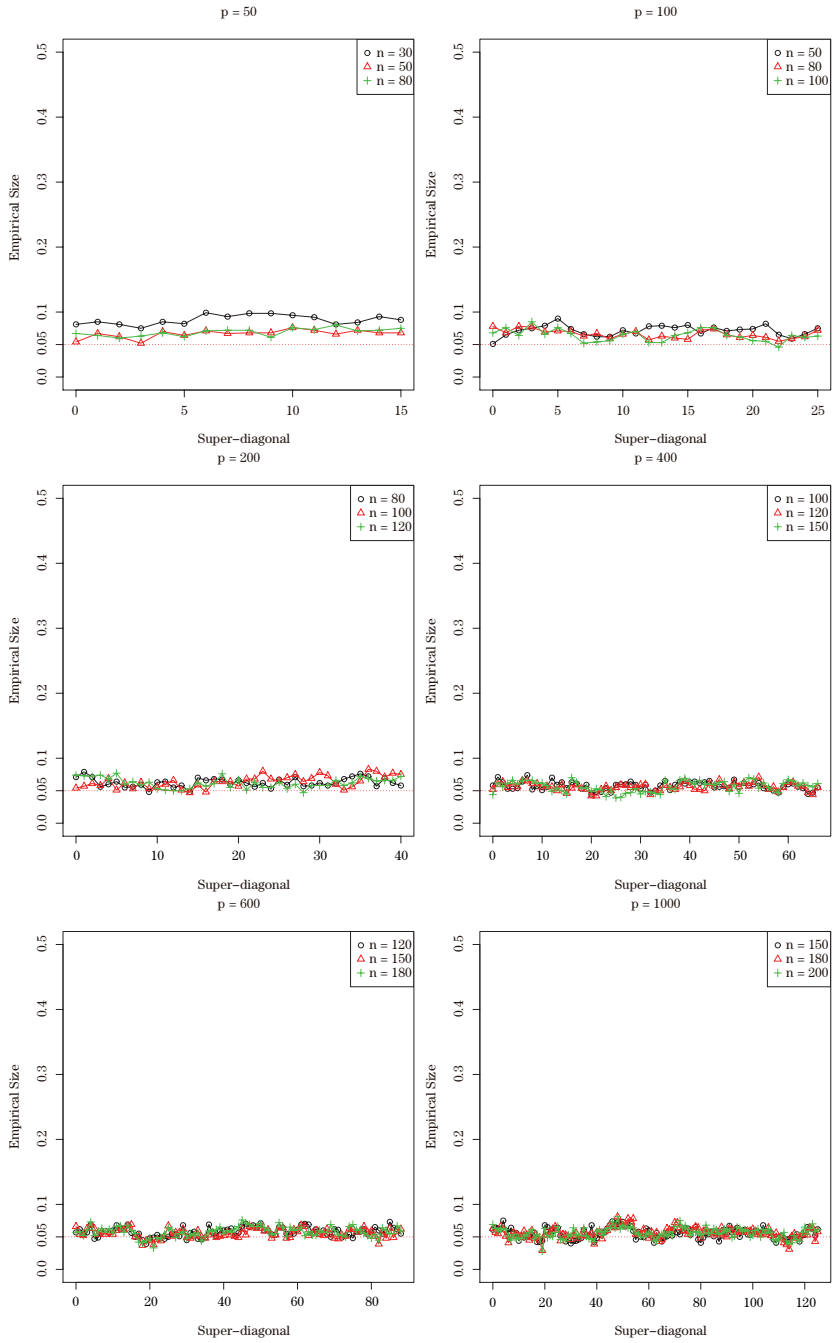

Figure 1. Empirical sizes of the individual tests $H_{0,q} : S_q = 0$ for Gaussian distributed data generated from model (4.1). The range of the horizontal axis is from $q = 0$ to $q = \lfloor p^{0.7} \rfloor$.

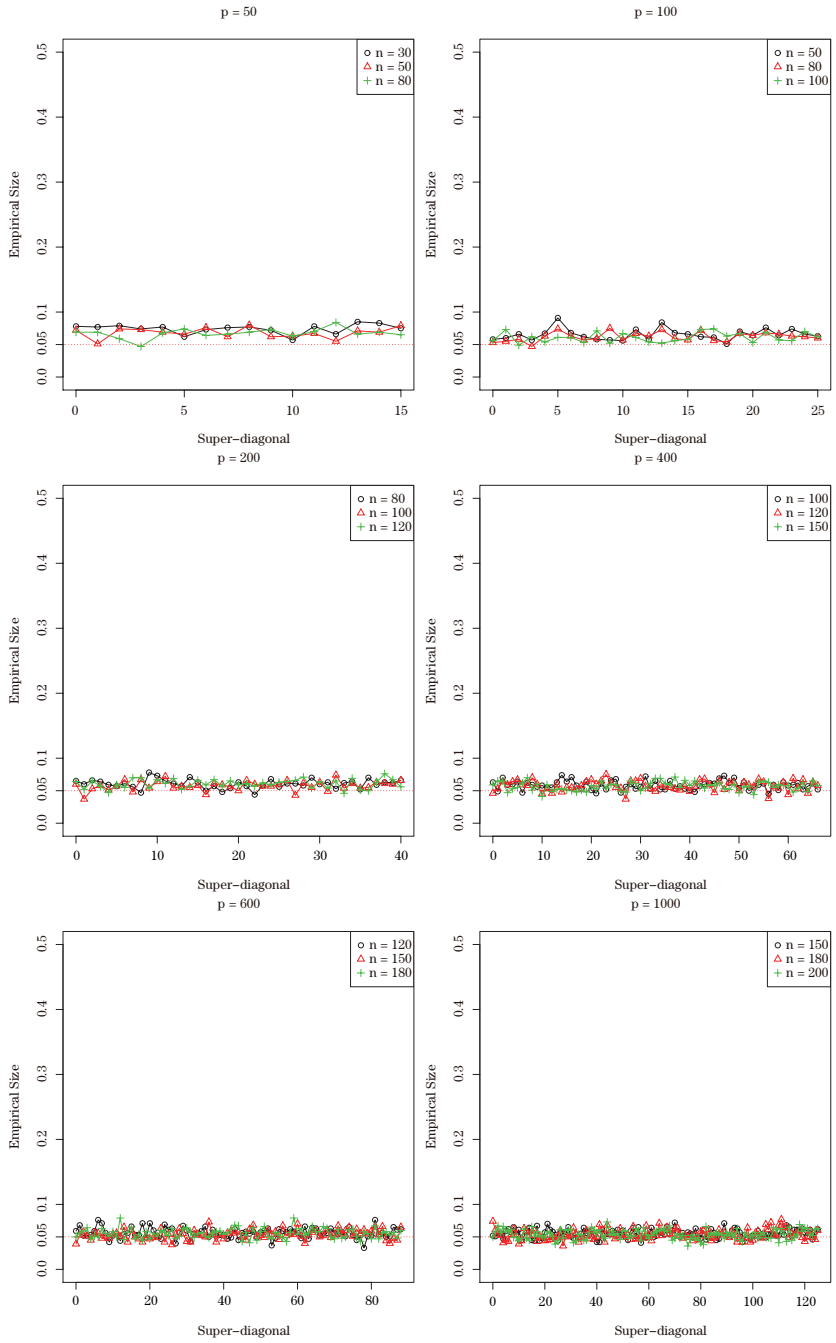

Figure 2. Empirical sizes of the individual tests $H_{0,q} : S_q = 0$ for Gaussian distributed data generated from model (4.3). The range of the horizontal axis is from $q = 0$ to $q = \lfloor p^{0.7} \rfloor$.

for $H_{0,q} : S_q = 0$, $q = 0, \ldots, N$, where $N = \lfloor p^{0.7} \rfloor$, using the Storey, Taylor and Siegmund (2004) procedure with $\lambda = 0.5$ and controlling the FDR at 0.05. All simulation results were based on 1,000 replications. Due to a limited space, we only report simulation results of the proposed test for the Gaussian data; those for the Gamma-distributed innovations are reported in the supplementary document.

Figures 1 and 2 display the empirical sizes of the individual tests $H_{0,q} : S_q = 0$ for q from 0 to $\lfloor p^{0.7} \rfloor$ for $\mathbf{X}_{i,j}$ generated from models (4.1) and (4.3), respectively. The empirical sizes were close to the nominal level 0.05 as n and p increase simultaneously in both models, which indicates that the asymptotic normality established in Theorem 1 is a good approximation for the null distribution of $\hat{S}_{nq}$. The individual empirical powers are shown in Figures 3 and 4 for models (4.2) and (4.4), respectively. Larger powers in Figures 3 and 4 were observed on the super-diagonals with larger h_q shown in Figure 5, and they increased rapidly and converged to one as n and p grew. Thus, these results are consistent with the theoretical analyses conveyed in Theorems 1 and 2.

Tables 1 and 2 provide the empirical sizes and powers of testing $H_0 : \Sigma_1 = \Sigma_2$ for models (4.2) and (4.4) respectively. For the proposed test, H_0 is rejected if there exists any $H_{0,q} : S_q = 0$ rejected in the multiple test procedure. It is observed that the empirical sizes of the proposed method were larger than 0.05 when n and p were both small, due to the error in the asymptotic distribution for small samples. This was in accordance with the findings in Figure 1 and 2 that the individual empirical sizes were also above 0.05 when $p = 50$ and 100. As n and p grow simultaneously, the sizes of the proposed test were close to 0.05. Most of the sizes were smaller than 0.05 because the FDR controlling methods tend to be conservative, as it controls the FDR rather than the familywise error rate. To alleviate any advantage of the proposed test due to having a larger size, we re-adjusted the nominal level for each (p, n)-setting so that the size of the proposed test was the smallest among the four tests considered in the simulation experiments. The adjusted sizes and powers are reported in the parentheses in Tables 1 and 2. As n and p increased simultaneously, the empirical sizes of the proposed test became much more accurate, thus the adjustment had become smaller. Although the sizes were the smallest, the proposed test had consistently higher powers than the other three tests. The improvement in the power was substantial in most cases.

We observe that in both Tables 1 and 2 the powers of CLX's test decreased as the dimension p grew. This is because the number of nonzero elements in $\Sigma_1 - \Sigma_2$

Table 1. Empirical sizes and powers of the proposed test in conjunction with the Storey, Taylor and Siegmund (2004) procedure and the tests of Srivastava and Yanagihara (2010) (SY), Li and Chen (2012) (LC) and Cai, Liu and Xia (2013) (CLX) for Gaussian distributed data generated from model (4.1) for size and (4.2) for power. The multiple test procedure is conducted by controlling the FDR at $\alpha = 0.05$ and $N = \lfloor p^{0.7} \rfloor$. The figures in the parentheses are the adjusted empirical sizes and powers so that the empirical sizes are smaller than the other three tests.

		Empirical Size				Empirical Power			
p	n	SY	LC	CLX	Proposed	SY	LC	CLX	Proposed
50	30	0.068	0.072	0.053	0.108(0.044)	0.250	0.158	0.271	0.679(0.499)
	50	0.059	0.060	0.041	0.099(0.035)	0.423	0.219	0.288	0.906(0.694)
	80	0.054	0.058	0.044	0.090(0.042)	0.563	0.391	0.405	0.993(0.930)
100	50	0.073	0.053	0.056	0.069(0.042)	0.502	0.240	0.269	0.989(0.921)
	80	0.049	0.055	0.048	0.064(0.045)	0.746	0.412	0.349	1.000(0.990)
	100	0.049	0.059	0.047	0.068(0.046)	0.854	0.520	0.409	1.000(1.000)
200	80	0.047	0.054	0.041	0.066(0.041)	0.853	0.426	0.301	1.000(1.000)
	100	0.046	0.050	0.057	0.063(0.046)	0.896	0.531	0.381	1.000(1.000)
	120	0.045	0.055	0.048	0.061(0.044)	0.919	0.698	0.439	1.000(1.000)
400	100	0.048	0.052	0.041	0.042(0.041)	0.957	0.555	0.330	1.000(1.000)
	120	0.049	0.044	0.034	0.058(0.030)	0.990	0.695	0.388	1.000(1.000)
	150	0.047	0.049	0.040	0.059(0.036)	0.999	0.841	0.523	1.000(1.000)
600	120	0.044	0.044	0.047	0.040(0.040)	1.000	0.670	0.364	1.000(1.000)
	150	0.041	0.044	0.043	0.055(0.040)	1.000	0.860	0.497	1.000(1.000)
	180	0.053	0.038	0.041	0.041(0.035)	1.000	0.941	0.608	1.000(1.000)
1,000	150	0.061	0.047	0.045	0.044(0.044)	1.000	0.852	0.383	1.000(1.000)
	180	0.056	0.048	0.040	0.043(0.038)	1.000	0.954	0.513	1.000(1.000)
	200	0.046	0.048	0.039	0.050(0.039)	1.000	0.975	0.598	1.000(1.000)

gets large as p increases in Models (4.2) and (4.4). The CLX's test is based on the maximal norm type statistic and tends to be more powerful when Σ_1 and Σ_2 differ only in a small proportion of elements with relative high magnitude. The proposed test, as well as the SY and LC tests, had increasing empirical powers as n and p increased. The reason for SY and LC tests having lower power than the proposed test is due to the aggregation of more noise from the whole covariance matrices Σ_1 and Σ_2 in the variance of their test statistics. The larger variance of the test statistic reduced the signal to noise ratio, and hence reduced the power of the test. The powers of the proposed test were significantly larger than the other three tests since we have larger signal to noise ratios.

Table 3 reports the empirical FDR and the Correct Rejectionc Rates (CRRs) of the proposed test under models (4.1) and (4.2). Table 4 reports the results for models (4.3) and (4.4). The empirical FDRs were controlled to be under 0.05 in

Table 2. Empirical sizes and powers of the proposed test in conjunction with the Storey, Taylor and Siegmund (2004) procedure and the tests of SY, LC and CLX for Gaussian distributed data generated from model (4.3) for size and (4.4) for power. The multiple test procedure is conducted by controlling the FDR at $\alpha = 0.05$ and $N = \lfloor p^{0.7} \rfloor$. The figures in the parentheses are the adjusted empirical sizes and powers so that the empirical sizes are smaller than the other three tests.

		Empirical Size				Empirical Power			
p	n	SY	LC	CLX	Proposed	SY	LC	CLX	Proposed
50	30	0.047	0.057	0.067	0.108(0.041)	0.495	0.372	0.248	0.609(0.498)
	50	0.063	0.069	0.048	0.099(0.043)	0.711	0.589	0.256	0.814(0.729)
	80	0.045	0.081	0.059	0.092(0.041)	0.849	0.858	0.339	0.952(0.925)
100	50	0.058	0.059	0.044	0.069(0.041)	0.806	0.637	0.234	0.926(0.900)
	80	0.051	0.063	0.051	0.061(0.047)	0.953	0.905	0.311	0.996(0.994)
	100	0.041	0.042	0.042	0.057(0.040)	0.990	0.966	0.369	0.998(0.996)
200	80	0.043	0.071	0.045	0.065(0.038)	1.000	0.953	0.260	1.000(1.000)
	100	0.043	0.054	0.042	0.071(0.041)	1.000	0.990	0.300	1.000(1.000)
	120	0.049	0.045	0.046	0.068(0.045)	1.000	1.000	0.369	1.000(1.000)
400	100	0.061	0.039	0.061	0.045(0.035)	1.000	0.992	0.253	1.000(1.000)
	120	0.055	0.062	0.038	0.042(0.038)	1.000	1.000	0.297	1.000(1.000)
	150	0.054	0.057	0.040	0.040(0.040)	1.000	1.000	0.437	1.000(1.000)
600	120	0.046	0.049	0.056	0.045(0.045)	1.000	1.000	0.263	1.000(1.000)
	150	0.049	0.048	0.038	0.046(0.038)	1.000	1.000	0.355	1.000(1.000)
	180	0.063	0.057	0.053	0.041(0.041)	1.000	1.000	0.503	1.000(1.000)
1,000	150	0.049	0.054	0.048	0.046(0.046)	1.000	1.000	0.303	1.000(1.000)
	180	0.055	0.053	0.043	0.043(0.043)	1.000	1.000	0.426	1.000(1.000)
	200	0.049	0.047	0.046	0.044(0.044)	1.000	1.000	0.517	1.000(1.000)

most cases. Still, the CRRs increased substantially as n and p grew. It is noted that the CRRs increased more quickly to 1 under model (4.4) than under (4.2). Here, the average signals in (4.4) are relatively larger, as shown in Figure 5. Moreover, the performances of the test, as reflected by the FDRs and the CRRs, were robust with respect to different choices of the constant C in deciding the number of the super-diagonals in the test.

The results for the Gamma-distributed data were largely similar to those for the Gaussian data. Furthermore, since the correlations between the S_q are positive under models (4.1) - (4.4) in our simulation study, the proposed test can also be used in conjunction with the Benjamini and Hochberg (1995) procedure. The corresponding results are in the supplementary document; they are similar to those using the Storey, Taylor and Siegmund (2004) procedure.

The empirical powers under model (4.5) are given in Table 5. Considering that the powers of the SY, LC and the proposed tests were all equal to 1 when

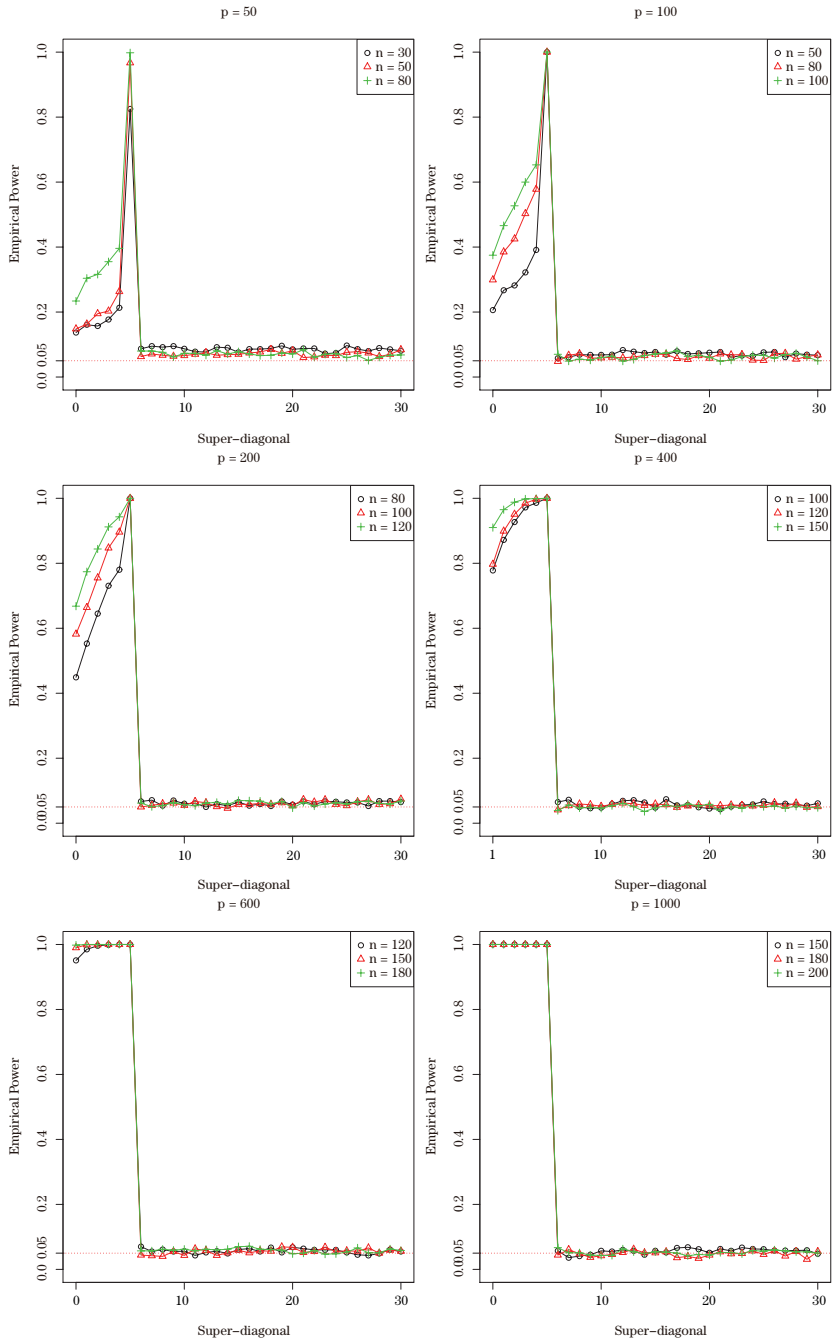

Figure 3. Empirical powers of the individual tests $H_{0,q} : S_q = 0$ for Gaussian distributed data with the first sample generated from model (4.1) while the second sample generated from model (4.2).

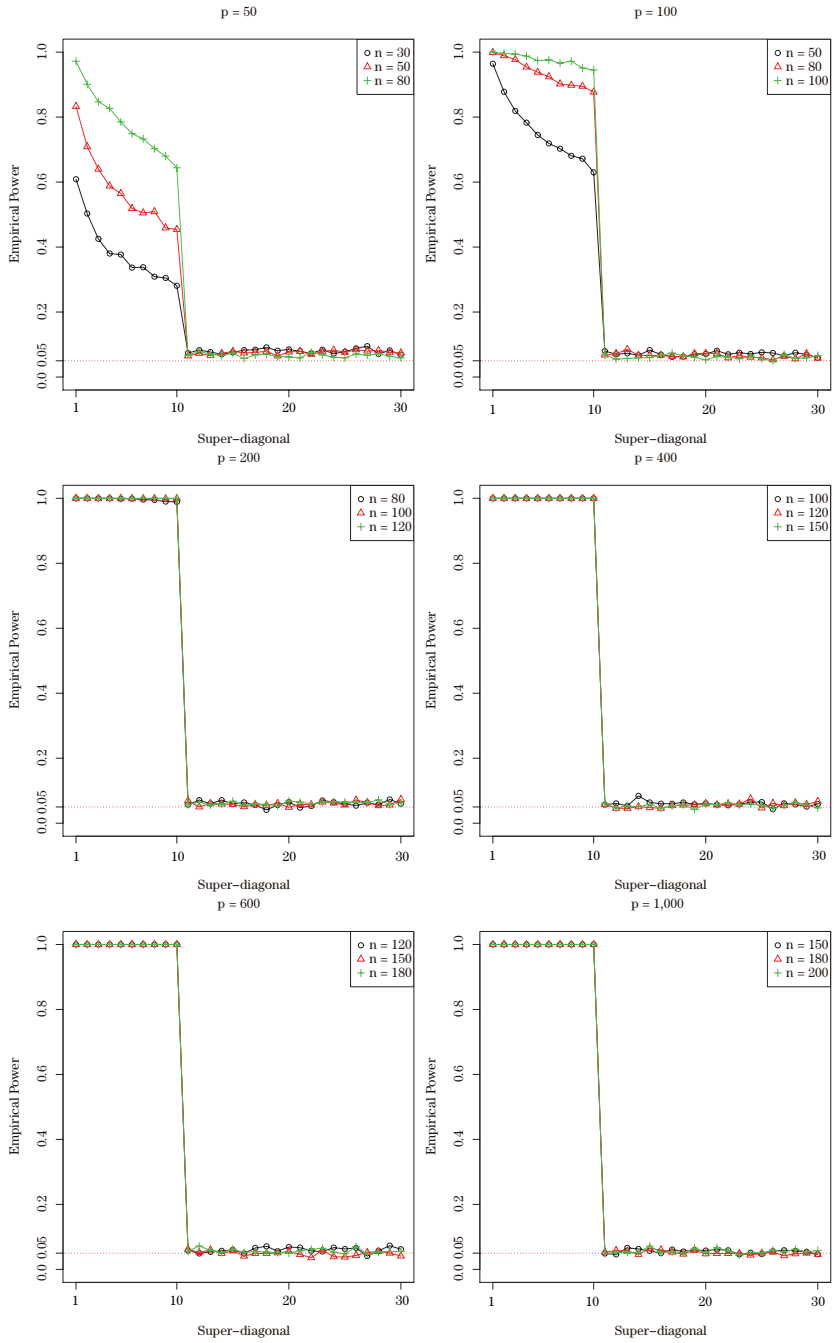

Figure 4. Empirical powers of the individual tests $H_{0,q} : S_q = 0$ for Gaussian distributed data with the first sample generated from model (4.3) while the second sample generated from model (4.4).

Table 3. False discovery rates and correct rejection rates of the proposed test in conjunction with the Storey, Taylor and Siegmund (2004) procedure for Gaussian distributed data where the first sample is generated from model (4.1) and the second generated from model (4.2). The multiple test procedure is performed by controlling the FDR at $\alpha = 0.05$ and $N = \lfloor Cp^{0.7} \rfloor$. Three dimensions are considered for each dimension, that is $n = 30, 50, 80$ for $p = 50$; $n = 50, 80, 100$ for $p = 100$; $n = 80, 100, 120$ for $p = 200$; $n = 100, 120, 150$ for $p = 400$; $n = 120, 150, 180$ for $p = 600$; and $n = 150, 180, 200$ for $p = 1{,}000$, respectively.

False Discovery Rate						Correct Rejection Rate					
p						p					
50	100	200	400	600	1,000	50	100	200	400	600	1,000
$C = 1$						$C = 1$					
0.048	0.057	0.040	0.058	0.058	0.054	0.157	0.249	0.205	0.738	0.934	1.000
0.030	0.036	0.046	0.055	0.057	0.054	0.214	0.267	0.533	0.768	0.967	1.000
0.051	0.056	0.045	0.057	0.058	0.053	0.356	0.397	0.638	0.818	0.994	1.000
$C = 1.5$						$C = 1.5$					
0.063	0.050	0.050	0.045	0.055	0.046	0.134	0.151	0.228	0.686	0.920	1.000
0.043	0.033	0.044	0.052	0.059	0.048	0.194	0.205	0.482	0.703	0.946	1.000
0.054	0.046	0.045	0.056	0.057	0.048	0.328	0.349	0.628	0.867	0.993	1.000
$C = 2$						$C = 2$					
0.053	0.068	0.060	0.044	0.057	0.047	0.147	0.250	0.222	0.618	0.897	1.000
0.043	0.047	0.040	0.058	0.055	0.049	0.203	0.252	0.440	0.746	0.933	1.000
0.051	0.042	0.057	0.051	0.060	0.049	0.312	0.345	0.617	0.882	0.991	1.000
$C = 2.5$						$C = 2.5$					
0.069	0.032	0.057	0.036	0.054	0.040	0.138	0.156	0.217	0.556	0.901	1.000
0.053	0.031	0.032	0.053	0.048	0.041	0.196	0.202	0.443	0.737	0.923	1.000
0.044	0.041	0.047	0.051	0.053	0.040	0.308	0.348	0.567	0.880	0.991	1.000
$C = 3$						$C = 3$					
0.094	0.022	0.058	0.044	0.056	0.047	0.134	0.183	0.251	0.561	0.861	1.000
0.070	0.030	0.036	0.058	0.042	0.046	0.190	0.201	0.484	0.695	0.963	1.000
0.049	0.045	0.048	0.054	0.052	0.044	0.305	0.354	0.501	0.863	0.986	1.000

$p = 600$ and 1,000, we only report the results with p ranging from 50 to 400. In this case, $\Sigma_1 - \Sigma_2$ was denser than the previous models and did not have monotone decreasing signals Table 5 shows that the proposed method was consistently more powerful than the other three tests. The table also reports the empirical FDRs and CRRs of the proposed test, which shows that the FDRs were largely controlled around 5% while the CRRs increased quickly as n and p grow. Since there are extremely small signals for some q, as observed in Figure 5, the CRRs were not as high as those in Tables 3 and 4. This provides some empirical evidence for using the proposed test when $\Sigma_1 - \Sigma_2$ does not have the bandable structure.

Table 4. False discovery rates and correct rejection rates of the proposed test in conjunction with the Storey, Taylor and Siegmund (2004) procedure for Gaussian distributed data where the first sample is generated from model (4.3) and the second generated from model (4.4). The multiple test procedure is performed by controlling the FDR at $\alpha = 0.05$ and $N = \lfloor Cp^{0.7} \rfloor$. Three dimensions are considered for each dimension, that is $n = 30, 50, 80$ for $p = 50$; $n = 50, 80, 100$ for $p = 100$; $n = 80, 100, 120$ for $p = 200$; $n = 100, 120, 150$ for $p = 400$; $n = 120, 150, 180$ for $p = 600$; and $n = 150, 180, 200$ for $p = 1{,}000$, respectively.

False Discovery Rate						Correct Rejection Rate					
p						p					
50	100	200	400	600	1,000	50	100	200	400	600	1,000
$C = 1$						$C = 1$					
0.025	0.037	0.067	0.043	0.057	0.054	0.290	0.601	0.987	1.000	1.000	1.000
0.028	0.047	0.040	0.059	0.052	0.054	0.500	0.900	0.977	1.000	1.000	1.000
0.054	0.028	0.066	0.058	0.059	0.053	0.784	0.944	1.000	1.000	1.000	1.000
$C = 1.5$						$C = 1.5$					
0.056	0.051	0.069	0.043	0.053	0.046	0.212	0.570	0.983	1.000	1.000	1.000
0.042	0.047	0.045	0.062	0.051	0.048	0.406	0.834	0.996	1.000	1.000	1.000
0.063	0.036	0.041	0.055	0.057	0.048	0.702	0.935	0.999	1.000	1.000	1.000
$C = 2$						$C = 2$					
0.069	0.064	0.046	0.045	0.052	0.047	0.189	0.552	0.980	1.000	1.000	1.000
0.063	0.049	0.045	0.053	0.051	0.049	0.391	0.858	0.994	1.000	1.000	1.000
0.064	0.039	0.052	0.058	0.057	0.049	0.668	0.929	0.999	1.000	1.000	1.000
$C = 2.5$						$C = 2.5$					
0.037	0.052	0.046	0.045	0.053	0.040	0.149	0.511	0.976	1.000	1.000	1.000
0.067	0.053	0.046	0.051	0.051	0.041	0.360	0.846	0.994	1.000	1.000	1.000
0.066	0.042	0.053	0.057	0.054	0.040	0.630	0.923	0.999	1.000	1.000	1.000
$C = 3$						$C = 3$					
0.055	0.059	0.048	0.048	0.056	0.047	0.148	0.498	0.974	1.000	1.000	1.000
0.050	0.053	0.066	0.052	0.056	0.047	0.320	0.835	0.993	1.000	1.000	1.000
0.063	0.047	0.056	0.057	0.055	0.044	0.648	0.919	0.999	1.000	1.000	1.000

5. Empirical Study

We considered an application of the proposed test to a prostate cancer dataset from Adam et al. (2002). Blood serum samples were procured from either patients diagnosed with prostate cancer (the cancer group) or age-matched healthy men (the healthy group) and were analyzed in protein mass spectroscopy. For each blood sample i, the intensity $X_{i,j}$ for many time-of-flight values t_j were observed. Time of flight is related to the mass over charge ratio m/z of the constituent proteins in the blood. The order of the intensities are pre-determined by the value of m/z. The widely used mass spectroscopy technology allows one to find m/z-sites that discriminate between the two groups and thus to detect

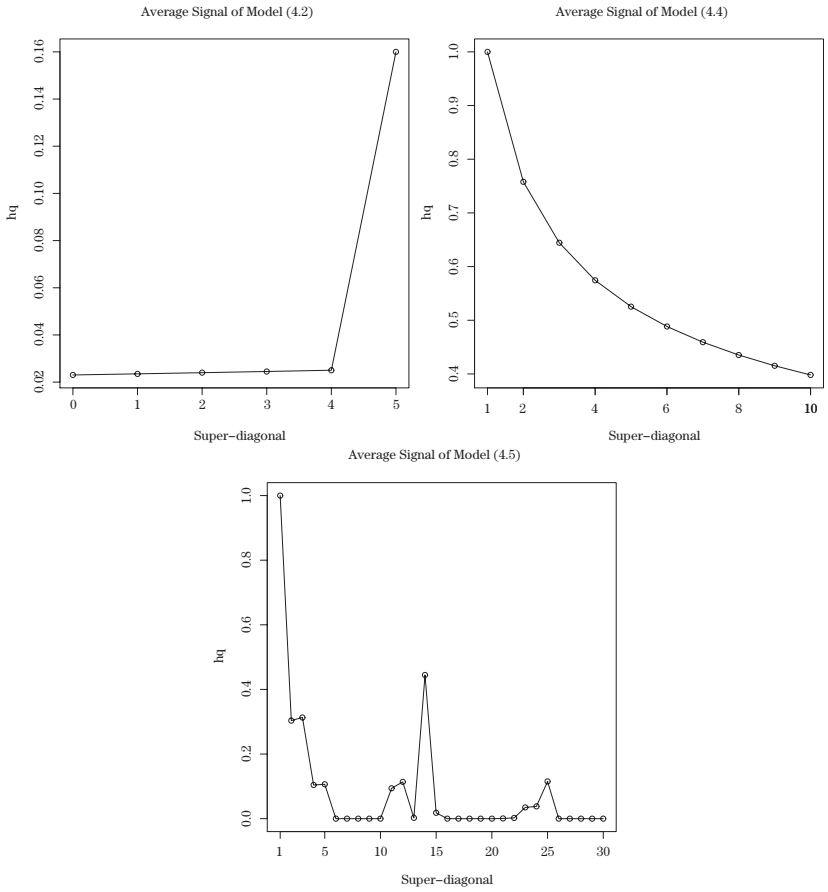

Figure 5. Average signals $h_q = \mathrm{S}_q/(p-q)$ for model (4.2) (the top left panel), model (4.4) (the top right panel) and model (4.5) (the bottom panel). In Model (4.2), $h_q = 0$ for $q > 5$; In Model (4.4), $h_q = 0$ for $q = 0$ and $q > 10$; And in Model (4.5), $h_q = 0$ for $q \notin \mathcal{K} = \{1, \ldots, 5,\ 11, \ldots, 15,\ 21, \ldots, 25,\ 31, \ldots, \lfloor 0.5p^{0.7} \rfloor\}$.

prostate cancer. The full dataset consists of 167 patients in the cancer group and 157 in the healthy group. Following the original researchers, we ignored m/z-sites below 2,000 to avoid chemical artifacts and averaged the data in consecutive blocks of 20 to smooth the intensity profile. This gave a total of 2,181 dimensions, a relatively large number for the sample sizes $n_1 = 157$ and $n_2 = 167$.

Qiu and Chen (2012) carried out additional averaging on this dataset in consecutive blocks of 10, which gave 218 dimensions in total. For the original 2,181-dimensional data we used, the left panel of Figure 6 plots the estimated average signal $\hat{h}_{i,q} = \hat{\mathrm{D}}_{i,nq}/(p-q)$ for $i = 1$ and 2, representing the healthy and the cancer group respectively. It shows that $\hat{h}_{i,q}$ decays rapidly as q increases.

Table 5. Empirical powers of the proposed test in conjunction with the Storey, Taylor and Siegmund (2004) procedure and the tests of SY, LC and CLX for Gaussian distributed data generated from models (4.3) and (4.5). The multiple test procedure is conducted by controlling the FDR at $\alpha = 0.05$ and $N = \lfloor p^{0.7} \rfloor$. The figures in the parentheses are the adjusted empirical powers corresponding to the adjusted empirical sizes in Table 2. The last two columns report the empirical FDR and Correct Rejection Rate (CRR) of the proposed method.

		Empirical Power				the Proposed test	
p	n	SY	LC	CLX	Proposed	FDR	CRR
50	30	0.266	0.187	0.108	0.489(0.374)	0.057	0.124
	50	0.407	0.300	0.212	0.742(0.625)	0.056	0.180
	80	0.607	0.514	0.236	0.935(0.870)	0.057	0.201
100	50	0.557	0.332	0.197	0.905(0.862)	0.048	0.153
	80	0.767	0.559	0.204	0.996(0.989)	0.047	0.239
	100	0.872	0.690	0.359	0.999(0.996)	0.049	0.279
200	80	0.948	0.607	0.292	1.000(1.000)	0.051	0.328
	100	0.982	0.736	0.323	1.000(1.000)	0.050	0.373
	120	0.994	0.866	0.364	1.000(1.000)	0.048	0.432
400	100	0.999	0.746	0.294	1.000(1.000)	0.051	0.443
	120	1.000	0.953	0.313	1.000(1.000)	0.055	0.506
	150	1.000	0.995	0.392	1.000(1.000)	0.056	0.577

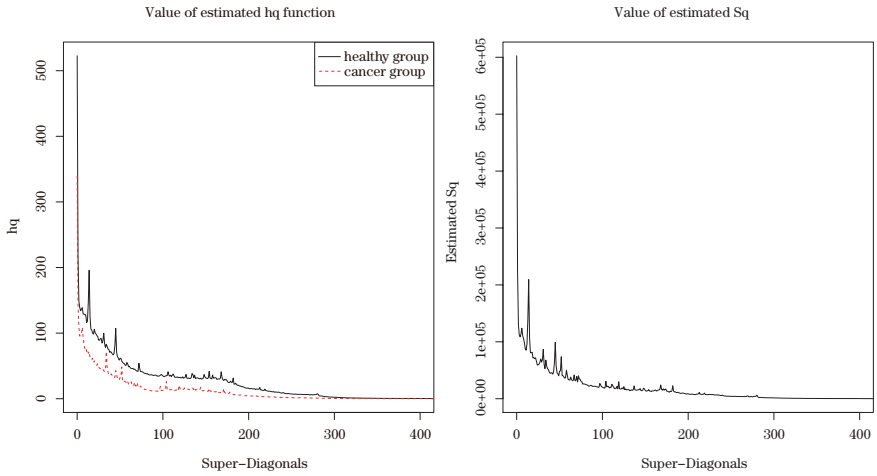

Figure 6. Estimated average signals $\hat{h}_{i,q} = \hat{\mathrm{D}}_{i,nq}/(p-q)$ for the healthy and the cancer groups (left panel) for different q and the estimated $\hat{\mathrm{S}}_{nq}$ representing the signals on super-diagonals of $\Sigma_1 - \Sigma_2$ (right panel).

This agrees with the finding in Qiu and Chen (2012) that the elements of Σ_1 and Σ_2 decay as they move away from the main diagonal, and is eligible for the

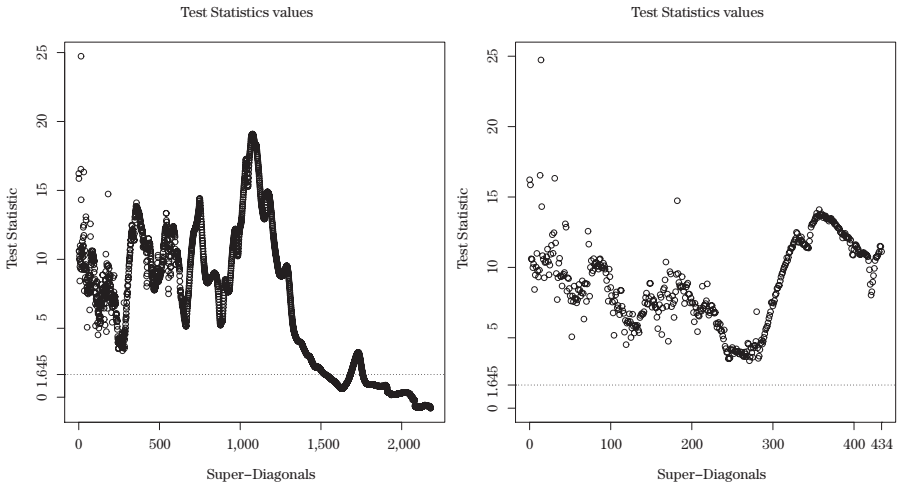

Figure 7. Standardized test statistics $\hat{\mathrm{S}}_{nq}/\hat{V}_{0,nq}$ for q ranging from 0 to $(p-1)$ (the left panel) and for q from 0 to $N = \lfloor 2p^{0.7} \rfloor = 434$ (the right panel). In both panels, the black dotted horizontal line is the critical value $z_{0.95} = 1.645$ for individual tests at 5% level of significance.

bandable assumption. Meanwhile, $\Sigma_1 - \Sigma_2$ appears to obey a bandable structure from the right panel of Figure 6.

The proposed test was applied to test $\mathrm{H}_0 : \Sigma_1 = \Sigma_2$ by implementing the multiple test $\mathrm{H}_{0,q} : \mathrm{S}_q = 0$ for $q = 0, \ldots, N$, and controlling the FDR at 0.05. Figure 7 provides the values of the test statistics $\hat{\mathrm{S}}_{nq}/\hat{V}_{0,nq}$ for q ranging from 0 to $(p-1)$. To show more clearly when the signal is relatively stronger, we plotted $\hat{\mathrm{S}}_{nq}/\hat{V}_{0,nq}$ from q from 0 to $N = \lfloor 2p^{0.7} \rfloor = 434$ in the right panel.

Comparing with Figure 6, it can be seen that the $\hat{\mathrm{S}}_{nq}/\hat{V}_{0,nq}$ effectively reflects the amount of dissimilarity between Σ_1 and Σ_2. For $q = 1$ to 300, where obvious difference in the signals between the healthy and the cancer samples can be observed in Figure 6, the tests for $H_{0,q} : S_q = 0$ were all rejected. For $q \geq 300$, although the $\hat{\mathrm{S}}_{nq}$ were quite small, since the variances $\hat{V}_{np}^2$ were also small the signal to noise ratios for single super-diagonals turned out to be quite large leading to rejection of $\mathrm{H}_{0,q}$, as well.

Using the Storey, Taylor and Siegmund (2004) procedure with the tuning parameter $\lambda = 0.5$, the joint hypothesis $\mathrm{H}_{0,q} : \mathrm{S}_q = 0$ for $q = 0, \ldots, N$ was rejected for $N = \lfloor 2p^{0.7} \rfloor = 434$. The Benjamini and Hochberg (1995) procedure gave the same result. This suggests that the dependence structure among the healthy and the cancer groups is significantly different. We also used the SY, LC and CLX tests for the null hypothesis $\mathrm{H}_0 : \Sigma_1 = \Sigma_2$. The LC test rejects H_0

at the smallest p-value 0.0000, while The CLX test rejects H_0 at p-value 0.0006, and the SY test also rejects H_0 at p-value 0.015. The conclusion of the proposed method is in accordance with that of these tests.

Supplementary Materials

Proofs of the main results are in the supplementary materials. We also provide more simulation results of the proposed test method in conjunction with the Benjamini and Hochberg (1995) procedure, and for Gamma-distributed data.

Acknowledgment

The authors thank the Editor and two reviewers for constructive comments which led to improvement in the presentation of the paper. The first author was supported by China's National Natural Science Foundation grants 11701466 and the Fundamental Research Funds for the Central Universities JBK1801066. The second author acknowledge support from China's National Key Research and Development Grants SQ2016YFC207701 and 2015CB856000, and China's National Natural Science Foundation grants 11131002, 71371016 and 71532001.

References

Adam, B. L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z. and Wright, G. L. Jr. (2002). Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research* **62**, 3609–3614.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd Edition. Wiley, New York.

Bai, Z., Jiang, D., Yao, J. F. and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **37**, 3822–3840.

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

Bai, Z. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.* **21**, 1275–1294.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B Stat. Methodol* **57**, 289–300.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

Cai, T. T., Liu, W. and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108**, 265–277.

Cai, T. T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under

dependence. *J. Roy. Statist. Soc. Ser. B Stat. Methodol* **76**, 349–372.

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

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

Chen, S. X., Zhang, L. X. and Zhong, P. S. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105**, 810–819.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

Doukhan, P. (1994). *Mixing: Properties and Examples*, Lecture Notes in Statistics. Springer-Verlag, New York.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *Amer. Statistician* **56**, 316–324.

Gupta, A. and Tang, J. (1984). Distribution of likelihood ratio statistic for testing equality of covariance matrices of multivariate Gaussian models. *Biometrika* **71**, 555–559.

Hall, P. and Jin, J. (2009). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38**, 1686–1732.

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.* **2**, 360–378.

John, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika* **59**, 169–173.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**, 295–327.

Lee, L. F. and Yu, J. (2010). Estimation of spatial autoregressive panel data models with fixed effects. *J. Econometrics* **154**, 165–185.

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40**, 908–940.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Nagao, H. (1973). On some test criteria for covariance matrix. *Ann. Statist.* **1**, 700–709.

Qiu, Y. and Chen, S. X. (2012). Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *Ann. Statist.* **40**, 1285–1314.

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

Rodríguez, J. and Bárdossy, A. (2014). Multivariate interactions modeling through their manifestations: low dimensional model building via the Cumulant Generating Function. arXiv preprint arXiv:1406.2815.

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. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99**, 386–402.

Srivastava, M. S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *J. Multivariate Anal.* **101**, 1319–1329.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. Ser. B Stat. Methodol* **66**, 187–205.

School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China.

E-mail: he_jing@swufe.edu.cn

Department of Business Statistics and Econometrics, Guanghua School of Management and The Center for Statistical Science, Peking University, Beijing, China.

E-mail: csx@gsm.pku.edu.cn

(Received May 2017; accepted January 2018)