2013

Two-Sample Covariance Matrix Testing and Support Recovery

Tony Cai
University of Pennsylvania

Weidong Liu

Yin Xia
University of Pennsylvania

Recommended Citation

Cai, T., Liu, W., & Xia, Y. (2013). Two-Sample Covariance Matrix Testing and Support Recovery. *Journal of the American Statistical Association, 108* (501), 265-277. http://dx.doi.org/10.1080/01621459.2012.758041

Two-Sample Covariance Matrix Testing and Support Recovery

Abstract

This paper proposes a new test for testing the equality of two covariance matrices Σ_1 and Σ_2 in the high-dimensional setting and investigates its theoretical and numerical properties. The limiting null distribution of the test statistic is derived. The test is shown to enjoy certain optimality and to be especially powerful against sparse alternatives. The simulation results show that the test significantly outperforms the existing methods both in terms of size and power. Analysis of prostate cancer datasets is carried out to demonstrate the application of the testing procedures. When the null hypothesis of equal covariance matrices is rejected, it is often of significant interest to further investigate in which way they differ. Motivated by applications in genomics, we also consider two related problems, recovering the support of $\Sigma_1 - \Sigma_2$ and testing the equality of the two covariance matrices row by row. New testing procedures are introduced and their properties are studied. Applications to gene selection is also discussed.

Keywords

extreme value Type 1 distribution, gene selection, hypothesis testing, sparsity

Disciplines

Biostatistics | Genetics and Genomics

Two-Sample Covariance Matrix Testing And Support Recovery [*]

Tony Cai, Weidong Liu and Yin Xia

Abstract

This paper proposes a new test for testing the equality of two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in the high-dimensional setting and investigates its theoretical and numerical properties. The limiting null distribution of the test statistic is derived. The test is shown to enjoy certain optimality and to be especially powerful against sparse alternatives. The simulation results show that the test significantly outperforms the existing methods both in terms of size and power. Analysis of prostate cancer datasets is carried out to demonstrate the application of the testing procedures.

When the null hypothesis of equal covariance matrices is rejected, it is often of significant interest to further investigate in which way they differ. Motivated by applications in genomics, we also consider two related problems, recovering the support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ and testing the equality of the two

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

covariance matrices row by row. New testing procedures are introduced and their properties are studied. Applications to gene selection is also discussed.

Keywords: Covariance matrix, extreme value type I distribution, gene selection, hypothesis testing, sparsity, support recovery.

1 Introduction

Testing the equality of two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ is an important problem in multivariate analysis. Many statistical procedures including the classical Fisher's linear discriminant analysis rely on the fundamental assumption of equal covariance matrices. This testing problem has been well studied in the conventional low-dimensional setting. See, for example, Sugiura and Nagao (1968), Gupta and Giri (1973), Perlman (1980), Gupta and Tang (1984), O'Brien (1992), and Anderson (2003). In particular, the likelihood ratio test (LRT) is commonly used and enjoys certain optimality under regularity conditions.

Driven by a wide range of contemporary scientific applications, analysis of high dimensional data is of significant current interest. In the high dimensional setting, where the dimension can be much larger than the sample size, the conventional testing procedures such as the LRT perform poorly or are not even well defined. Several tests for the equality of two large covariance matrices have been proposed. For example, Schott (2007) introduced a test based on the Frobenius norm of the difference of the two covariance matrices. Srivastava and Yanagihara (2010) constructed a test that relied on a measure of distance by $tr(\mathbf{\Sigma}_1^2)/(tr(\mathbf{\Sigma}_1))^2 - tr(\mathbf{\Sigma}_2^2)/(tr(\mathbf{\Sigma}_2))^2$. Both of these two tests are designed for the multivariate normal populations. Chen and Li (2011) proposed a test using a linear combination of three one-sample U-statistics which was also motivated by an unbiased estimator of the Frobenius norm of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$.

In many applications such as gene selection in genomics, the covariance matrices of the two populations can be either equal or quite similar in the sense that they only possibly differ in a small number of entries. In such a setting, under the alternative the difference of the two covariance matrices $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ is sparse. The above mentioned tests, which are all based on the Frobenius norm, are not powerful against such sparse alternatives.

The main goal of this paper is to develop a test that is powerful against sparse alternatives and robust with respect to the population distributions. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two p variate random vectors with covariance matrices $\mathbf{\Sigma}_1 = (\sigma_{ij1})_{p\times p}$ and $\mathbf{\Sigma}_2 = (\sigma_{ij2})_{p\times p}$ respectively. Let $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}\}$ be i.i.d. random samples from $\boldsymbol{X}$ and let $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}\}$ be i.i.d. random samples from $\boldsymbol{Y}$ that are independent of $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}\}$. We wish to test the hypotheses

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

based on the two samples. We are particularly interested in the high dimensional setting where p can be much larger than $n = \max(n_1, n_2)$. Define the sample covariance matrices

$$(\hat{\sigma}_{ij1})_{p\times p} := \hat{\mathbf{\Sigma}}_1 = \frac{1}{n_1}\sum_{k=1}^{n_1}(\boldsymbol{X}_k - \bar{\boldsymbol{X}})(\boldsymbol{X}_k - \bar{\boldsymbol{X}})',$$
$$(\hat{\sigma}_{ij2})_{p\times p} := \hat{\mathbf{\Sigma}}_2 = \frac{1}{n_2}\sum_{k=1}^{n_2}(\boldsymbol{Y}_k - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_k - \bar{\boldsymbol{Y}})',$$

where $\bar{\boldsymbol{X}} = \frac{1}{n_1}\sum_{k=1}^{n_1}\boldsymbol{X}_k$ and $\bar{\boldsymbol{Y}} = \frac{1}{n_2}\sum_{k=1}^{n_2}\boldsymbol{Y}_k$.

In this paper we propose a test based on the maximum of the standardized differences between the entries of the two sample covariance matrices and investigate its theoretical and numerical properties. The limiting null distribution of the test statistic is derived. It is shown that the distribution of the test statistic converges to a type I extreme value distribution under the null H_0. This fact implies that the proposed test has the pre-specified significance level asymptotically. The theoretical analysis shows that the test enjoys certain optimality against a class of sparse alternatives in terms of the power. In addition to the theoretical properties, we also investigate the numerical performance of the proposed testing procedure using both simulated and real datasets. The numerical results show that the new test significantly outperforms the existing methods both in terms of size and power. We also use prostate cancer datasets to demonstrate the application of our testing procedures for gene selection.

In many applications, if the null hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is rejected, it is often of significant interest to further investigate in which way the two covariance matrices differ from each other. Motivated by applications to gene selection, in this paper we also consider two related problems, recovering the support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ and testing the equality of the two covariance matrices row by row. Support recovery can also be viewed as simultaneous testing of the equality of individual entries between the two covariance matrices. We introduce a procedure for support recovery based on the thresholding of the standardized differences between the entries of the two covariance matrices. It is shown that under certain conditions, the procedure recovers the true support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ exactly with probability tending to 1. The procedure is also shown to be minimax rate optimal.

The problem of testing the equality of two covariance matrices row by row is motivated by applications in genomics. A commonly used approach in microarray analysis is to select "interesting" genes by applying multiple testing procedures on the two-sample t-statistics. This approach has been successful in finding genes with significant changes in the mean expression levels between diseased and non-diseased populations. It has been noted recently that these mean-based methods lack the ability to discover genes that change their relationships with other genes and new methods that are based on the change in the gene's dependence structure are thus needed to identify these genes. See, for example, Ho, et al. (2008), Hu, et al. (2009) and Hu, et al. (2010). In this paper, we propose a procedure which simultaneously tests the p null hypotheses that the corresponding rows of the two covariance matrices are equal to each other. Asymptotic null distribution of the test statistics is derived and properties of the test is studied. It is shown that the procedure controls the family-wise error rate at a pre-specified level. Applications to gene selection is also considered.

The rest of this paper is organized as follows. Section 2 introduces the procedure for testing the equality of two covariance matrices. Theoretical properties of the test are investigated in Section 3. After Section 3.1 in which basic definitions and assumptions are given, Section 3.2 develops the asymptotic null distribution of the test statistic and presents the optimality results in terms of the power of the test against sparse alternatives. Section 4 considers support recovery of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ and testing the two covariance matrices row by row. Applications to gene selection is also discussed. Section 5 investigates the numerical performance of the proposed tests by simulations and by applications to the analysis of prostate cancer datasets. Section 6 discusses our results and other related work. The proofs of the main results are given in Section 7.

2 The testing procedure

The null hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is equivalent to $H_0 : \max_{1\leq i\leq j\leq p} |\sigma_{ij1} - \sigma_{ij2}| = 0$. A natural approach to testing this hypothesis is to compare the sample covariances $\hat{\sigma}_{ij1}$ and $\hat{\sigma}_{ij2}$ and to base the test on the maximum differences. It is important to note that the sample covariances $\hat{\sigma}_{ij1}$'s and $\hat{\sigma}_{ij2}$'s are in general heteroscedastic and can possibly have a wide range of variability. It is thus necessary to first standardize $\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2}$ before making a comparison among different entries.

To be more specific, define the variances $\theta_{ij1} = \mathsf{Var}((X_i - \mu_{1i})(X_j - \mu_{1j}))$ and $\theta_{ij2} = \mathsf{Var}((Y_i - \mu_{2i})(Y_j - \mu_{2j}))$. Given the two samples $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}\}$ and $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}\}$, θ_{ij1} and θ_{ij2} can be respectively estimated by

$$\hat{\theta}_{ij1} = \frac{1}{n_1} \sum_{k=1}^{n_1} \Big[(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j) - \hat{\sigma}_{ij1} \Big]^2, \quad \bar{X}_i = \frac{1}{n_1} \sum_{k=1}^{n_1} X_{ki},$$

and

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

Such an estimator of the variance have been used in Cai and Liu (2011) in the context of adaptive estimation of a sparse covariance matrix. Given $\hat{\theta}_{ij1}$ and $\hat{\theta}_{ij2}$, the variance of $\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2}$ can then be estimated by $\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2$.

Define the standardized statistics

$$W_{ij} =: \frac{\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2}}{\sqrt{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2}} \quad \text{and} \quad M_{ij} =: W_{ij}^2, \quad , \quad 1 \le i \le j \le p. \tag{2}$$

We consider the following test statistic for testing the hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$,

$$M_n =: \max_{1 \le i \le j \le p} M_{ij} = \max_{1 \le i \le j \le p} \frac{(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2})^2}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2}. \tag{3}$$

The asymptotic behavior of the test statistic M_n will be studied in detail in Section 3. Intuitively, W_{ij} are approximately standard normal variables under the null H_0 and the M_{ij} are only "weakly dependent" under suitable conditions. The test statistic M_n is the maximum of p^2 such variables and so the value of M_n is close to $2 \log p^2$ under H_0, based on the extreme values of normal random variables. More precisely, we shall show in Section 3 that, under certain regularity conditions, $M_n - 4 \log p + \log \log p$ converges to a type I extreme value distribution under the null hypothesis H_0. Based on this result, for a given significance level $0 < \alpha < 1$, we define the test Φ_α by

$$\Phi_\alpha = I(M_n \ge q_\alpha + 4 \log p - \log \log p) \tag{4}$$

where q_α is the $1-\alpha$ quantile of the type I extreme value distribution with the cumulative distribution function $\exp(-\frac{1}{\sqrt{8\pi}}\exp(-\frac{x}{2}))$, i.e.,

$$q_\alpha = -\log(8\pi) - 2\log\log(1-\alpha)^{-1}. \tag{5}$$

The hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is rejected whenever $\Phi_\alpha = 1$.

The test Φ_α is particularly well suited for testing against sparse alternatives. It is consistent if one of the entries of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ has a magnitude no less than $C\sqrt{\log p/n}$ for some constant $C > 0$ and no other special structure of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ is required. It will be shown that the test Φ_α is an asymptotically α-level test and enjoys certain optimality against sparse alternatives.

In addition, the standardized statistics M_{ij} are useful for identifying the support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$. That is, they can be used to estimate the positions at which the two covariance matrices differ from each other. This is of particular interest in many applications including gene selection. We shall show in Section 4.1 that under certain conditions, the support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ can be correctly recovered by thresholding the M_{ij}.

3 Theoretical analysis of size and power

We now turn to an analysis of the properties of the test Φ_α including the asymptotic size and power. The asymptotic size of the test is obtained by deriving the limiting distribution of the test statistic M_n under the null. We are particularly interested in the power of the test Φ_α under the sparse alternatives where $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ are sparse in the sense that $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ only contains a small number of nonzero entries.

3.1 Definitions and assumptions

Before we present the main results, we need to introduce some definitions and technical assumptions. Throughout the paper, denote $|\mathbf{a}|_2 = \sqrt{\sum_{j=1}^{p} a_j^2}$ for the usual Euclidean norm of a vector $\mathbf{a} = (a_1, \ldots, a_p)^T \in I\!R^p$. For a matrix $\boldsymbol{A} = (a_{ij}) \in I\!R^{p\times q}$, define the spectral norm $\|\boldsymbol{A}\|_2 = \sup_{|\mathbf{x}|_2 \le 1} |\boldsymbol{A}\mathbf{x}|_2$ and the Frobenius norm $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \le C|b_n|$ holds for all sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there exist constants $C > c > 0$ such that $c|b_n| \le |a_n| \le C|b_n|$ for all sufficiently large n.

The two sample sizes are assumed to be comparable, i.e., $n_1 \asymp n_2$. Let $n = \max(n_1, n_2)$ and let $\boldsymbol{R}_1 =: (\rho_{ij1})$ and $\boldsymbol{R}_2 =: (\rho_{ij2})$ be the correlation matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. For $0 < r < 1$, define the set

$$\boldsymbol{\Lambda}(r) = \{1 \le i \le p : |\rho_{ij1}| > r \text{ or } |\rho_{ij2}| > r \text{ for some } j \ne i\}.$$

So $\boldsymbol{\Lambda}(r)$ is the set of indices i such that either the ith variable of $\boldsymbol{X}$ is highly correlated ($> r$) with some other variable of $\boldsymbol{X}$ or the ith variable of $\boldsymbol{Y}$ is highly correlated ($> r$) with some other variable of $\boldsymbol{Y}$.

To obtain the asymptotic distribution of M_n, we assume that the eigenvalues of the correlation matrices are bounded from above, which is a commonly used assumption, and the set $\boldsymbol{\Lambda}(r)$ is not too large for some $r < 1$.

(C1). Suppose that $\lambda_{\max}(\boldsymbol{R}_1) \le C_0$ and $\lambda_{\max}(\boldsymbol{R}_2) \le C_0$ for some $C_0 > 0$. Moreover, there exist some constant $r < 1$ and a sequence of numbers $\Lambda_{p,r}$ such that $\mathrm{Card}(\boldsymbol{\Lambda}(r)) \le \Lambda_{p,r} = o(p)$.

It is easy to see that the condition $\mathrm{Card}(\boldsymbol{\Lambda}(r)) = o(p)$ for some $r < 1$ is trivially satisfied if all the correlations are bounded away from ±1, i.e.,

$$\max_{1\le i<j\le p} |\rho_{ij1}| \le r < 1 \text{ and } \max_{1\le i<j\le p} |\rho_{ij2}| \le r < 1. \tag{6}$$

We do not require the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be Gaussian. Instead we shall impose moment conditions.

(C2). *Sub-Gaussian type tails:* Suppose that $\log p = o(n^{1/5})$. There exist some constants $\eta > 0$ and $K > 0$ such that

$$\mathsf{E}\exp(\eta(X_i - \mu_{i1})^2/\sigma_{ii1}) \le K, \quad \mathsf{E}\exp(\eta(Y_i - \mu_{i2})^2/\sigma_{ii2}) \le K \quad \text{for all } i.$$

Furthermore, we assume that for some constants $\tau_1 > 0$ and $\tau_2 > 0$,

$$\min_{1\le i\le j\le p} \frac{\theta_{ij1}}{\sigma_{ii1}\sigma_{jj1}} \ge \tau_1 \text{ and } \min_{1\le i\le j\le p} \frac{\theta_{ij2}}{\sigma_{ii2}\sigma_{jj2}} \ge \tau_2. \tag{7}$$

(C2*). *Polynomial-type tails:* Suppose that for some $\gamma_0, c_1 > 0$, $p \le c_1 n^{\gamma_0}$, and for some $\epsilon > 0$

$$\mathsf{E}|(X_i - \mu_{i1})/\sigma_{ii1}^{1/2}|^{4\gamma_0+4+\epsilon} \le K, \quad \mathsf{E}|(Y_i - \mu_{i2})/\sigma_{ii2}^{1/2}|^{4\gamma_0+4+\epsilon} \le K \quad \text{for all } i.$$

Furthermore, we assume (7) holds.

In addition to the moment conditions, as in Bai and Saranadasa (1996), we assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be written as the transforms of white noise.

(C3). For any four distinct indices $S = (i_1, i_2, i_3, i_4)$, the vector $\boldsymbol{X}_S = (X_{i_1}, X_{i_2}, X_{i_3}, X_{i_4})^T$ can be written as

$$\boldsymbol{X}_S = \Gamma_S \boldsymbol{Z}_S + \boldsymbol{\mu}_S, \tag{8}$$

where $\boldsymbol{\mu}_S = (\mu_{i_1}, \ldots, \mu_{i_4})^T$, Γ_S is a lower triangular matrix satisfying $\Gamma_S\Gamma_S^T = \mathsf{Cov}(\boldsymbol{X}_S)$ by Cholesky decomposition, and $\boldsymbol{Z}_S = (Z_{i_1}, \ldots, Z_{i_4})^T$ satisfies that $\mathsf{E}\boldsymbol{Z}_S = 0$, $\mathsf{Cov}(\boldsymbol{Z}_S) = I_{4\times 4}$ and for nonnegative integers $\alpha_1, \ldots, \alpha_4$,

$$\mathsf{E}\prod_{j=1}^{4} Z_{i_j}^{\alpha_j} = \prod_{j=1}^{4} \mathsf{E}Z_{i_j}^{\alpha_j} \tag{9}$$

whenever $\sum_{k=1}^4 \alpha_k \le 4$. The same conditions (8) and (9) hold for $\boldsymbol{Y}$.

Remark 2. Conditions (C2) and (C2*) are moment conditions on $\boldsymbol{X}$ and $\boldsymbol{Y}$. They are much weaker than the Gaussian assumption required in the existing literature such as Schott (2007) and Srivastava and Yanagihara (2010). Condition (7) is satisfied with $\tau_1 = \tau_2 = 1$ if $\boldsymbol{X} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{Y} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. For non-Gaussian distributions, if (C3) and (6) hold, then (7) holds with $\tau_1 = \tau_2 = 1 - r^2$. (C3) is a technical condition for the asymptotic distribution of M_n. It holds if $\boldsymbol{X} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{Y} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.

3.2 Limiting null distribution and optimality

We are now ready to introduce the asymptotic null distribution of M_n. The following theorem shows that $M_n - 4\log p + \log\log p$ converges weakly under H_0 to an extreme value distribution of type I with distribution function $F(t) = \exp(-\frac{1}{\sqrt{8\pi}}e^{-t/2})$, $t \in I\!R$.

Theorem 1 *Suppose that (C1), (C2) (or (C2*)) and (C3) hold. Then under H_0, as $n, p \to \infty$, we have for any $t \in I\!R$,*

$$\mathsf{P}\Big(M_n - 4\log p + \log\log p \le t\Big) \to \exp\Big(-\frac{1}{\sqrt{8\pi}}\exp\Big(-\frac{t}{2}\Big)\Big). \tag{10}$$

Furthermore, the convergence in (10) is uniformly for all $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfying (C1), (C2) (or (C2*)) and (C3) and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

The limiting null distribution given in (10) shows that the test Φ_α defined in (4) is an asymptotically level α test. Numerical results show that for moderately large n and p the distribution of M_n is already well approximated by its asymptotic distribution and consequently the actual size of the test is close to the nominal level.

The limiting behavior of M_n is similar to that of the largest off-diagonal entry L_n of the sample correlation matrix in Jiang (2004), wherein the paper derived the asymptotic distribution of L_n under the assumption that the components $X_1, \ldots, X_p$ are independent. Some further extensions and improvements can be found in Zhou

(2007), Liu, Lin and Shao (2008) and Cai and Jiang (2011). A key assumption in these papers is the independence between the components. The techniques in their proofs can not be used to obtain (10), since dependent components are allowed in Theorem 1. The proof of (10) requires quite different techniques.

We now turn to an analysis of the power of the test Φ_α given in (4). We shall define the following class of matrices:

$$\mathcal{U}(c) = \Big\{(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) : \max_{1\le i\le j\le p} \frac{|\sigma_{ij1} - \sigma_{ij2}|}{\sqrt{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}} \ge c\sqrt{\log p}\Big\}. \tag{11}$$

The next theorem shows that the null parameter set in which $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is asymptotically distinguishable from $\mathcal{U}(4)$ by the test Φ_α. That is, H_0 is rejected by Φ_α with overwhelming probability if $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathcal{U}(4)$.

Theorem 2 *Suppose that (C2) or (C2*) holds. Then we have*

$$\inf_{(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in\mathcal{U}(4)} \mathsf{P}\Big(\Phi_\alpha = 1\Big) \to 1, \tag{12}$$

as $n, p \to \infty$.

It can be seen from Theorem 2 that it only requires one of the entries of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ having a magnitude no less than $C\sqrt{\log p/n}$ in order for the test Φ_α to correctly reject H_0. This lower bound is rate-optimal. Let $\mathcal{T}_\alpha$ be the set of α-level tests, i.e., $\mathsf{P}(T_\alpha = 1) \le \alpha$ under H_0 for any $T_\alpha \in \mathcal{T}_\alpha$.

Theorem 3 *Suppose that* $\log p = o(n)$, $\boldsymbol{X} \sim N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ *and* $\boldsymbol{Y} \sim N(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$. *Let* $\alpha, \beta > 0$ *and* $\alpha + \beta < 1$. *There exists some constant* $c_0 > 0$ *such that for all large* n *and* p,

$$\inf_{(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in\mathcal{U}(c_0)} \sup_{T_\alpha\in\mathcal{T}_\alpha} \mathsf{P}\Big(T_\alpha = 1\Big) \le 1 - \beta. \tag{13}$$

The difference between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ is measured by $\max_{1\le i\le j\le p}|\sigma_{ij1}-\sigma_{ij2}|$ in Theorem 3. Another measurement is based on the Frobenius norm $\|\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\|_F$. Suppose that $\mathbf{\Sigma}_1-\mathbf{\Sigma}_2$ is a sparse matrix with $c_0(p)$ nonzero entries. That is,

$$c_0(p)=\sum_{i=1}^{p}\sum_{j=1}^{p} I\{\sigma_{ij1}-\sigma_{ij2}\neq 0\}.$$

We introduce the following class of matrices for $\mathbf{\Sigma}_1-\mathbf{\Sigma}_2$:

$$\mathcal{V}(c)=\Big\{(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2):\max_i|\sigma_{ii1}|\le K,\max_i|\sigma_{ii2}|\le K,\|\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\|_F^2\ge cc_0(p)\frac{\log p}{n}\Big\}.$$

Note that on $\mathcal{V}(c)$, we have $\max_{1\le i\le j\le p}|\sigma_{ij1}-\sigma_{ij2}|\ge\sqrt{c\log p/n}$. Thus, for sufficiently large c,

$$\inf_{(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in\mathcal{V}(c)}\mathsf{P}\Big(\Phi_\alpha=1\Big)\to 1$$

as $n,p\to\infty$. The following theorem shows that the lower bound for $\|\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\|_F^2$ in $\mathcal{V}(c)$ is rate-optimal. That is, no α-level test can reject H_0 with overwhelming probability uniformly over $\mathcal{V}(c_0)$ for some $c_0>0$.

Theorem 4 *Suppose that* $\log p=o(n)$, $\boldsymbol{X}\sim N(\boldsymbol{\mu}_1,\mathbf{\Sigma}_1)$ *and* $\boldsymbol{Y}\sim N(\boldsymbol{\mu}_2,\mathbf{\Sigma}_2)$. *Assume that* $c_0(p)\le p^r$ *for some* $0<r<1/2$. *Let* $\alpha,\beta>0$ *and* $\alpha+\beta<1$. *There exists a* $c_0>0$ *such that for all large* n *and* p,

$$\inf_{(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in\mathcal{V}(c_0)}\sup_{T_\alpha\in\mathcal{T}_\alpha}P\Big(T_\alpha=1\Big)\le 1-\beta.$$

Note that under the multivariate normality assumption, for every $c>0$, there exists some constant $K(c)>0$ such that for any $0<c_0<K(c)$, $\mathcal{V}(c)\subset\mathcal{U}(c_0)$. Thus, Theorem 3 follows from Theorem 4 directly.

4 Support recovery of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ and application to gene selection

We have focused on testing the equality of two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in Sections 2 and 3. As mentioned in the introduction, if the null hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is rejected, further exploring in which ways they differ is also of significant interest in practice. Motivated by applications in gene selection, we consider in this section two related problems, one is recovering the support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ and another is identifying the rows on which the two covariance matrices differ from each other.

4.1 Support recovery of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$

The goal of support recovery is to find the positions at which the two covariance matrices differ from each other. The problem can also be viewed as simultaneous testing of equality of individual entries between the two covariance matrices. Denote the support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ by

$$\Psi = \Psi(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = \{(i,j) : \sigma_{ij1} \neq \sigma_{ij2}\}. \tag{14}$$

In certain applications, the variances along the diagonal play a more important role than the covariances. For example, in a differential variability analysis of gene expression Ho, et al (2008) proposed to select genes based on the differences in the variances. In this section we shall treat the variances along the diagonal differently from the off-diagonal covariances for support recovery. Since there are p diagonal elements, M_{ii} along the diagonal are thresholded at $2\log p$ based on the extreme values of normal variables. The off-diagonal entries M_{ij} with $i \neq j$ are thresholded at a different level. More specifically, set

$$\hat{\Psi}(\tau) = \{(i,i) : M_{ii} \geq 2\log p\} \cup \{(i,j) : M_{ij} \geq \tau \log p, \quad i \neq j\}, \tag{15}$$

where M_{ij} are defined in (2) and τ is the threshold constant for the off-diagonal entries.

The following theorem shows that with $\tau = 4$ the estimator $\hat{\Psi}(4)$ recovers the support Ψ exactly with probability tending to 1 when the magnitudes of nonzero entries are above certain thresholds. Define

$$\mathcal{W}_0(c) = \Big\{(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) : \min_{(i,j)\in\Psi} \frac{|\sigma_{ij1} - \sigma_{ij2}|}{\sqrt{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}} \geq c\sqrt{\log p}\Big\}.$$

We have the following result.

Theorem 5 *Suppose that (C2) or (C2*) holds. Then*

$$\inf_{(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\in\mathcal{W}_0(4)} P\Big(\hat{\Psi}(4) = \Psi\Big) \to 1$$

as $n, p \to \infty$.

The choice of the threshold constant $\tau = 4$ is optimal for the exact recovery of the support Ψ. Consider the class of $s_0(p)$-sparse matrices,

$$\mathcal{S}_0 = \Big\{A = (a_{ij})_{p\times p} : \max_i \sum_{j=1}^{p} I\{a_{ij} \neq 0\} \leq s_0(p)\Big\}.$$

Theorem 6 *Suppose that (C1), (C2) (or (C2*)) and (C3) hold. Let $0 < \tau < 4$. If $s_0(p) = O(p^{1-\tau_1})$ with some $\tau/4 < \tau_1 < 1$, then as $n, p \to \infty$,*

$$\sup_{\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\in\mathcal{S}_0} P\Big(\hat{\Psi}(\tau) = \Psi\Big) \to 0.$$

In other words, for any threshold constant $\tau < 4$, the probability of recovering the support exactly tends to 0 for all $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \in \mathcal{S}_0$. This is mainly due to the fact that the threshold level $\tau \log p$ is too small to ensure that the zero entries of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ are estimated by zero.

In addition, the minimum separation of the magnitudes of the nonzero entries must be at least of order $\sqrt{\log p/n}$ to enable the exact recovery of the support.

Consider, for example, $s_0(p) = 1$ in Theorem 4. Then Theorem 4 indicates that no test can distinguish H_0 and H_1 uniformly on the space $\mathcal{V}(c_0)$ with probability tending to one. It is easy to see that $\mathcal{V}(c_0) \subseteq \mathcal{W}_0(c_1)$ for some $c_1 > 0$. Hence, there is no estimate that can recover the support Ψ exactly uniformly on the space $\mathcal{W}_0(c_1)$ with probability tending to one.

We have so far focused on the exact recovery of the support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ with probability tending to 1. It is sometimes desirable to recover the support under a less stringent criterion such as the family-wise error rate (FWER), defined by FWER $= \mathsf{P}(V \geq 1)$, where V is the number of false discoveries. The goal of using the threshold level $4 \log p$ is to ensure FWER $\to 0$. To control the FWER at a pre-specified level α for some $0 < \alpha < 1$, a different threshold level is needed. For this purpose, we shall set the threshold level at $4 \log p - \log\log p + q_\alpha$ where q_α is given in (5). Define

$$\hat{\Psi}^{\star} = \{(i,i) : M_{ii} \geq 2\log p\} \cup \{(i,j) : M_{ij} \geq 4\log p - \log\log p + q_\alpha, \quad i \neq j\}.$$

We have the following result.

Proposition 1 *Suppose that (C1), (C2) (or (C2*)) and (C3) hold. Under* $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \in \mathcal{S}_0 \cap \mathcal{W}_0(4)$ *with* $s_0(p) = o(p)$*, we have as* $n, p \to \infty$*,*

$$\mathsf{P}\Big(\hat{\Psi}^{\star} \neq \Psi\Big) \to \alpha.$$

4.2 Testing rows of two covariance matrices

As discussed in the introduction, the standard methods for gene selection are based on the comparisons of the means and thus lack the ability to select genes that change their relationships with other genes. It is of significant practical interest to develop methods for gene selection which capture the changes in the gene's dependence structure.

Several methods have been proposed in the literature. It was noted in Ho, et al. (2008) that the changes of variances are biologically interesting and are associated with changes in coexpression patterns in different biological states. Ho, et al. (2008) proposed to test $H_{0i} : \sigma_{ii1} = \sigma_{ii2}$ and select the i-th gene if H_{0i} is rejected, and Hu, et al. (2009) and Hu, et al. (2010) introduced methods which are based on simultaneous testing of the equality of the joint distributions of each row between two sample correlation matrices/covariance distance matrices.

The covariance provides a natural measure of the association between two genes and it can also reflect the changes of variances. Motivated by these applications, in this section we consider testing the equality of two covariance matrices row by row. Let $\sigma_{i\cdot 1}$ and $\sigma_{i\cdot 2}$ be the i-th row of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ respectively. We consider testing simultaneously the hypotheses

$$H_{0i} : \sigma_{i\cdot 1} = \sigma_{i\cdot 2},\ 1 \leq i \leq p.$$

We shall use the family-wise error rate (FWER) to measure the type I errors for the p tests. The support estimate $\hat{\Psi}(4)$ defined in (15) can be used to test the hypotheses H_{0i} by rejecting H_{0i} if the i-th row of $\hat{\Psi}(4)$ is nonzero. Suppose that (C1) and (C2) (or (C2*)) hold. Then it can be shown easily that for this test, the FWER $\to 0$.

A different test is needed to simultaneously test the hypotheses H_{0i}, $1 \leq i \leq p$, at a pre-specified FWER level α for some $0 < \alpha < 1$. Define

$$M_{i\cdot} = \max_{1 \leq j \leq p, j \neq i} \frac{(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2})^2}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2}.$$

The null distribution of $M_{i\cdot}$ can be derived similarly under the same conditions as in Theorem 1.

Proposition 2 *Suppose the null hypothesis H_{0i} holds. Then under the conditions* *in Theorem 1,*

$$P\Big(M_{i\cdot} - 2\log p + \log\log p \le x\Big) \to \exp\Big(-\frac{1}{\sqrt{\pi}}\exp\Big(-\frac{x}{2}\Big)\Big). \tag{16}$$

Based on the limiting null distribution of $M_{i\cdot}$ given in (16), we introduce the following test for testing a single hypothesis H_{0i},

$$\Phi_{i,\alpha} = I(M_{i\cdot} \ge \alpha_p \text{ or } M_{ii} \ge 2\log p) \tag{17}$$

with $\alpha_p = 4\log p - \log\log p + q_\alpha$, where q_α is given in (5).

H_{0i} is rejected and the i-th gene is selected if $\Phi_{i,\alpha} = 1$. It can be shown that the tests $\Phi_{i,\alpha}$, $1 \le i \le p$ together controls the overall FWER at level α asymptotically.

Proposition 3 *Let $F = \{i : \sigma_{i\cdot 1} \ne \sigma_{i\cdot 2}, 1 \le i \le p\}$ and suppose Card$(F) = o(p)$. Then under the conditions in Theorem 1,*

$$\text{FWER} = P\Big(\max_{i\in F^c} \Phi_{i,\alpha} = 1\Big) \to \alpha.$$

5 Numerical results

In this section, we turn to the numerical performance of the proposed methods. The goal is to first investigate the numerical performance of the test Φ_α and support recovery through simulation studies and then illustrate our methods by applying them to the analysis of a prostate cancer dataset. The test Φ_α is compared with four other tests, the likelihood ratio test (LRT), the tests given in Schott (2007) and Chen and Li (2011) which are both based on an unbiased estimator of $\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_F^2$, the test proposed in Srivastava and Yanagihara (2009) which is based on a measure of distance by $tr(\mathbf{\Sigma}_1^2)/(tr(\mathbf{\Sigma}_1))^2 - tr(\mathbf{\Sigma}_2^2)/(tr(\mathbf{\Sigma}_2))^2$.

We first introduce the matrix models used in the simulations. Let $\boldsymbol{D} = (d_{ij})$ be a diagonal matrix with diagonal elements $d_{ii} = \text{Unif}(0.5, 2.5)$ for $i = 1, ..., p$. Denote by $\lambda_{\min}(A)$ the minimum eigenvalue of a symmetric matrix A. The following four models under the null, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}^{(i)}$, $i = 1, 2, 3$ and 4, are used to study the size of the tests.

- Model 1: $\mathbf{\Sigma}^{*(1)} = (\sigma_{ij}^{*(1)})$ where $\sigma_{ii}^{*(1)} = 1$, $\sigma_{ij}^{*(1)} = 0.5$ for $5(k-1)+1 \le i \ne j \le 5k$, where $k = 1, ..., [p/5]$ and $\sigma_{ij}^{*(1)} = 0$ otherwise. $\mathbf{\Sigma}^{(1)} = \boldsymbol{D}^{1/2}\mathbf{\Sigma}^{*(1)}\boldsymbol{D}^{1/2}$.
- Model 2: $\mathbf{\Sigma}^{*(2)} = (\sigma_{ij}^{*(2)})$ where $\omega_{ij}^{*(2)} = 0.5^{|i-j|}$ for $1 \le i, j \le p$. $\mathbf{\Sigma}^{(2)} = \boldsymbol{D}^{1/2}\mathbf{\Sigma}^{*(2)}\boldsymbol{D}^{1/2}$.
- Model 3: $\mathbf{\Sigma}^{*(3)} = (\sigma_{ij}^{(3)})$ where $\sigma_{ii}^{*(3)} = 1$, $\sigma_{ij}^{*(3)} = 0.5 * \text{Bernoulli}(1, 0.05)$ for $i < j$ and $\sigma_{ji}^{*(3)} = \sigma_{ij}^{*(3)}$. $\mathbf{\Sigma}^{(3)} = \boldsymbol{D}^{1/2}(\mathbf{\Sigma}^{*(3)} + \delta\boldsymbol{I})/(1+\delta)\boldsymbol{D}^{1/2}$ with $\delta = |\lambda_{\min}(\mathbf{\Sigma}^{*(3)})| + 0.05$.
- Model 4: $\mathbf{\Sigma}^{(4)} = \boldsymbol{O}\Delta\boldsymbol{O}$, where $\boldsymbol{O} = \text{diag}(\omega_1, ..., \omega_p)$ and $\omega_1, ..., \omega_p \overset{iid}{\sim} \text{Unif}(1, 5)$ and $\Delta = (a_{ij})$ and $a_{ij} = (-1)^{i+j}0.4^{|i-j|^{1/10}}$. This model was used in Srivastava and Yanagihara (2009).

To evaluate the power of the tests, let $\boldsymbol{U} = (u_{kl})$ be a matrix with 8 random nonzero entries. The locations of 4 nonzero entries are selected randomly from the upper triangle of $\boldsymbol{U}$, each with a magnitude generated from $\text{Unif}(0, 4) * \max_{1 \le j \le p} \sigma_{jj}$. The other 4 nonzero entries in the lower triangle are determined by symmetry. We use the following four pairs of covariance matrices $(\mathbf{\Sigma}_1^{(i)}, \mathbf{\Sigma}_2^{(i)})$, $i = 1, 2, 3$ and 4, to compare the power of the tests, where $\mathbf{\Sigma}_1^{(i)} = \mathbf{\Sigma}^{(i)} + \delta\boldsymbol{I}$ and $\mathbf{\Sigma}_2^{(i)} = \mathbf{\Sigma}^{(i)} + \boldsymbol{U} + \delta\boldsymbol{I}$, with $\delta = |\min\{\lambda_{\min}(\mathbf{\Sigma}^{(i)} + \boldsymbol{U}), \lambda_{\min}(\mathbf{\Sigma}^{(i)})\}| + 0.05$.

The sample sizes are taken to be $n_1 = n_2 = n$ with $n = 60$ and 100, while the dimension p varies over the values 50, 100, 200, 400 and 800. For each model, data are generated from multivariate normal distributions with mean zero and covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. The nominal significant level for all the tests is set at $\alpha = 0.05$.

The actual sizes and powers for the four models, reported in Table 1, are estimated from 5000 replications.

It can be seen from Table 1 that the estimated sizes of our test Φ_α are close to the nominal level 0.05 in all the cases. This reflects the fact that the null distribution of the test statistic M_n is well approximated by its asymptotic distribution. For Models 1-3, the estimated sizes of the tests in Schott (2007) and Chen and Li (2011) are also close to 0.05. But both tests suffer from the size distortion for Model 4, the actual sizes are around 0.10 for both tests. The likelihood ratio test has serious size distortion (all is equal to 1). Srivastava and Yanagihara (2010)'s test has actual sizes close to the nominal significance level only in the fourth model, with the actual sizes all close to 0 in the first three models.

The power results in Table 1 show that the proposed test has much higher power than the other tests in all settings. The number of nonzero off-diagonal entries of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ does not change when p grows, so the estimated powers of all tests tend to decrease when the dimension p increases. It can be seen in Table 1 that the powers of Schott (2007) and Chen and Li (2011)'s tests decrease extremely fast as p grows. Srivastava and Yanagihara (2010)'s test has trivial powers no matter how large p is. However, the power of the proposed test Φ_α remains high even when $p = 800$, especially in the case of $n = 100$. Overall, for the sparse models, the new test significantly outperforms all the other three tests.

We also carried out simulations for non-Gaussian distributions including Gamma distribution, t distribution and normal distribution contaminated with exponential distribution. Similar phenomena as those in the Gaussian case are observed. For reasons of space, these simulation results are given in the supplementary material, Cai, Liu and Xia (2011).

n	p	50	100	200	400	800	50	100	200	400	800	50	100	200	400	800	50	100	200	400	800
		Model 1					Model 2					Model 3					Model 4				
		Empirical size																			
60	Φ_α	0.052	0.049	0.051	0.056	0.061	0.055	0.054	0.050	0.053	0.060	0.055	0.053	0.056	0.059	0.059	0.045	0.044	0.043	0.047	0.055
	likelihood	1.000	1.000	NA	NA	NA	1.000	1.000	NA	NA	NA	1.000	1.000	NA	NA	NA	1.000	1.000	NA	NA	NA
	Schott	0.071	0.064	0.054	0.050	0.053	0.066	0.061	0.056	0.051	0.046	0.061	0.055	0.050	0.054	0.051	0.089	0.100	0.096	0.099	0.103
	Chen	0.071	0.058	0.053	0.055	0.056	0.066	0.060	0.052	0.048	0.049	0.063	0.052	0.053	0.054	0.051	0.100	0.109	0.102	0.091	0.104
	Srivastava	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.026	0.029	0.023	0.026	0.029
100	Φ_α	0.048	0.041	0.044	0.049	0.048	0.045	0.042	0.043	0.050	0.049	0.046	0.051	0.045	0.045	0.049	0.042	0.039	0.037	0.040	0.042
	likelihood	1.000	1.000	1.000	NA	NA	1.000	1.000	1.000	NA	NA	1.000	1.000	1.000	NA	NA	1.000	1.000	1.000	NA	NA
	Schott	0.065	0.052	0.049	0.052	0.049	0.056	0.050	0.049	0.050	0.052	0.050	0.051	0.050	0.049	0.046	0.092	0.092	0.092	0.099	0.097
	Chen	0.065	0.053	0.051	0.051	0.047	0.057	0.050	0.051	0.052	0.052	0.049	0.051	0.050	0.051	0.049	0.091	0.094	0.092	0.099	0.096
	Srivastava	0.002	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.032	0.036	0.035	0.035	0.032
		Empirical power																			
60	Φ_α	0.879	0.741	0.476	0.404	0.388	0.867	0.562	0.609	0.463	0.258	0.903	0.776	0.497	0.425	0.401	0.818	0.721	0.482	0.405	0.327
	Schott	0.323	0.148	0.086	0.066	0.064	0.347	0.094	0.090	0.067	0.057	0.354	0.157	0.087	0.069	0.069	0.284	0.135	0.082	0.071	0.072
	Chen	0.312	0.139	0.081	0.064	0.059	0.337	0.089	0.087	0.067	0.054	0.340	0.154	0.083	0.067	0.064	0.278	0.131	0.079	0.069	0.070
	Srivastava	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000
100	Φ_α	0.999	0.994	0.941	0.916	0.930	0.999	0.946	0.981	0.959	0.793	1.000	0.997	0.950	0.929	0.937	0.996	0.992	0.939	0.925	0.894
	Schott	0.576	0.249	0.100	0.078	0.070	0.638	0.141	0.132	0.090	0.058	0.643	0.280	0.109	0.077	0.067	0.513	0.219	0.094	0.079	0.065
	Chen	0.571	0.243	0.098	0.076	0.071	0.625	0.141	0.128	0.089	0.054	0.632	0.271	0.102	0.080	0.069	0.504	0.214	0.094	0.079	0.062
	Srivastava	0.021	0.000	0.000	0.000	0.000	0.028	0.000	0.000	0.000	0.000	0.038	0.003	0.000	0.000	0.000	0.038	0.007	0.000	0.000	0.000

Table 1: $N(0,1)$ random variables. Model 1-4 Empirical sizes and powers. $\alpha = 0.05$. $n = 60$ and 100. 5000 replications.

5.1 Support recovery

We now consider the simulation results on recovering the support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ in the first three models with $\boldsymbol{D} = \boldsymbol{I}$ and the fourth model with $\boldsymbol{O} = \boldsymbol{I}$ under the normal distribution. For $i = 1, 2, 3$ and 4, let $U^{(i)}$ be a matrix with 50 random nonzero entries, each with a magnitude of 2 and let $\mathbf{\Sigma}_1^{(i)} = (\mathbf{\Sigma}^{(i)} + \delta \boldsymbol{I})/(1+\delta)$ and $\mathbf{\Sigma}_2^{(i)} = (\mathbf{\Sigma}^{(i)} + \boldsymbol{U}^{(i)} + \delta \boldsymbol{I})/(1+\delta)$ with $\delta = |\min(\lambda_{\min}(\mathbf{\Sigma}^{(i)} + \boldsymbol{U}^{(i)}), \lambda_{\min}(\mathbf{\Sigma}^{(i)}))| + 0.05$. After normalization, the nonzero elements of $\mathbf{\Sigma}_2^{(i)} - \mathbf{\Sigma}_1^{(i)}$ have magnitude between 0.74 and 0.86 for $i = 1, 2, 3$ and 4 in our generated models.

The accuracy of the support recovery is evaluated by the similarity measure $s(\hat{\Psi}, \Psi)$ defined by

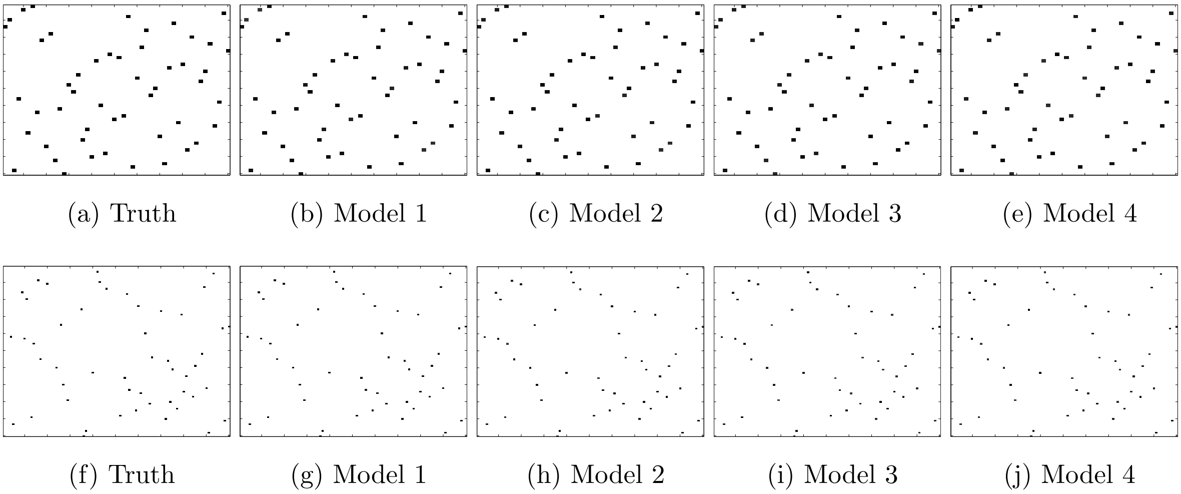
$$s(\hat{\Psi}, \Psi) = \frac{|\hat{\Psi} \cap \Psi|}{\sqrt{|\hat{\Psi}||\Psi|}},$$

where $\hat{\Psi} = \hat{\Psi}(4)$ and Ψ is the support of $\mathbf{\Sigma}_1^{(i)} - \mathbf{\Sigma}_2^{(i)}$ and $|\cdot|$ denotes the cardinality. Note that $0 \leq s(\hat{\Psi}, \Psi) \leq 1$ with $s(\hat{\Psi}, \Psi) = 0$ indicating disjointness and $s(\hat{\Psi}, \Psi) = 1$ indicating exact recovery. We summarize the average (standard deviation) values of $s(\hat{\Psi}, \Psi)$ for all models with $n = 100$ in Table 2 based on 100 replications. The values are close to one, and hence the supports are well recovered by our procedure.

	Model 1	Model 2	Model 3	Model 4
p=50	0.974(0.025)	0.898(0.043)	0.913(0.041)	0.930(0.038)
p=100	0.953(0.031)	0.868(0.051)	0.862(0.053)	0.897(0.044)
p=200	0.788(0.056)	0.801(0.059)	0.824(0.058)	0.840(0.064)
p=400	0.756(0.065)	0.732(0.077)	0.794(0.061)	0.778(0.070)
p=800	0.640(0.075)	0.618(0.088)	0.729(0.061)	0.681(0.076)

Table 2: Average (standard deviation) of $s(\hat{\Psi}, \Psi)$.

To better illustrate the elementwise recovery performance, heat maps of the nonzeros identified out of 100 replications for $p = 50$ and 100 are shown in Figures 1 and 2. These heat maps suggest that, in all models, the estimated support by $\hat{\Psi}(4)$ is close to the true support.


(a) Truth (b) Model 1 (c) Model 2 (d) Model 3 (e) Model 4

(f) Truth (g) Model 1 (h) Model 2 (i) Model 3 (j) Model 4

Figure 1: Heat maps of the frequency of the 0s identified for each entry of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ (n=100, p=50 for the top row and p=100 for the second row) out of 100 replications. White indicates 100 0s identified out of 100 runs; black, 0/100.

5.2 Real data analysis

We now illustrate our methods by applying them to the analysis of a prostate cancer dataset (Singh et al. (2002)) which is available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. The dataset consists of two classes of gene expression data that came from 52 prostate tumor patients and 50 prostate normal patients. This dataset has been analyzed in several papers on classification in which the two covariance matrices are assumed to be equal; see, for example, Xu, Brock and Parrish (2009). The equality of the two covariance matrices is an important assumption for the validity of these classification methods. It is thus interesting to test whether such an assumption is valid.

There are a total of 12600 genes. To control the computational costs, only the 5000 genes with the largest absolute values of the t-statistics are used. Let $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ be respectively the covariance matrices of these 5000 genes in tumor and normal samples. We apply the test Φ_α defined in (4) to test the hypotheses $H_0: \ \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ versus $H_1: \ \mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. Based on the asymptotic distribution of the test statistic M_n, the p-value is calculated to be 0.0058 and the null hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is thus rejected at commonly used significant levels such as $\alpha = 0.05$ or $\alpha = 0.01$. Based on this test result, it is therefore not reasonable to assume $\Sigma_1 = \Sigma_2$ in applying a classifier to this dataset.

We then apply three methods to select genes with changes in variances/covariances between the two classes. The first is the differential variability analysis (Ho, et al., 2008) which chooses the genes with different variances between two classes. In our procedure the i-th gene is selected if $M_{ii} \geq 2\log p$. As a result, 21 genes are selected. The second and third methods are based on the differential covariance analysis, which is similar to the differential covariance distance vector analysis in Hu, et al. (2010), but replacing the covariance distance matrix in their paper by the covariance matrix. The second method selects the i-th gene if the i-th row of $\hat{\Psi}(4)$ is nonzero. This leads to the selection of 43 genes. The third method, defined in (17), controls the family-wise error rate at $\alpha = 0.1$, and is able to find 52 genes. As expected, the gene selection based on the covariances could be more powerful than the one that is only based on the variances.

Finally, we apply the support recovery procedure $\hat{\Psi}(4)$ to $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$. For a visual comparison between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, Figure 2 plots the heat map of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ of the 200 largest absolute values of two sample t statistics. It can be seen from Figure 2 that the estimated support of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ is quite sparse.

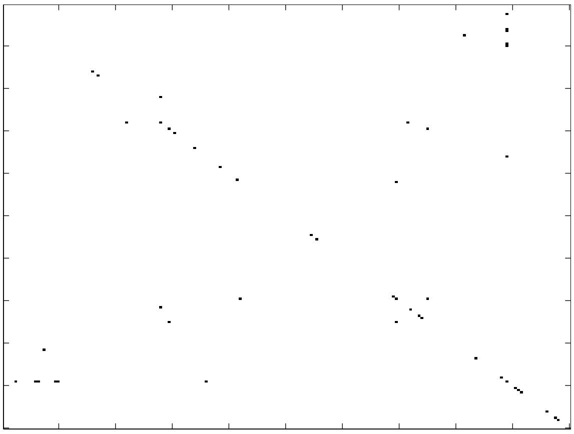


Figure 2: Heat map of the the selected genes by exactly recovery.

6 Discussion

We introduced in this paper a test for the equality of two covariance matrices which is proved to have the prespecified significance level asymptotically and to enjoy certain optimality in terms of its power. In particular, we show both theoretically and numerically that the test is especially powerful against sparse alternatives. Support recovery and testing two covariance matrices row by row with applications to gene selection are also considered. There are several possible extensions of our method. For example, an interesting direction is to generalize the procedure to testing the hypothesis of homogeneity of several covariance matrices, $H_0 : \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_K$, where $K \geq 2$. A test statistic similar to M_n can be constructed to test this hypothesis and analogous theoretical results can be developed. We shall report the details of the results elsewhere in the future as a significant amount of additional work is still needed.

The existing tests for testing the equality of two covariance matrices such as those proposed in Schott (2007), Srivastava and Yanagihara (2010), and Chen and Li (2011) are based on the Frobenius norm and require $n\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_F^2/p \to \infty$ for the tests to be able to distinguish between the null and the alternative with probability tending to 1. These tests are not suited for testing sparse alternatives. If $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ is a sparse matrix and the number of nonzero entries of $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ is of order $o(p/\log p)$ and their magnitudes are of order $C\sqrt{\log p/n}$, then it can be shown that the powers of these tests are trivial. This fact was also illustrated in some of the simulation results given in Section 5.

Much recent attention has focused on the estimation of large covariance and precision matrices. The current work here is related to the estimation of covariance matrices. An adaptive thresholding estimator of sparse covariance matrices was introduced in Cai and Liu (2011). The procedure is based on the standardized statistics $\hat{\sigma}_{ij}/\hat{\theta}_{ij}^{1/2}$, which is closely related to M_{ij}. In this paper, the asymptotic distribution of $M_n = \max_{1\le i\le j\le p} M_{ij}$ is obtained. It gives an justification on the family-wise error of simultaneous tests $H_{0ij} : \sigma_{ij1} = \sigma_{ij2}$ for $1 \le i \le j \le p$. For example, by thresholding M_{ij} at level $4\log p - \log\log p + q_\alpha$, the family-wise error is controlled asymptotically at level α, i.e.

$$\text{FWER} = \mathsf{P}\Big(\max_{(i,j)\in G} M_{ij} \ge 4\log p - \log\log p + q_\alpha\Big) \to \alpha,$$

where $G = \{(i,j) : \sigma_{ij1} = \sigma_{ij2}\}$ and $\text{Card}(G^c) = o(p^2)$. These tests are useful in the studies of differential coexpression in genetics; see de la Fuentea (2010).

7 Proofs

In this section, we will prove the main results. The proofs of Propositions 1-3 are given in the supplementary material Cai, Liu and Xia (2011). Throughout this section, we denote by C, C_1, $C_2, \ldots$, constants which do not depend on n, p and may vary from place to place. We begin by collecting some technical lemmas that will be used in the proofs of the main results. These technical lemmas are proved in the supplementary material, Cai, Liu and Xia (2011).

7.1 Technical lemmas

The first lemma is the classical Bonferroni's inequality.

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

$$\sum_{t=1}^{2k} (-1)^{t-1} E_t \leq \mathsf{P}(B) \leq \sum_{t=1}^{2k-1} (-1)^{t-1} E_t,$$

where $E_t = \sum_{1 \leq i_1 < \cdots < i_t \leq p} \mathsf{P}(B_{i_1} \cap \cdots \cap B_{i_t})$.

The second lemma comes from Berman (1962).

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

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

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

The next lemma is on the large deviations for $\hat{\theta}_{ij1}$ and $\hat{\theta}_{ij2}$.

Lemma 3 *Under the conditions of (C2) or $(C2^*)$, there exists some constant $C > 0$ such that*

$$\mathsf{P}\Big(\max_{i,j} |\hat{\theta}_{ij1} - \theta_{ij1}|/\sigma_{ii1}\sigma_{jj1} \geq C\frac{\varepsilon_n}{\log p}\Big) = O(p^{-1} + n^{-\epsilon/8}), \tag{18}$$

and

$$\mathsf{P}\Big(\max_{i,j} |\hat{\theta}_{ij2} - \theta_{ij2}|/\sigma_{ii2}\sigma_{jj2} \geq C\frac{\varepsilon_n}{\log p}\Big) = O(p^{-1} + n^{-\epsilon/8}), \tag{19}$$

where $\varepsilon_n = \max((\log p)^{1/6}/n^{1/2}, (\log p)^{-1}) \to 0$ as $n, p \to \infty$.

Define

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_1 &= (\tilde{\sigma}_{ij1})_{p\times p} = \frac{1}{n_1}\sum_{k=1}^{n_1} (\boldsymbol{X} - \boldsymbol{\mu}_1)(\boldsymbol{X} - \boldsymbol{\mu}_1)^T, \\
\tilde{\mathbf{\Sigma}}_2 &= (\tilde{\sigma}_{ij2})_{p\times p} = \frac{1}{n_2}\sum_{k=1}^{n_2} (\boldsymbol{Y} - \boldsymbol{\mu}_2)(\boldsymbol{Y} - \boldsymbol{\mu}_2)^T.
\end{aligned}$$

Let Λ be any subset of $\{(i,j) : 1 \leq i \leq j \leq p\}$ and $|\Lambda|$ =Card(Λ).

Lemma 4 *Under the conditions of (C2) or (C2*), we have for some constant $C > 0$ that*

$$P\Big(\max_{(i,j)\in\Lambda}\frac{(\tilde{\sigma}_{ij1}-\tilde{\sigma}_{ij2}-\sigma_{ij1}+\sigma_{ij2})^2}{\theta_{ij1}/n_1+\theta_{ij2}/n_2}\geq x^2\Big)\leq C|\Lambda|(1-\Phi(x))+O(p^{-1}+n^{-\epsilon/8})$$

uniformly for $0\leq x\leq (8\log p)^{1/2}$ and $\Lambda\subseteq\{(i,j):1\leq i\leq j\leq p\}$.

The proofs of Lemmas 3 and 4 are given in the supplementary material Cai, Liu and Xia (2011).

7.2 Proof of Theorem 1

Without loss of generality, we assume that $\boldsymbol{\mu}_1=0$, $\boldsymbol{\mu}_2=0$, $\sigma_{ii1}=\sigma_{ii2}=1$ for $1\leq i\leq p$. Instead of the boundedness condition on the eigenvalues, we shall prove the theorem under the more general condition that for any $\gamma>0$, $\max_j s_j=O(p^{\gamma})$, where

$$s_j:=\mathrm{card}\{i:|\rho_{ij1}|\geq(\log p)^{-1-\alpha_0}\text{ or }|\rho_{ij2}|\geq(\log p)^{-1-\alpha_0}\}$$

with some $\alpha_0>0$. (Note that $\lambda_{\max}(\boldsymbol{R}_1)\leq C_0$ and $\lambda_{\max}(\boldsymbol{R}_1)\leq C_0$ implies that $\max_j s_j\leq C(\log p)^{2+2\alpha_0}$.) Let

$$\hat{M}_n=\max_{1\leq i\leq j\leq p}\frac{(\hat{\sigma}_{ij1}-\hat{\sigma}_{ij2})^2}{\theta_{ij1}/n_1+\theta_{ij2}/n_2},\quad \tilde{M}_n=\max_{1\leq i\leq j\leq p}\frac{(\tilde{\sigma}_{ij1}-\tilde{\sigma}_{ij2})^2}{\theta_{ij1}/n_1+\theta_{ij2}/n_2}.$$

Note that under the event $\{|\hat{\theta}_{ij1}/\theta_{ij1}-1|\leq C\varepsilon_n/\log p, |\hat{\theta}_{ij2}/\theta_{ij2}-1|\leq C\varepsilon_n/\log p\}$, we have

$$\begin{aligned}&\Big|M_n-\hat{M}_n\Big|\leq C\hat{M}_n\frac{\varepsilon_n}{\log p},\\&\Big|\hat{M}_n-\tilde{M}_n\Big|\leq C(n_1+n_2)(\max_{1\leq i\leq p}\bar{X}_i^4+\max_{1\leq i\leq p}\bar{Y}_i^4)+C(n_1+n_2)^{1/2}\tilde{M}_n^{1/2}(\max_{1\leq i\leq p}\bar{X}_i^2+\max_{1\leq i\leq p}\bar{Y}_i^2).\end{aligned}$$

By the exponential inequality, $\max_{1\leq i\leq p}|\bar{X}_i|+\max_{1\leq i\leq p}|\bar{Y}_i|=O_{\mathsf{P}}(\sqrt{\log p/n})$. Thus, by Lemma 3, it suffices to show that for any $x\in R$,

$$\mathsf{P}\Big(\tilde{M}_n-4\log p+\log\log p\leq x\Big)\to\exp\Big(-\frac{1}{\sqrt{8\pi}}\exp\Big(-\frac{x}{2}\Big)\Big)\qquad(20)$$

as $n, p \to \infty$. Let $\rho_{ij} = \rho_{ij1} = \rho_{ij2}$ under H_0. Define

$$\begin{aligned}
&A_j = \Big\{i : |\rho_{ij}| \geq (\log p)^{-1-\alpha_0}\Big\},\\
&A = \{(i,j) : 1 \leq i \leq j \leq p\},\\
&A_0 = \{(i,j) : 1 \leq i \leq p, i \leq j, j \in A_i\},\\
&B_0 = \{(i,j) : i \in \Lambda(r), i < j \leq p\} \cup \{(i,j) : j \in \Lambda(r), 1 \leq i < j\},\\
&D_0 = A_0 \cup B_0.
\end{aligned}$$

By the definition of D_0, for any $(i_1, j_1) \in A \setminus D_0$ and $(i_2, j_2) \in A \setminus D_0$, we have $|\rho_{kl}| \leq r$ for any $k \neq l \in \{i_1, j_1, i_2, j_2\}$. Set

$$\tilde{M}_{A\setminus D_0} = \max_{(i,j)\in A\setminus D_0} \frac{(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}, \quad \tilde{M}_{D_0} = \max_{(i,j)\in D_0} \frac{(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}.$$

Let $y_p = x + 4\log p - \log\log p$. Then

$$\Big|\mathsf{P}\Big(\tilde{M}_n \geq y_p\Big) - \mathsf{P}\Big(\tilde{M}_{A\setminus D_0} \geq y_p\Big)\Big| \leq \mathsf{P}\Big(\tilde{M}_{D_0} \geq y_p\Big).$$

Note that $\mathrm{Card}(\Lambda(r)) = o(p)$ and for any $\gamma > 0$, $\max_j s_j = O(p^\gamma)$. This implies that $\mathrm{Card}(D_0) \leq C_0 p^{1+\gamma} + o(p^2)$ for any $\gamma > 0$. By Lemma 4, we obtain that for any fixed $x \in R$,

$$\mathsf{P}\Big(\tilde{M}_{D_0} \geq y_p\Big) \leq \ \mathrm{Card}(D_0) \times Cp^{-2} + o(1) = o(1).$$

Set

$$\tilde{M}_{A\setminus D_0} = \max_{(i,j)\in A\setminus D_0} \frac{(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2} =: \max_{(i,j)\in A\setminus D_0} U_{ij}^2,$$

where

$$U_{ij} = \frac{\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2}}{\sqrt{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}}.$$

It is enough to show that for any $x \in R$,

$$\mathsf{P}\Big(\tilde{M}_{A\setminus D_0} - 4\log p + \log\log p \leq x\Big) \to \exp\Big(-\frac{1}{\sqrt{8\pi}}\exp\Big(-\frac{x}{2}\Big)\Big)$$

as $n, p \to \infty$. We arrange the two dimensional indices $\{(i,j) : (i,j) \in A \setminus D_0\}$ in any ordering and set them as $\{(i_m, j_m) : 1 \le m \le q\}$ with q =Card$(A \setminus D_0)$. Let $n_2/n_1 \le K_1$ with $K_1 \ge 1$, $\theta_{k1} = \theta_{i_k j_k 1}$, $\theta_{k2} = \theta_{i_k j_k 2}$ and define

$$\begin{aligned}
Z_{lk} &= \frac{n_2}{n_1}(X_{li_k}X_{lj_k} - \sigma_{i_kj_k1}), \quad 1 \le l \le n_1,\\
Z_{lk} &= -(Y_{li_k}Y_{lj_k} - \sigma_{i_kj_k2}), \quad n_1 + 1 \le l \le n_1 + n_2,\\
V_k &= \frac{1}{\sqrt{n_2^2\theta_{k1}/n_1 + n_2\theta_{k2}}} \sum_{l=1}^{n_1+n_2} Z_{lk},\\
\hat{V}_k &= \frac{1}{\sqrt{n_2^2\theta_{k1}/n_1 + n_2\theta_{k2}}} \sum_{l=1}^{n_1+n_2} \hat{Z}_{lk},
\end{aligned}$$

where

$$\hat{Z}_{lk} = Z_{lk}I\{|Z_{lk}| \le \tau_n\} - \mathsf{E}Z_{lk}I\{|Z_{lk}| \le \tau_n\},$$

and $\tau_n = \eta^{-1}8K_1 \log(p+n)$ if (C2) holds, $\tau_n = \sqrt{n}/(\log p)^8$ if (C2*) holds. Note that

$$\max_{(i,j)\in A\setminus D_0} U_{ij}^2 = \max_{1\le k\le q} V_k^2.$$

We have, if (C2) holds, then

$$\begin{aligned}
&\max_{1\le k\le q} \frac{1}{\sqrt{n}} \sum_{l=1}^{n_1+n_2} \mathsf{E}|Z_{lk}|I\{|Z_{lk}| \ge \eta^{-1}8K_1\log(p+n)\}\\
&\quad\le C\sqrt{n} \max_{1\le l\le n_1+n_2} \max_{1\le k\le q} \mathsf{E}|Z_{lk}|I\{|Z_{lk}| \ge \eta^{-1}8K_1\log(p+n)\}\\
&\quad\le C\sqrt{n}(p+n)^{-4} \max_{1\le l\le n_1+n_2} \max_{1\le k\le q} \mathsf{E}|Z_{lk}|\exp(\eta|Z_{lk}|/(2K_1))\\
&\quad\le C\sqrt{n}(p+n)^{-4},
\end{aligned}$$

and if (C2*) holds, then

$$\begin{aligned}
&\max_{1\le k\le q} \frac{1}{\sqrt{n}} \sum_{l=1}^{n_1+n_2} \mathsf{E}|Z_{lk}|I\{|Z_{lk}| \ge \sqrt{n}/(\log p)^8\}\\
&\quad\le C\sqrt{n} \max_{1\le l\le n_1+n_2} \max_{1\le k\le q} \mathsf{E}|Z_{lk}|I\{|Z_{lk}| \ge \sqrt{n}/(\log p)^8\}\\
&\quad\le Cn^{-\gamma_0-\epsilon/8}.
\end{aligned}$$

Hence we have

$$\mathsf{P}\Big(\max_{1\le k\le q} |V_k - \hat{V}_k| \ge (\log p)^{-1}\Big) \;\le\; \mathsf{P}\Big(\max_{1\le k\le q} \max_{1\le l\le n_1+n_2} |Z_{lk}| \ge \tau_n\Big)$$

$$
\begin{aligned}
&\leq \; np \max_{1\leq j\leq p}\Big[\mathsf{P}\Big(X_j^2 \geq \tau_n/2\Big) + \mathsf{P}\Big(Y_j^2 \geq \tau_n/2\Big)\Big] \\
&= \; O(p^{-1} + n^{-\epsilon/8}). \qquad (21)
\end{aligned}
$$

Note that

$$
\Big|\max_{1\leq k\leq q} V_k^2 - \max_{1\leq k\leq q} \hat{V}_k^2\Big| \leq 2\max_{1\leq k\leq q}|\hat{V}_k| \max_{1\leq k\leq q}|V_k - \hat{V}_k| + \max_{1\leq k\leq q}|V_k - \hat{V}_k|^2. \qquad (22)
$$

By (21) and (22), it is enough to prove that for any $x \in R$,

$$
\mathsf{P}\Big(\max_{1\leq k\leq q} \hat{V}_k^2 - 4\log p + \log\log p \leq x\Big) \to \exp\Big(-\frac{1}{\sqrt{8\pi}}\exp\Big(-\frac{x}{2}\Big)\Big) \qquad (23)
$$

as $n, p \to \infty$. By Bonferroni inequality (see Lemma 1), we have for any fixed integer m with $0 < m < q/2$,

$$
\begin{aligned}
&\sum_{d=1}^{2m}(-1)^{d-1}\sum_{1\leq k_1<\cdots<k_d\leq q}\mathsf{P}\Big(\bigcap_{j=1}^{d}E_{k_j}\Big) \\
&\quad \leq \mathsf{P}\Big(\max_{1\leq k\leq q}\hat{V}_k^2 \geq y_p\Big) \\
&\quad\quad \leq \sum_{d=1}^{2m-1}(-1)^{d-1}\sum_{1\leq k_1<\cdots<k_d\leq q}\mathsf{P}\Big(\bigcap_{j=1}^{d}E_{k_j}\Big), \qquad (24)
\end{aligned}
$$

where $E_{k_j} = \{\hat{V}_{k_j}^2 \geq y_p\}$. Let $\tilde{Z}_{lk} = \hat{Z}_{lk}/(n_2\theta_{k1}/n_1 + \theta_{k2})^{1/2}$ for $1 \leq k \leq q$ and $\boldsymbol{W}_l = (\tilde{Z}_{lk_1}, \ldots, \tilde{Z}_{lk_d})$, for $1 \leq l \leq n_1 + n_2$. Define $|\boldsymbol{a}|_{\min} = \min_{1\leq i\leq d}|a_i|$ for any vector $\boldsymbol{a} \in R^d$. Then we have

$$
\mathsf{P}\Big(\bigcap_{j=1}^{d}E_{k_j}\Big) = \mathsf{P}\Big(\Big|n_2^{-1/2}\sum_{l=1}^{n_1+n_2}\boldsymbol{W}_l\Big|_{\min} \geq y_n^{1/2}\Big).
$$

By Theorem 1 in Zaitsev (1987), we have

$$
\begin{aligned}
\mathsf{P}\Big(\Big|n_2^{-1/2}\sum_{l=1}^{n_1+n_2}\boldsymbol{W}_l\Big|_{\min} \geq y_n^{1/2}\Big) \;&\leq\; \mathsf{P}\Big(|\boldsymbol{N}_d|_{\min} \geq y_p^{1/2} - \epsilon_n(\log p)^{-1/2}\Big) \\
&\quad + c_1 d^{5/2}\exp\Big(-\frac{n^{1/2}\epsilon_n}{c_2 d^3\tau_n(\log p)^{1/2}}\Big), \qquad (25)
\end{aligned}
$$

where $c_1 > 0$ and $c_2 > 0$ are absolutely constants, $\epsilon_n \to 0$ which will be specified later and $\boldsymbol{N}_d =: (N_{k_1}, \ldots, N_{k_d})$ is a d dimensional normal vector with $\mathsf{E}\boldsymbol{N}_d = 0$ and

$\mathsf{Cov}(\boldsymbol{N}_d) = \frac{n_1}{n_2}\mathsf{Cov}(\boldsymbol{W}_1) + \mathsf{Cov}(\boldsymbol{W}_{n_1+1})$. Recall that d is a fixed integer which does not depend on n, p. Because $\log p = o(n^{1/5})$, we can let $\epsilon_n \to 0$ sufficiently slow such that

$$c_1 d^{5/2} \exp\Big(- \frac{n^{1/2}\epsilon_n}{c_2 d^3 \tau_n (\log p)^{1/2}}\Big) = O(p^{-M}) \tag{26}$$

for any large $M > 0$. It follows from (24), (25) and (26) that

$$\begin{aligned} &\mathsf{P}\Big(\max_{1\le k\le q} \hat{V}_k^2 \ge y_p\Big) \\ &\quad \le \sum_{d=1}^{2m-1} (-1)^{d-1} \sum_{1\le k_1<\cdots<k_d\le q} \mathsf{P}\Big(|\boldsymbol{N}_d|_{\min} \ge y_p^{1/2} - \epsilon_n (\log p)^{-1/2}\Big) + o(1). \end{aligned} \tag{27}$$

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

$$\begin{aligned} &\mathsf{P}\Big(\max_{1\le k\le q} \hat{V}_k^2 \ge y_p\Big) \\ &\quad \ge \sum_{d=1}^{2m} (-1)^{d-1} \sum_{1\le k_1<\cdots<k_d\le q} \mathsf{P}\Big(|\boldsymbol{N}_d|_{\min} \ge y_p^{1/2} + \epsilon_n (\log p)^{-1/2}\Big) - o(1). \end{aligned} \tag{28}$$

Lemma 5 *For any fixed integer $d \ge 1$ and real number $x \in R$,*

$$\sum_{1\le k_1<\cdots<k_d\le q} P\Big(|\boldsymbol{N}_d|_{\min} \ge y_p^{1/2} \pm \epsilon_n (\log p)^{-1/2}\Big) = (1+o(1))\frac{1}{d!}\Big(\frac{1}{\sqrt{8\pi}}\exp(-\frac{x}{2})\Big)^d. \tag{29}$$

The proof of Lemma 5 is very complicated, and is given in the supplementary material Cai, Liu and Xia (2011). Submitting (29) into (27) and (28), we can get

$$\limsup_{n,p\to\infty} \mathsf{P}\Big(\max_{1\le k\le q} \hat{V}_k^2 \ge y_p\Big) \le \sum_{d=1}^{2m} (-1)^{d-1} \frac{1}{d!}\Big(\frac{1}{\sqrt{8\pi}}\exp(-\frac{x}{2})\Big)^d$$

and

$$\liminf_{n,p\to\infty} \mathsf{P}\Big(\max_{1\le k\le q} \hat{V}_k^2 \ge y_p\Big) \ge \sum_{d=1}^{2m-1} (-1)^{d-1} \frac{1}{d!}\Big(\frac{1}{\sqrt{8\pi}}\exp(-\frac{x}{2})\Big)^d$$

for any positive integer m. Letting $m \to \infty$, we obtain (23). The theorem is proved. ∎

7.3 Proof of the other theorems

Proof of Theorem 2. Recall that

$$M_n^1 = \max_{1\le i\le j\le p} \frac{(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2} - \sigma_{ij1} + \sigma_{ij2})^2}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2}.$$

By Lemmas 3 and 4,

$$\mathsf{P}\Big(M_n^1 \le 4\log p - \frac{1}{2}\log\log p\Big) \to 1 \tag{30}$$

as $n, p \to \infty$. By Lemma 3, the inequalities

$$\max_{1\le i\le j\le p} \frac{(\sigma_{ij1} - \sigma_{ij2})^2}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2} \le 2M_n^1 + 2M_n \tag{31}$$

and

$$\max_{1\le i\le j\le p} \frac{(\sigma_{ij1} - \sigma_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2} \ge 16\log p,$$

we obtain that

$$\mathsf{P}\Big(M_n \ge q_\alpha + 4\log p - \log\log p\Big) \to 1$$

as $n, p \to \infty$. The proof of Theorem 2 is complete. ▮

Proof of Theorem 4 Let $\mathcal{M}$ denote the set of all subsets of $\{1, ..., p\}$ with cardinality $c_0(p)$. Let $\hat{m}$ be a random subset of $\{1, ..., p\}$, which is uniformly distributed on $\mathcal{M}$. We construct a class of $\mathbf{\Sigma}_1$, $\mathcal{N} = \{\mathbf{\Sigma}_{\hat{m}}, \hat{m} \in \mathcal{M}\}$, such that

$$\sigma_{ij} = 0 \text{ for } i \ne j \text{ and } \sigma_{ii} - 1 = \rho \mathbf{1}_{i\in\hat{m}},$$

for $i, j = 1, ..., p$ and $\rho = c\sqrt{\log p/n}$, where $c > 0$ will be specified later. Let $\mathbf{\Sigma}_2 = \boldsymbol{I}$ and $\mathbf{\Sigma}_1$ be uniformly distributed on $\mathcal{N}$. Let μ_ρ be the distribution of $\mathbf{\Sigma}_1 - \boldsymbol{I}$. Note that μ_ρ is a probability measure on $\{\Delta \in \mathcal{S}(c_0(p)) : \|\Delta\|_F^2 = c_0(p)\rho^2\}$, where $\mathcal{S}(c_0(p))$ is the class of matrices with $c_0(p)$ nonzero entries. Let $d\mathsf{P}_1(\{\boldsymbol{X}_n, \boldsymbol{Y}_n\})$ be the likelihood function given $\mathbf{\Sigma}_1$ being uniformly distributed on $\mathcal{N}$ and

$$L_{\mu_\rho} := L_{\mu_\rho}(\{\boldsymbol{X}_n, \boldsymbol{Y}_n\}) = \mathsf{E}_{\mu_\rho}\Big(\frac{d\mathsf{P}_1(\{\boldsymbol{X}_n, \boldsymbol{Y}_n\})}{d\mathsf{P}_0(\{\boldsymbol{X}_n, \boldsymbol{Y}_n\})}\Big),$$

where $\mathsf{E}_{\mu_\rho}(\cdot)$ is the expectation on $\boldsymbol{\Sigma}_1$. By the arguments in Baraud(2002), page 595, it suffices to show that

$$\mathsf{E}L^2_{\mu_\rho} \le 1 + o(1).$$

It is easy to see that

$$L_{\mu_\rho} = \mathsf{E}_{\hat{m}}\Big(\prod_{i=1}^{n} \frac{1}{|\boldsymbol{\Sigma}_{\hat{m}}|^{1/2}} \exp\Big(-\frac{1}{2}\boldsymbol{Z}_i^T(\boldsymbol{\Omega}_{\hat{m}} - \boldsymbol{I})\boldsymbol{Z}_i\Big)\Big),$$

where $\boldsymbol{\Omega}_{\hat{m}} = \boldsymbol{\Sigma}_{\hat{m}}^{-1}$ and $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ are i.i.d multivariate normal vectors with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{I}$. Thus,

$$\begin{aligned}
\mathsf{E}L^2_{\mu_\rho} &= \mathsf{E}\Big(\frac{1}{\binom{p}{k_p}} \sum_{m\in\mathcal{M}} \Big(\prod_{i=1}^{n} \frac{1}{|\boldsymbol{\Sigma}_m|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}\boldsymbol{Z}_i^T(\boldsymbol{\Omega}_m - \boldsymbol{I})\boldsymbol{Z}_i\Big)\Big)\Big)^2 \\
&= \frac{1}{\binom{p}{k_p}^2} \sum_{m,m'\in\mathcal{M}} \mathsf{E}\Big(\prod_{i=1}^{n} \frac{1}{|\boldsymbol{\Sigma}_m|^{\frac{1}{2}}} \frac{1}{|\boldsymbol{\Sigma}_{m'}|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}\boldsymbol{Z}_i^T(\boldsymbol{\Omega}_m + \boldsymbol{\Omega}_{m'} - 2\boldsymbol{I})\boldsymbol{Z}_i\Big)\Big).
\end{aligned}$$

Set $\boldsymbol{\Omega}_m + \boldsymbol{\Omega}_{m'} - 2\boldsymbol{I} = (a_{ij})$. Then $a_{ij} = 0$ for $i \neq j$, $a_{jj} = 0$ if $j \in (m \cup m')^c$, $a_{jj} = 2(\frac{1}{1+\rho} - 1)$ if $j \in m \cap m'$ and $a_{jj} = \frac{1}{1+\rho} - 1$ if $j \in m \setminus m'$ and $m' \setminus m$. Let $t = |m \cap m'|$. Then we have

$$\begin{aligned}
\mathsf{E}L^2_{\mu_\rho} &= \frac{1}{\binom{p}{k_p}} \sum_{t=0}^{k_p} \binom{k_p}{t}\binom{p-k_p}{k_p-t} \frac{1}{(1+\rho)^{k_p n}} (1+\rho)^{(k_p-t)n} \Big(\frac{1+\rho}{1-\rho}\Big)^{tn/2} \\
&\le p^{k_p}(p-k_p)!/p! \sum_{t=0}^{k_p} \binom{k_p}{t} \Big(\frac{k_p}{p}\Big)^t \Big(\frac{1}{1-\rho^2}\Big)^{tn/2} \\
&= (1+o(1))\Big(1 + \frac{k_p}{p(1-\rho^2)^{n/2}}\Big)^{k_p},
\end{aligned}$$

for $r < 1/2$. So

$$\begin{aligned}
\mathsf{E}L^2_{\mu_\rho} &\le (1+o(1)) \exp\Big(k_p \log(1 + k_p p^{c^2-1})\Big) \\
&\le (1+o(1)) \exp\Big(k_p^2 p^{c^2-1}\Big) \\
&= 1 + o(1)
\end{aligned}$$

by letting c be sufficiently small. Theorem 4 is proved. ∎

Proof of Theorem 5. The proof of Theorem 5 is similar to that of Theorem 2. In fact, by (30) and a similar inequality as (31), we can get

$$\mathsf{P}\Big(\min_{(i,j)\in\Psi} M_{ij} \geq 4\log p\Big) \to 1 \tag{32}$$

uniformly for $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathcal{W}_0(4)$. ▮

Proof of Theorem 6. Let A_1 be the largest subset of $\{1, \cdots, p\} \setminus \{1\}$ such that $\sigma_{1k1} = \sigma_{1k2}$ for all $k \in A_1$. Let $i_1 = \min\{j : j \in A_1, j > 1\}$. Then we have $|i_1 - 1| \leq s_0(p)$. Also, $\mathrm{Card}(A_1) \geq p - s_0(p)$. Similarly, let A_l be the largest subset of $A_{l-1} \setminus \{i_{l-1}\}$ such that $\sigma_{i_{l-1}k1} = \sigma_{i_{l-1}k2}$ for all $k \in A_l$ and $i_l = \min\{j : j \in A_l, j > i_{l-1}\}$. We can see that $i_l - i_{l-1} \leq s_0(p)$ for $l < p/s_0(p)$ and $\mathrm{Card}(A_l) \geq \mathrm{Card}(A_{l-1}) - s_0(p) \geq p - (s_0(p) + 1)l$. Let $l = [p^{\tau_2}]$ with $\tau/4 < \tau_2 < \tau_1$. Let $\mathbf{\Sigma}_{1l}$ and $\mathbf{\Sigma}_{2l}$ be the covariance matrices of $(X_{i_0}, \ldots, X_{i_l})$ and $(Y_{i_0}, \ldots, Y_{i_l})$. Then the entries of $\mathbf{\Sigma}_{1l}$ and $\mathbf{\Sigma}_{2l}$ are the same except for the diagonal. Hence by the proof of Theorem 1, we can show that

$$\mathsf{P}\Big(\max_{0\leq j<k\leq l} M_{i_j i_k} - 4\log l + \log\log l \leq x\Big) \to \exp\Big(-\frac{1}{\sqrt{8\pi}}\exp\Big(-\frac{x}{2}\Big)\Big)$$

uniformly for all $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \in \mathcal{S}_0$. This implies that

$$\inf_{\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \in \mathcal{S}_0} \mathsf{P}\Big(\max_{0\leq j<k\leq l} M_{i_j i_k} \geq c\log p\Big) \to 1$$

for all $\tau < c < 4\tau_2$. By the definition of $\hat{\Psi}(\tau)$ and the fact $\sigma_{i_j i_k 1} = \sigma_{i_j i_k 2}$ for all $0 \leq j < k \leq l$, Theorem 6 is proved. ▮

References

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

[2] Berman, S.M. (1962), Limiting distribution of the maximum term in sequences of dependent random variables. *The Annals of Mathematical Statistics*, 33, 894-908.

[3] Cai, T. and Jiang, T. (2011), Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39, 1496-1525.

[4] Cai, T. and Liu, W.D. (2011), Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106, 672-684.

[5] Cai, T., Liu, W.D. and Xia, Y. (2011), Supplement to "Two-sample covariance matrix testing and support recovery". Technical report.

[6] Chen, S. and Li, J. (2011), Two sample tests for high dimensional covariance matrices. Technical report.

[7] de la Fuente, A. (2010), From "differential expression" to "differential networking"-identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26, 326-333.

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

[9] Gupta, D.S. and Giri, N. (1973), Properties of tests concerning covariance matrices of normal distributions. *The Annals of Statistics*, 6, 1222-1224.

[10] Ho, J.W., Stefani, M., dos Remedios, C.G. and Charleston, M.A. (2008), Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, 24, 390-398.

[11] Hu, R, Qiu, X. and Glazko, G. (2010), A new gene selection procedure based on the covariance distance. *Bioinformatics*, 26, 348-354.

[12] Hu, R, Qiu, X., Glazko, G., Klevanov, L. and Yakovlev, A. (2009), Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. *BMC Bioinformatics*, 10:20.

[13] Jiang, T. (2004), The asymptotic distributions of the largest entries of sample correlation matrices. *The Annals of Applied Probability*, 14, 865-880.

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

[15] O'Brien, P.C. (1992), Robust procedures for testing equality of covariance matrices. *Biometrics*, 48, 819-827.

[16] Perlman, M.D. (1980), Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. *The Annals of Statistics*, 8, 247-263.

[17] Sugiura, N. and Nagao, H. (1968), Unbiasedness of some test criteria for the equality of one or two covariance matrices. *The Annals of Mathematical Statistics*, 39, 1682-1692.

[18] Schott, J. R. (2007), A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics and Data Analysis*, 51, 6535-6542.

[19] Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D$'$Amico, A. and Richie, J. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1, 203-209.

[20] Srivastava, M. S. and Yanagihara, H. (2010), Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, 101, 1319-1329.

[21] Xu, P., Brock, G.N. and Parrish, R.S. (2009), Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics and Data Analysis*, 53, 1674-1687,

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

[23] Zhou, W. (2007), Asymptotic distribution of the largest off-diagonal entry of correlation matrices. *Transactions of the American Mathematical Society*, 359, 5345-5364.