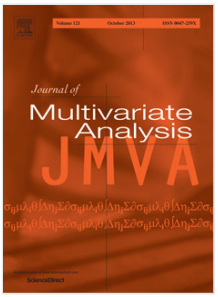
Accepted Manuscript



Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data

Kai Dong, Herbert Pang, Tiejun Tong, Marc G. Genton

PII: S0047-259X(15)00214-6
DOI: http://dx.doi.org/10.1016/j.jmva.2015.08.022
Reference: YJMVA 3997

To appear in: *Journal of Multivariate Analysis*

Received date: 25 June 2014

Please cite this article as: K. Dong, H. Pang, T. Tong, M.G. Genton, Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data, *Journal of Multivariate Analysis* (2015), http://dx.doi.org/10.1016/j.jmva.2015.08.022

Shrinkage-Based Diagonal Hotelling's Tests for High-Dimensional Small Sample Size Data

Kai Dong[1], Herbert Pang[2,3], Tiejun Tong[1,*] and Marc G. Genton[4]

[1]Department of Mathematics, Hong Kong Baptist University, Hong Kong

[2]School of Public Health, The University of Hong Kong, Hong Kong

[3]Department of Biostatistics and Bioinformatics, Duke University, USA

[4]CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia

∗Email: tongt@hkbu.edu.hk

July 3, 2015

Abstract

DNA sequencing techniques bring novel tools and also statistical challenges to genetic research. In addition to detecting differentially expressed genes, testing the significance of gene sets or pathway analysis has been recognized as an equally important problem. Owing to the "large p small n" paradigm, the traditional Hotelling's T^2 test suffers from the singularity problem and therefore is not valid in this setting. In this paper, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. We also suggest several different ways to derive the approximate null distribution under different scenarios of p and n for our proposed shrinkage-based test. Simulation studies show that the proposed method performs comparably to existing competitors when n is moderate or large, but it is better when n is small. In addition, we analyze four gene expression data sets and they demonstrate the advantage of our proposed shrinkage-based diagonal Hotelling's test.

KEY WORDS: Diagonal Hotelling's test; High-dimensional data; Microarray data; Null distribution; Optimal variance estimation.

1 Introduction

DNA microarrays allow us to acquire thousands or even millions of gene expression values simultaneously, which introduces novel approaches to genetic research. One important goal of analyzing gene expression microarray data is to detect differentially expressed genes. Recently, biologists and medical scientists have also recognized that testing the significance of gene sets or pathway analysis is an equally important problem (Efron and Tibshirani 2007, Newton et al. 2007, Chen and Qin 2010, Maciejewski 2013). Specifically, if we want to know whether a certain gene set, Z, is significantly differentially expressed in two different treatments, A and B, the testing hypothesis is $H_0 : \boldsymbol{\mu}_{\mathbf{Z}A} = \boldsymbol{\mu}_{\mathbf{Z}B}$, where $\boldsymbol{\mu}_{\mathbf{Z}A}$ and $\boldsymbol{\mu}_{\mathbf{Z}B}$ are the mean vectors of $\mathbf{Z}$ in A and B, respectively. In statistics, this is essentially a two-sample multivariate testing problem. One classical method used to solve such testing problems is Hotelling's T^2 test (Hotelling 1931), which is a generalization of Student's t test. This method works when the sample size, n, is larger than the data dimension, p. More generally, in a k-sample experiment, we are interested in whether or not there exist some differences among the k mean vectors of populations.

In this paper, we focus on one-sample and two-sample multivariate testing problems for high-dimensional small sample size data, or equivalently, for "large p small n" data. In such settings, Hotelling's T^2 test suffers from a singularity problem in the covariance matrix estimation and therefore is not valid in this setting. To overcome the singularity problem, some remedies have been proposed in the literature; see, for example, the non-exact significance test and the randomization test in Dempster (1960). These approaches, however, are known to perform poorly in practice due to their complicated estimation of the degrees of freedom and some related issues (Bai and Saranadasa 1996). In recent years, a number of approaches to improve Hotelling's T^2 test have emerged for testing high-dimensional data. In essence, these approaches can be classified into the following three categories, with the main difference among them how the covariance matrix is handled:

1) In the first category, the covariance matrix is removed from Hotelling's T^2 statistic to avoid the covariance matrix estimation. This idea was first considered by Bai and Saranadasa (1996). Specifically, they proposed to use $(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$ to replace $(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T\mathbf{S}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$ in Hotelling's T^2 statistic, where $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ are the sample mean vectors and $\mathbf{S}$ is the pooled sample covariance matrix. They demonstrated that the proposed test has better power than Hotelling's T^2 test under the requirement of p and n being of the same order. Recently, Zhang and Xu (2009) and Chen and Qin (2010) extended this method to "large p small n" data. We refer to the methods in this category as *the unscaled Hotelling's tests*.

2) In the second category, a regularization method is applied to the covariance matrix estimation to resolve the singularity problem. In this direction, Chen et al. (2011) have made a major contribution. They proposed a regularized Hotelling's T^2 test that estimates the covariance matrix by $\mathbf{S} + \lambda\mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix and $\lambda > 0$ is a shrinkage parameter. This test works for both $p < n$ and $p \geq n$ cases. Note that a similar method was also proposed in Shen, Lin and Zhu (2011), where the form of $\lambda\mathbf{S} + (1 - \lambda)\mathbf{I}_p$ is used to estimate the covariance matrix with $0 \leq \lambda < 1$. In the special case of $\lambda = 0$, the test reduces to an unscaled Hotelling's test. We refer to the methods in this category as *the regularized Hotelling's tests*.

3) In the third category, the covariance matrix is assumed to be diagonal. Under this assumption, the singularity problem is circumvented since a diagonal matrix is always invertible for non-zero entries, whether or not p is larger than n. This idea was first considered by Wu, Genton and Stefanski (2006) and then revisited by several other researchers; see, for example, Srivastava and Du (2008), Srivastava (2009), Park and Nag Ayyala (2013), and Srivastava, Katayama and Kano (2013). For more details, see Section 2.1 below. These methods are essentially all the same and we refer to them as *the diagonal Hotelling's tests*.

In our simulation studies, we note that the unscaled Hotelling's tests are often sensitive to the deviation of equal eigenvalues of the covariance matrix. If one eigenvalue is extremely large, then the performance of the test will be dominated by that individual component and thus a lower power will result. For more details, see the simulation studies in Section 4. In addition, even for the case of equal eigenvalues, Chen and Qin (2010) suggested $n = [20\log(p)]$ to have a reasonably large power. For instance, n needs to be at least 46, 92 and 138 for $p = 10$, 100 and 1000, respectively. For high-dimensional data such as gene expression microarray data, however, it is not uncommon that n is very small, say for example less than 10 samples per group (Pomeroy et al. 2002, Dong et al. 2005). This has motivated researchers to consider more realistic testing methods for high-dimensional small sample size data, e.g., the regularized Hotelling's tests and the diagonal Hotelling's tests. Our additional simulation studies indicate that the existing regularized Hotelling's tests do not perform comparably to the diagonal Hotelling's tests when n is relatively small.

In view of the good performance of the diagonal Hotelling's tests, we also assume that the covariance matrix is diagonal in this paper. Before moving forward, we note that this diagonal covariance matrix assumption has been commonly used for high-dimensional small sample size data, e.g., Dudoit, Fridlyand and Speed (2002), Bickel and Levina (2004) and Tong and Wang (2007). In particular, Bickel and Levina (2004) pointed out that if the estimated correlations are all very noisy, then we are probably better off without estimating them. This, in essence, is the assumption of a diagonal covariance matrix when n is relatively small. In discriminant analysis, Lee et al. (2005) have also observed that discriminant rules with an inverse generalized matrix may not perform as well as diagonal discriminant rules for microarray data. Although very promising, the performance of the diagonal Hotelling's tests themselves can be suboptimal due to the unreliable estimates of the sample variances from the limited number of observations. This suggests that some modifications to the diagonal Hotelling's tests are necessary to further improve their performance. We note that one such attempt

has already been made by Dinu et al. (2007). They proposed a modified diagonal Hotelling's test, called "SAM-GS", by adding a small constant to each gene-specific variance estimate to stabilize the variance estimation, an idea originated in the SAM test of Tusher, Tibshirani and Chu (2001).

In this paper, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. The test is structured by replacing the sample variances in the diagonal Hotelling's tests by the optimal shrinkage estimation of variances in Tong and Wang (2007). For the proposed shrinkage-based test, we then consider several different ways to derive the approximate null distribution under different scenarios of p and n. Simulation results show that the proposed method always performs comparably to existing competitors, especially when n is less than 10. In addition, to assess the performance of the proposed method using real data, we consider four gene expression data sets. A case study also demonstrates the advantage of the proposed shrinkage-based diagonal Hotelling's test. The remainder of the paper is organized as follows. The shrinkage-based diagonal Hotelling's tests are introduced in Section 2. In Section 3, we derive both a scaled chi-squared null distribution and a normal null distribution. Simulation studies and real data analysis are conducted in Sections 4 and 5, respectively.

2 Improving the Diagonal Hotelling's Tests

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$, $i = 1, \ldots, n$, be independent and identically distributed (i.i.d.) random vectors from a multivariate normal distribution, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the population mean vector and $\boldsymbol{\Sigma}$ is the population covariance matrix. Let also $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ be the sample mean vector and $\mathbf{S} = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T/(n-1)$ be the sample covariance matrix. For the one-sample testing problem, the hypothesis is

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{1}$$

where $\boldsymbol{\mu}_0$ is a fixed vector. To test hypothesis (1), the one-sample Hotelling's T^2 statistic is defined as

$$T_1^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0). \tag{2}$$

When $p \leq n-1$ so that $\mathbf{S}$ is invertible, under H_0, the scaled test statistic, $\{(n-p)/p(n-1)\}T_1^2$, follows an $F_{p,n-p}$ distribution with p and $n-p$ degrees of freedom.

For the two-sample testing problem, similarly, we assume that $\mathbf{X}_{ki} = (X_{ki1}, \ldots, X_{kip})^T$, $i = 1, \ldots, n_k$, are i.i.d. from a multivariate normal distribution, $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, for $k = 1$ and 2, respectively, where $\boldsymbol{\mu}_k$ are the population mean vectors and $\boldsymbol{\Sigma}$ is the common covariance matrix. Let $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ denote the class-specific sample means, and let $\mathbf{S}_{pool} = \{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2\}/(n_1+n_2-2)$ be the pooled sample covariance matrix. Then, to test the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{3}$$

the two-sample Hotelling's T^2 statistic is given by

$$T_2^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \mathbf{S}_{pool}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2). \tag{4}$$

When $p \leq n_1 + n_2 - 2$ so that $\mathbf{S}_{pool}$ is invertible, under H_0, the scaled test statistic, $\{(n_1+n_2-p-1)/p(n_1+n_2-2)\}T_2^2$, follows an F_{p,n_1+n_2-p-1} distribution with p and n_1+n_2-p-1 degrees of freedom.

2.1 The Diagonal Hotelling's Tests

We note that Hotelling's T^2 statistics require $n-1 \geq p$ for the one-sample case and $n_1+n_2-2 \geq p$ for the two-sample case to ensure that the sample covariance matrix is nonsingular. Hence, these methods do not work under the "large p small n" paradigm. To avoid the singularity problem, Wu, Genton and Stefanski (2006) proposed a pooled component test for the two-sample case, which essentially is a diagonal version of

Hotelling's T^2 statistic (4). Specifically, their proposed test statistic is

$$\begin{aligned} T^2_{\mathrm{D2}} &= \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^T\{\mathrm{diag}(\mathbf{S}_{pool})\}^{-1}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2) \\ &= \frac{n_1 n_2}{n_1+n_2}\sum_{j=1}^{p}(\bar{X}_{1j}-\bar{X}_{2j})^2/s^2_{j,pool}, \end{aligned} \tag{5}$$

where $\mathrm{diag}(\mathbf{S}_{pool}) = \mathrm{diag}(s^2_{1,pool},\ldots,s^2_{p,pool})$ with $s^2_{j,pool} = \{(n_1-1)s^2_{1j}+(n_2-1)s^2_{2j})\}/(n_1+n_2-2)$ for $j=1,\ldots,p$. Note that, for simplicity, the missing data problem is not considered in (5) unlike in Wu, Genton and Stefanski (2006). Although only the two-sample case was considered in their paper, the diagonal idea can be readily extended to the one-sample case with T^2_{D1} taking the form

$$\begin{aligned} T^2_{\mathrm{D1}} &= n(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)^T\{\mathrm{diag}(\mathbf{S})\}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0) \\ &= n\sum_{j=1}^{p}(\bar{X}_j-\mu_{0j})^2/s^2_j. \end{aligned}$$

To define the rejection region of the diagonal Hotelling's tests, we reject H_0 if $T^2_{\mathrm{D1}} > C_1$ for the one-sample case and $T^2_{\mathrm{D2}} > C_2$ for the two-sample case, where C_1 and C_2 are two critical values.

Besides the aforementioned pooled components test, Srivastava and Du (2008) proposed scalar transformation invariant tests for both one-sample and two-sample cases, which essentially are the functions of the diagonal Hotelling's test. Srivastava (2009) constructed the test statistic using the diagonal Hotelling's test under non-normality. This test statistic is similar to Srivastava and Du (2008), and the only difference is that Srivastava (2009) deleted the adjustment coefficient appearing in Srivastava and Du (2008). Park and Nag Ayyala (2013) proposed new scalar transformation invariant tests for both one-sample and two-sample cases. Their tests modified the test statistic of Srivastava (2009). This test statistic still assumes that the covariance matrix is diagonal. Srivastava, Katayama and Kano (2013) proposed a two-sample test under the condition of unequal covariance matrices, which essentially is also a function of the diagonal Hotelling's test.

2.2 Shrinkage-Based Diagonal Hotelling's Tests

To establish the diagonal Hotelling's tests, the aforementioned references used the diagonal matrix of sample variances to estimate the covariance matrix. However, when the number of observations is limited, such as when there are fewer than 10 observations, the sample variances are not reliable estimations any more, and the diagonal Hotelling's tests are thus unreliable. This point has been demonstrated by the simulation studies of Srivastava et al. (2013). Therefore, it is necessary to find an improved variance estimation. Dinu et al. (2007) made such an attempt. In this section, we use the optimal shrinkage estimation in Tong and Wang (2007) to improve the variance estimation.

Let $\text{diag}(\mathbf{\Sigma})=\text{diag}(\sigma_1^2,\ldots,\sigma_p^2)$, and $(\sigma_j^2)^t=\sigma_j^{2t}$, $j=1,\ldots,p$. The shrinkage estimator of σ_j^{2t} is

$$\tilde{\sigma}_j^{2t}(\alpha)=\{h_{\nu,p}(t)\hat{\sigma}_{pool}^{2t}\}^{\alpha}\{h_{\nu,1}(t)\hat{\sigma}_j^{2t}\}^{1-\alpha}, \tag{6}$$

where $\hat{\sigma}_j^{2t}$ estimates σ_j^{2t}, $\hat{\sigma}_{pool}^{2t}=\prod_{j=1}^{p}(\hat{\sigma}_j^2)^{t/p}$, $\nu=n-1$, $\Gamma(\cdot)$ is the gamma function, and

$$h_{\nu,p}(t)=\left(\frac{\nu}{2}\right)^t\left\{\frac{\Gamma(\nu/2)}{\Gamma(\nu/2+t/p)}\right\}^p.$$

This shrinkage estimator includes a shrinkage parameter $\alpha\in[0,1]$. The estimator degenerates to the unbiased estimation of σ_j^{2t} if $\alpha=0$, and it shrinks to the pooled variance estimation if $\alpha=1$. Under the Stein loss function, $L(\sigma^2,\tilde{\sigma}^2)=\tilde{\sigma}^2/\sigma^2-\ln(\tilde{\sigma}^2/\sigma^2)-1$, Tong and Wang (2007) proved that there exists a unique optimal α in (0,1], denoted by α^*, to achieve the minimum average risk for any fixed p, ν and $t>-\nu/2$. A two-step procedure, proposed by Tong and Wang (2007), can be useful to estimate α^*.

Now, we move to the shrinkage-based diagonal Hotelling's tests. For ease of distinction, let σ_j^2 and $\sigma_{j,pool}^2$ denote the jth variance in $\text{diag}(\mathbf{\Sigma})$ for one-sample and two-sample cases, respectively. Then, $\hat{\sigma}_j^2=s_j^2$ and $\hat{\sigma}_{j,pool}^2=s_{j,pool}^2$. Moreover, since the sample variances appear in the denominator of the diagonal Hotelling's tests, we consider estimating $\sigma_j^{-2}=1/\sigma_j^2$ instead of σ_j^2, which is the case of $t=-1$. Pang, Tong

and Zhao (2009) found that results for $t = 1$ and $t = -1$ were similar with the latter slightly better. Therefore, we focus on estimating σ_j^{-2} in the remaining of the paper.

Let $\tilde{\alpha}^*$ denote the estimated optimal shrinkage parameter. For the one-sample test, we define the shrinkage-based diagonal Hotelling's test statistic as

$$\begin{aligned} T^2_{\mathrm{SD1}}(\tilde{\alpha}^*) &= n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T \tilde{\mathbf{S}}^* (\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \\ &= n \sum_{j=1}^{p} (\bar{X}_j - \mu_{0j})^2 \tilde{\sigma}_j^{-2}(\tilde{\alpha}^*), \end{aligned} \tag{7}$$

where $\tilde{\mathbf{S}}^* = \mathrm{diag}\{\tilde{\sigma}_1^{-2}(\tilde{\alpha}^*), \ldots, \tilde{\sigma}_p^{-2}(\tilde{\alpha}^*)\}$. Similarly, for the two-sample test, the shrinkage-based diagonal Hotelling's test statistic is

$$\begin{aligned} T^2_{\mathrm{SD2}}(\tilde{\alpha}^*) &= \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \tilde{\mathbf{S}}^*_{pool} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \\ &= \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{p} (\bar{X}_{1j} - \bar{X}_{2j})^2 \tilde{\sigma}^{-2}_{j,pool}(\tilde{\alpha}^*), \end{aligned} \tag{8}$$

where $\tilde{\mathbf{S}}^*_{pool} = \mathrm{diag}\{\tilde{\sigma}^{-2}_{1,pool}(\tilde{\alpha}^*), \ldots, \tilde{\sigma}^{-2}_{p,pool}(\tilde{\alpha}^*)\}$.

Tong and Wang (2007) showed that $\tilde{\alpha}^* \to 0$ for $n \to \infty$ and fixed p. This property demonstrates that, when the sample size is very large, on the one hand, our methods degenerate to the diagonal Hotelling's tests and thus it is unnecessary to borrow information from other genes. Our simulation studies indicate that our methods perform comparably to current approaches, and the diagonal Hotelling's tests are hence appropriate for testing the significance of gene sets. On the other hand, the approximate null distributions in the diagonal Hotelling's tests can also be used. However, when the sample size is small, the above approximations may not be accurate. Hence, it is of great importance to derive the approximate null distribution in the case of small sample sizes.

3 Null Distributions of Shrinkage-Based Diagonal Hotelling's Tests for Small Sample Size

When the sample size is large, two types of distributions are derived to be the approximate null distributions. One is the chi-squared distribution. Wu, Genton and Stefanski (2006) considered a scaled chi-squared distribution as an approximation for both $p < n$ and $p \geq n$. If $(n, p) \to \infty$, the other possible choice, the normal distribution, is used as the asymptotic null distribution (Srivastava and Du 2008, Srivastava 2009, Park and Nag Ayyala 2013, Srivastava, Katayama and Kano 2013). This motivates us to derive approximate null distributions similarly when the sample size is very small. In this section, we follow Wu, Genton and Stefanski (2006) and derive the scaled chi-squared null distribution, and the normal null distribution is derived as $p \to \infty$.

To obtain the null distributions of the shrinkage-based diagonal Hotelling's tests, we first derive the means and variances for $T^2_{\mathrm{SD1}}(\alpha)$ and $T^2_{\mathrm{SD2}}(\alpha)$ in Lemmas 1 and 2, respectively.

Lemma 1. *For any $\nu = n - 1 > 4$ and $\alpha \in (0, 1]$, the mean and variance of the test statistic, T^2_{SD1}, under H_0 are*

$$E\left\{T^2_{\mathrm{SD1}}(\alpha)\right\} = C_1 \sigma_{pool}^{-2\alpha} \sum_{j=1}^{p} \sigma_j^{2\alpha},$$

and

$$Var\left\{T^2_{\mathrm{SD1}}(\alpha)\right\} = (3C_2 - C_3)\sigma_{pool}^{-4\alpha} \sum_{j=1}^{p} \sigma_j^{4\alpha} + (C_3 - C_1^2)\sigma_{pool}^{-4\alpha} (\sum_{j=1}^{p} \sigma_j^{2\alpha})^2,$$

where

$$C_1 = \frac{h^{\alpha}_{\nu,p}(-1) h^{1-\alpha}_{\nu,1}(-1)}{h^{p-1}_{\nu,1}(-\alpha/p) h_{\nu,1}\{-\alpha/p - (1-\alpha)\}},$$

$$C_2 = \frac{h^{2\alpha}_{\nu,p}(-1) h^{2(1-\alpha)}_{\nu,1}(-1)}{h^{p-1}_{\nu,1}(-2\alpha/p) h_{\nu,1}\{-2\alpha/p - 2(1-\alpha)\}},$$

$$C_3 = \frac{h^{2\alpha}_{\nu,p}(-1) h^{2(1-\alpha)}_{\nu,1}(-1)}{h^{p-2}_{\nu,1}(-2\alpha/p) h^2_{\nu,1}\{-2\alpha/p - (1-\alpha)\}}.$$

Lemma 2. *For any $\nu = n_1 + n_2 - 2 > 4$ and $\alpha \in (0, 1]$, the mean and variance of the test statistic, T^2_{SD2}, under H_0 are*

$$E\left\{T^2_{\text{SD2}}(\alpha)\right\} = C_1 \sigma_{pool}^{-2\alpha} \sum_{j=1}^{p} \sigma_{j,pool}^{2\alpha},$$

and

$$Var\left\{T^2_{\text{SD2}}(\alpha)\right\} = (3C_2 - C_3)\sigma_{pool}^{-4\alpha} \sum_{j=1}^{p} \sigma_{j,pool}^{4\alpha} + (C_3 - C_1^2)\sigma_{pool}^{-4\alpha}(\sum_{j=1}^{p} \sigma_{j,pool}^{2\alpha})^2,$$

where

$$C_1 = \frac{h^{\alpha}_{\nu,p}(-1)h^{1-\alpha}_{\nu,1}(-1)}{h^{p-1}_{\nu,1}(-\alpha/p)h_{\nu,1}\{-\alpha/p - (1-\alpha)\}},$$

$$C_2 = \frac{h^{2\alpha}_{\nu,p}(-1)h^{2(1-\alpha)}_{\nu,1}(-1)}{h^{p-1}_{\nu,1}(-2\alpha/p)h_{\nu,1}\{-2\alpha/p - 2(1-\alpha)\}},$$

$$C_3 = \frac{h^{2\alpha}_{\nu,p}(-1)h^{2(1-\alpha)}_{\nu,1}(-1)}{h^{p-2}_{\nu,1}(-2\alpha/p)h^2_{\nu,1}\{-2\alpha/p - (1-\alpha)\}}.$$

The proof of Lemma 1 is given in Appendix 1. The proof of Lemma 2 is omitted since it is essentially the same as that for Lemma 1. Both lemmas are however necessary for determining the parameters of the approximate null distributions.

3.1 Chi-squared Approximation

For small p, the chi-squared distribution can be used as a good approximate null distribution. In this section, we approximate the null distributions of the proposed test statistics as a scaled chi-squared distribution, $c\chi^2_d$, as in Wu, Genton and Stefanski (2006). To determine the scale parameter, c_1, and the degrees of freedom, d_1, for $T^2_{\text{SD1}}(\tilde{\alpha}^*)$, we equate the mean and variance of $c_1\chi^2_{d_1}$ with the mean and variance of $T^2_{\text{SD1}}(\tilde{\alpha}^*)$. Specifically, we have

$$E\{T^2_{\text{SD1}}(\tilde{\alpha}^*)\} = c_1 d_1 \qquad \text{and} \qquad \text{Var}\{T^2_{\text{SD1}}(\tilde{\alpha}^*)\} = 2c_1^2 d_1.$$

For $T^2_{\text{SD2}}(\tilde{\alpha}^*)$, we use the same approach to determine the corresponding scale parameter, c_2, and the degrees of freedom, d_2. The following theorems describe the approximate null distributions for our test statistics.

Theorem 1. *For any $n > 5$ and optimal shrinkage parameter estimation, $\tilde{\alpha}^*$, under the null hypothesis, we have*

$$T^2_{\mathrm{SD1}}(\tilde{\alpha}^*) \sim c_1 \chi^2_{d_1},$$

where

$$c_1 = \frac{(3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_j^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^{p} \sigma_j^{2\tilde{\alpha}^*})^2}{2C_1 \sigma_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_j^{2\tilde{\alpha}^*}},$$

$$d_1 = \frac{2C_1^2 \sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^{p} \sigma_j^{2\tilde{\alpha}^*})^2}{(3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_j^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^{p} \sigma_j^{2\tilde{\alpha}^*})^2}.$$

Theorem 2. *For any $n_1 + n_2 > 6$ and optimal shrinkage parameter estimation, $\tilde{\alpha}^*$, under the null hypothesis, we have*

$$T^2_{\mathrm{SD2}}(\tilde{\alpha}^*) \sim c_2 \chi^2_{d_2},$$

where

$$c_2 = \frac{(3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_{j,pool}^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^{p} \sigma_{j,pool}^{2\tilde{\alpha}^*})^2}{2C_1 \sigma_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_{j,pool}^{2\tilde{\alpha}^*}},$$

$$d_2 = \frac{2C_1^2 \sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^{p} \sigma_{j,pool}^{2\tilde{\alpha}^*})^2}{(3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_{j,pool}^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^{p} \sigma_{j,pool}^{2\tilde{\alpha}^*})^2}.$$

The proofs of Theorems 1 and 2 are simple and straightforward. They are thus omitted. Note that c_1, d_1, c_2 and d_2 involve some unknown quantities. Take c_1 and d_1 for example. Then, $b_1(\boldsymbol{\sigma}^2) = \sigma_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_j^{2\tilde{\alpha}^*}$ and $b_2(\boldsymbol{\sigma}^2) = \sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_j^{4\tilde{\alpha}^*}$ are the unknown quantities. In practice, we suggest the following rules for estimating $b_1(\boldsymbol{\sigma}^2)$ and $b_2(\boldsymbol{\sigma}^2)$, according to the different scenarios:

(i) For any fixed p but large n, by noting that $\hat{\sigma}_j^2 \xrightarrow{a.s.} \sigma_j^2$ as $n \to \infty$, where $\xrightarrow{a.s.}$ denotes the almost sure convergence, we have the following consistent estimators:

$$\hat{b}_1(\boldsymbol{\sigma}^2) = \hat{\sigma}_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^{p} \hat{\sigma}_j^{2\tilde{\alpha}^*} \qquad \text{and} \qquad \hat{b}_2(\boldsymbol{\sigma}^2) = \hat{\sigma}_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \hat{\sigma}_j^{4\tilde{\alpha}^*};$$

(ii) For any fixed n but large p, by Lemma 2 of Tong and Wang (2007), we estimate

$$\breve{b}_1(\boldsymbol{\sigma}^2) = w(\tilde{\alpha}^*)\hat{\sigma}_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^{p} \hat{\sigma}_j^{2\tilde{\alpha}^*} \qquad \text{and} \qquad \breve{b}_2(\boldsymbol{\sigma}^2) = w(2\tilde{\alpha}^*)\hat{\sigma}_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \hat{\sigma}_j^{4\tilde{\alpha}^*},$$

where $\Psi(t) = \Gamma'(t)/\Gamma(t)$ and $w(\alpha) = (\nu/2)^{-\alpha} h_{\nu,1}(\alpha) \exp\{\alpha\Psi(\nu/2)\}$. More specifically, under some mild conditions, we have $\breve{b}_1(\boldsymbol{\sigma}^2) \xrightarrow{a.s.} b_1(\boldsymbol{\sigma}^2)$ and $\breve{b}_2(\boldsymbol{\sigma}^2) \xrightarrow{a.s.} b_2(\boldsymbol{\sigma}^2)$ as $p \to \infty$;

(iii) Otherwise, we estimate $b_1(\boldsymbol{\sigma}^2)$ and $b_2(\boldsymbol{\sigma}^2)$ by replacing σ_j^2 with the estimated optimal shrinkage estimates $\tilde{\sigma}_j^2(\tilde{\alpha}^*)$. Specifically, we estimate them by

$$\tilde{b}_1(\boldsymbol{\sigma}^2) = \tilde{\sigma}_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^{p} \tilde{\sigma}_j^{2\tilde{\alpha}^*}(\tilde{\alpha}^*) \quad \text{and} \quad \tilde{b}_2(\boldsymbol{\sigma}^2) = \tilde{\sigma}_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \tilde{\sigma}_j^{4\tilde{\alpha}^*}(\tilde{\alpha}^*).$$

3.2 Normal Approximation

For large p, the normal distribution can be a good approximation. The following content of this section can illustrate this point.

Take the one-sample shrinkage-based diagonal Hotelling's test for example. Consider the variances, σ_j^2, as random variables and assume that they are i.i.d. from a common distribution, F, with $E(\sigma_1^4) < \infty$ and $E\{\ln(\sigma_1^2)\} < \infty$. Let $U_j(\alpha) = n(\bar{X}_j - \mu_{0j})^2 \{h_{\nu,1}(-1)\hat{\sigma}_j^{-2}\}^{1-\alpha}$, where $j = 1, \ldots, p$ and $\alpha \in (0, 1]$. Then,

$$T_{\mathrm{SD1}}^2(\alpha) = \{h_{\nu,p}(-1)\hat{\sigma}_{pool}^{-2}\}^{\alpha} \sum_{j=1}^{p} U_j(\alpha). \tag{9}$$

In Appendix 2, we show that

$$\hat{\sigma}_{pool}^{-2} \xrightarrow{a.s.} \exp\left[-E\left\{\ln(\sigma_1^2)\right\} + \ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right] \quad \text{as } p \to \infty. \tag{10}$$

This implies that the first term in (9), $\{h_{\nu,p}(-1)\hat{\sigma}_{pool}^{-2}\}^{\alpha}$, converges to a constant when p is large. In addition, given that σ_j^2 are i.i.d. random variables, under H_0, it is easy to see that $U_j(\alpha)$ are also i.i.d. random variables. Thus, by the central limit theorem, for any $\nu > 4$ and $\alpha \in (0, 1]$, we have

$$\frac{\sum_{j=1}^{p} U_j(\alpha) - pE\{U_1(\alpha)\}}{\sqrt{p\mathrm{Var}\{U_1(\alpha)\}}} \xrightarrow{\mathcal{D}} N(0, 1) \quad \text{as } p \to \infty, \tag{11}$$

where $\xrightarrow{\mathcal{D}}$ denotes the convergence in distribution, $E\{U_1(\alpha)\} = E\left[E\left\{U_1(\alpha)|\sigma_1^2\right\}\right] = h_{\nu,1}^{1-\alpha}(-1)E(\sigma_1^{2\alpha})/h_{\nu,1}\{-(1-\alpha)\}$ and

$$\begin{aligned}
\mathrm{Var}\{U_1(\alpha)\} &= E\left[\mathrm{Var}\left\{U_1(\alpha)|\sigma_1^2\right\}\right] + \mathrm{Var}\left[E\left\{U_1(\alpha)|\sigma_1^2\right\}\right] \\
&= \left[\frac{3h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}\{-2(1-\alpha)\}} - \frac{h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}^2\{-(1-\alpha)\}}\right] E(\sigma_1^{4\alpha}) + \frac{h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}^2\{-(1-\alpha)\}}\mathrm{Var}(\sigma_1^{2\alpha}) \\
&= \frac{3h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}\{-2(1-\alpha)\}}E(\sigma_1^{4\alpha}) - \frac{h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}^2\{-(1-\alpha)\}}\left\{E(\sigma_1^{2\alpha})\right\}^2.
\end{aligned}$$

By (10) and (11), together with Slutsky's Theorem, we can claim that the test statistic $T_{\mathrm{SD1}}^2(\tilde{\alpha}^*)$ is approximately normally distributed when p is large. The same conclusion can also be obtained for $T_{\mathrm{SD2}}^2(\tilde{\alpha}^*)$.

Now as in Section 3.1, to have the normal approximation, we equate the mean and variance of $N(\xi_1, \tau_1)$ with the mean and variance of $T_{\mathrm{SD1}}^2(\tilde{\alpha}^*)$. Similarly, for the two-sample comparison, we use the same method to determine the mean, ξ_2, and the variance, τ_2, of $T_{\mathrm{SD2}}^2(\tilde{\alpha}^*)$. The results are summarized as the following theorems.

Theorem 3. *For any $n > 5$ and optimal shrinkage parameter estimation $\tilde{\alpha}^*$, under the null hypothesis, we have*

$$T_{\mathrm{SD1}}^2(\tilde{\alpha}^*) \sim N(\xi_1, \tau_1), \quad \text{as } p \to \infty,$$

where

$$\xi_1 = C_1\sigma_{pool}^{-2\tilde{\alpha}^*}\sum_{j=1}^{p}\sigma_j^{2\tilde{\alpha}^*},$$

$$\tau_1 = (3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*}\sum_{j=1}^{p}\sigma_j^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*}(\sum_{j=1}^{p}\sigma_j^{2\tilde{\alpha}^*})^2.$$

Theorem 4. *For any $n_1 + n_2 > 6$ and optimal shrinkage parameter estimation $\tilde{\alpha}^*$, under the null hypothesis, we have*

$$T_{\mathrm{SD2}}^2(\tilde{\alpha}^*) \sim N(\xi_2, \tau_2), \quad \text{as } p \to \infty,$$

where

$$\xi_2 = C_1 \sigma_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_{j,pool}^{2\tilde{\alpha}^*},$$

$$\tau_2 = (3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \sigma_{j,pool}^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^{p} \sigma_{j,pool}^{2\tilde{\alpha}^*})^2.$$

The practical rules for estimating the unknown quantities are the same as those in Section 3.1.

4 Monte Carlo Simulation Studies

In this section, we compare the shrinkage-based diagonal Hotelling's tests, including the chi-squared null distribution ($SDchi$) and the normal null distribution ($SDnor$), with some current methods in the aforementioned three categories:

- One unscaled Hotelling's test: Chen and Qin (2010) (CQ).
- One regularized Hotelling's test: Chen et al. (2011) (RHT).
- Two diagonal Hotelling's tests: Wu, Genton and Stefanski (2006) (PCT) and Srivastava, Katayama and Kano (2013) (SR).

For RHT, we use the function "RHT.2samp" in the R package "RHT" provided by Chen et al. (2011). In our simulation studies, we simulate both the type I error rate and the power to assess the performances of all approaches. Moreover, we compare all methods by plotting the receiver operating characteristic (ROC) curves. The ROC curve describes the performance of the true positive rate (TPR) as the false positive rate (FPR) varies. The area under the curve (AUC) values are also provided. We mainly focus on small sample sizes in this section; but moderate to large sample sizes are also considered. Note that most existing methods were proposed for the two-sample case. For ease of comparison, we consider only the two-sample test in the simulation studies.

4.1 Simulation Design

In our simulation, we generate data from the multivariate normal distribution with a common covariance matrix $\mathbf{\Sigma}$. To assess the type I error rate, the data are generated for both groups from $N_p(\mathbf{0}, \mathbf{\Sigma})$. To assess the power, one group of data is generated from $N_p(\mathbf{0}, \mathbf{\Sigma})$ and the other one from $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mu_j = c\sigma_j^2$ for $j = 1, \ldots, p$ with c being the effect size, and σ_j^2 is randomly drawn from the scaled chi-squared distribution $(1/5)\chi_5^2$.

The structure of the common covariance matrix is $\mathbf{\Sigma} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$, where $\mathbf{D} = \text{diag}(\mathbf{\Sigma}) = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\mathbf{R}$ is the correlation matrix. We use the following block-diagonal matrix as the correlation matrix:

$$\mathbf{R} = \begin{pmatrix} \mathbf{\Sigma}_\rho & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{-\rho} & \mathbf{0} & \ddots & \vdots \\ \vdots & \mathbf{0} & \mathbf{\Sigma}_\rho & \mathbf{0} & \vdots \\ \vdots & \ddots & \mathbf{0} & \mathbf{\Sigma}_{-\rho} & \ddots \\ \mathbf{0} & \cdots & \cdots & \ddots & \ddots \end{pmatrix}_{p \times p},$$

where $\mathbf{\Sigma}_\rho$ is a $q \times q$ matrix and $q \leq p$. We consider the following two settings for $\mathbf{\Sigma}_\rho$:

- $\mathbf{\Sigma}_\rho = (\sigma_{ij})_{q \times q}$, where $\sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq q$. Let $\mathbf{\Sigma}_{AR}$ denote this type of common covariance matrix.

- $\mathbf{\Sigma}_\rho = (\sigma_{ij})_{q \times q}$, where $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = \rho$ for $1 \leq i \neq j \leq q$. Let $\mathbf{\Sigma}_{CS}$ denote this type of common covariance matrix.

For $\mathbf{\Sigma}_{AR}$, the correlation matrix is autoregressive of the order-1 structure (Guo, Hastie and Tibshirani 2007, Tong, Chen and Zhao 2012). For $\mathbf{\Sigma}_{CS}$, the correlation matrix takes the compound symmetry structure.

In our simulation, we set n_1=n_2=n from 5 to 10, and the effect sizes are c=0.55, 0.52, 0.50, 0.47, 0.43, 0.39 and 0.35. For different correlations, we set ρ=0, 0.2 and 0.4. Note that $\mathbf{\Sigma}_{CS}$ is not positive if both ± 0.4 are included in the block-diagonal correlation matrix. For the case of $\rho = 0.4$, we set all correlations in $\mathbf{\Sigma}_{CS}$ to be positive. The type

I error rate and power are obtained by running 1000 simulations under the settings of $p = 50$, $q = 5$ and $\alpha = 0.05$, where α is the significance level.

4.2 Simulation Results

We first focus on the performances of all approaches for small sample sizes. The type I error rate and power are reported in Tables 1 and 2, respectively. Two different structures of the correlation are considered, and we find that the results are similar for both $\mathbf{\Sigma}_{AR}$ and $\mathbf{\Sigma}_{CS}$. From the results in Tables 1 and 2, we observe that the shrinkage-based diagonal Hotelling's tests outperform the other methods for different ρ. When the correlation is weak, our methods control the type I error rate well and at the same time maintain high power. As the correlation increases, the type I error rate of our methods becomes higher but it is still better than those of the other methods. PCT and SR have high type I error rates when the sample size is smaller than 10. CQ selects the null hypothesis too often; RHT tends to be conservative; and both of their powers are low. For higher correlations, these four approaches perform similarly.

The superiority of the shrinkage-based diagonal Hotelling's tests is also demonstrated in Figure 1 and Table 3. Figure 1 shows the plots of the ROC curves and their respective AUC values are shown in Table 3. As in Si and Liu (2013), we plot ROC curves with a range of FPR values from 0 to 0.1. AUC values are also calculated in the same range as the FPR values. Without loss of generality, we plot the ROC curves for n=6 in Figure 1 to assess the overall performances of all approaches for small sample sizes. The figure shows that our methods, $SDchi$ and $SDnor$, have the largest AUC values and highest ROC curves for all three correlations. Additionally, we observe a very interesting and important result from Figure 1 and Table 3. The six curves appearing in Figure 1 can be divided into three groups, and these three groups clearly represent the aforementioned three categories. We can see that diagonal Hotelling's tests, including our shrinkage-based methods, have the highest ROC curves. The unscaled Hotelling's test has the second highest ROC curves and the regularized Hotelling's test has the lowest ROC curves. This demonstrates that with

Table 1: Type I error rates for p=50 under the null case.

ρ	$\boldsymbol{\Sigma}$	n	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	$\boldsymbol{\Sigma}_{AR} = \boldsymbol{\Sigma}_{CS}$	5	0.049	0.060	0.228	0.356	0.000	0.042
		6	0.044	0.059	0.132	0.250	0.000	0.039
		7	0.045	0.056	0.108	0.178	0.000	0.038
		8	0.052	0.056	0.072	0.148	0.000	0.040
		9	0.050	0.051	0.056	0.128	0.000	0.032
		10	0.046	0.053	0.047	0.120	0.000	0.023
		50	0.052	0.054	0.045	0.059	0.067	0.020
0.2	$\boldsymbol{\Sigma}_{AR}$	5	0.054	0.086	0.181	0.217	0.000	0.042
		6	0.051	0.075	0.135	0.253	0.000	0.026
		7	0.052	0.072	0.147	0.202	0.000	0.028
		8	0.050	0.076	0.105	0.154	0.000	0.034
		9	0.054	0.070	0.072	0.149	0.000	0.034
		10	0.055	0.068	0.049	0.140	0.000	0.016
		50	0.054	0.068	0.043	0.055	0.065	0.012
	$\boldsymbol{\Sigma}_{CS}$	5	0.055	0.090	0.273	0.338	0.000	0.044
		6	0.052	0.089	0.136	0.257	0.000	0.040
		7	0.052	0.094	0.110	0.204	0.000	0.038
		8	0.051	0.085	0.094	0.160	0.000	0.031
		9	0.056	0.085	0.060	0.156	0.000	0.023
		10	0.054	0.075	0.046	0.125	0.000	0.019
		50	0.059	0.076	0.044	0.064	0.070	0.008
0.4	$\boldsymbol{\Sigma}_{AR}$	5	0.069	0.106	0.280	0.257	0.000	0.063
		6	0.067	0.092	0.136	0.223	0.000	0.053
		7	0.073	0.087	0.110	0.175	0.000	0.033
		8	0.070	0.088	0.102	0.151	0.000	0.032
		9	0.069	0.083	0.070	0.116	0.000	0.020
		10	0.067	0.081	0.064	0.113	0.000	0.019
		50	0.068	0.083	0.046	0.058	0.064	0.010
	$\boldsymbol{\Sigma}_{CS}$	5	0.081	0.119	0.270	0.252	0.000	0.038
		6	0.085	0.119	0.136	0.214	0.000	0.036
		7	0.091	0.116	0.110	0.176	0.000	0.034
		8	0.085	0.114	0.106	0.143	0.000	0.034
		9	0.091	0.105	0.096	0.127	0.000	0.027
		10	0.092	0.106	0.063	0.113	0.000	0.015
		50	0.089	0.103	0.033	0.061	0.056	0.007

Table 2: Powers for p=50 under the alternative case.

ρ	$\boldsymbol{\Sigma}$	n	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	$\boldsymbol{\Sigma}_{AR} = \boldsymbol{\Sigma}_{CS}$	5	0.859	0.904	0.910	0.978	0.000	0.070
		6	0.912	0.927	0.925	0.969	0.000	0.059
		7	0.887	0.953	0.830	0.952	0.000	0.060
		8	0.818	0.950	0.842	0.928	0.000	0.076
		9	0.855	0.873	0.756	0.915	0.002	0.062
		10	0.788	0.810	0.713	0.869	0.003	0.045
		50	0.771	0.798	0.739	0.776	0.804	0.406
0.2	$\boldsymbol{\Sigma}_{AR}$	5	0.866	0.848	0.878	0.884	0.000	0.056
		6	0.823	0.853	0.824	0.941	0.000	0.038
		7	0.839	0.950	0.896	0.936	0.000	0.040
		8	0.842	0.888	0.752	0.914	0.000	0.036
		9	0.867	0.859	0.724	0.902	0.002	0.040
		10	0.789	0.847	0.668	0.823	0.002	0.033
		50	0.761	0.772	0.696	0.734	0.744	0.572
	$\boldsymbol{\Sigma}_{CS}$	5	0.843	0.864	0.963	0.965	0.000	0.061
		6	0.847	0.908	0.950	0.961	0.000	0.050
		7	0.897	0.874	0.796	0.924	0.000	0.053
		8	0.849	0.893	0.736	0.913	0.000	0.046
		9	0.876	0.855	0.800	0.892	0.002	0.042
		10	0.747	0.825	0.750	0.818	0.003	0.033
		50	0.750	0.838	0.733	0.728	0.750	0.690
0.4	$\boldsymbol{\Sigma}_{AR}$	5	0.793	0.868	0.834	0.824	0.000	0.051
		6	0.836	0.902	0.910	0.866	0.000	0.034
		7	0.873	0.940	0.713	0.889	0.000	0.040
		8	0.784	0.880	0.729	0.846	0.000	0.041
		9	0.859	0.861	0.730	0.823	0.003	0.040
		10	0.728	0.782	0.714	0.725	0.003	0.036
		50	0.713	0.697	0.589	0.588	0.571	0.668
	$\boldsymbol{\Sigma}_{CS}$	5	0.792	0.789	0.847	0.884	0.000	0.038
		6	0.882	0.837	0.732	0.841	0.000	0.038
		7	0.875	0.886	0.715	0.861	0.000	0.041
		8	0.770	0.783	0.690	0.782	0.000	0.033
		9	0.778	0.796	0.700	0.710	0.000	0.031
		10	0.726	0.685	0.570	0.694	0.002	0.037
		50	0.664	0.689	0.522	0.636	0.556	0.470

limited numbers of observations, the diagonal Hotelling's tests are the best options.

Table 3: AUC values for n=6 and p=50.

ρ	$\boldsymbol{\Sigma}$	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	$\boldsymbol{\Sigma}_{AR}$	0.0823	0.0822	0.0711	0.0707	0.0351	0.0076
0.2	$\boldsymbol{\Sigma}_{AR}$	0.0772	0.0771	0.0680	0.0656	0.0429	0.0074
0.4	$\boldsymbol{\Sigma}_{AR}$	0.0698	0.0706	0.0558	0.0581	0.0353	0.0070
0.4	$\boldsymbol{\Sigma}_{CS}$	0.0668	0.0683	0.0528	0.0579	0.0350	0.0071

Finally, we keep an eye on the case of the large sample size; for example, $n = 50$. The type I error rates and powers of all approaches are also reported in Tables 1 and 2. We find that the the shrinkage-based diagonal Hotelling's tests perform similarly to when the small sample size is small. However, the other approaches obtain satisfactory results that different greatly from the small sample size case. This demonstrates that for large sample sizes, it is unnecessary to borrow information across all variables.

4.3 Robustness of the Proposed Tests

To investigate the robustness of the shrinkage-based diagonal Hotelling's tests, we also conduct two more simulation studies with non-normal data and with unequal covariance matrices, respectively.

4.3.1 With non-normal data

We generate data from the multivariate t-distribution $t_\upsilon(\boldsymbol{\mu}, \boldsymbol{\Sigma}_s)$, where υ is the degrees of freedom and $\boldsymbol{\Sigma}_s$ is the scale matrix. Note that the covariance matrix $\boldsymbol{\Sigma} = \{\upsilon/(\upsilon - 2)\}\boldsymbol{\Sigma}_s$ for $\upsilon > 2$ and we conduct simulation studies under the common covariance matrix assumption. To assess the type I error rate, the data are generated for both groups from $t_\upsilon(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{s})$. To assess the power, one group of data is generated from $t_\upsilon(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{s})$ and the other one from $t_\upsilon(\boldsymbol{\mu}, \boldsymbol{\Sigma}_s)$. We follow the simulation settings in Section 4.1 and set $\upsilon = 4$. Also for simplicity, we only present the simulation results with the covariance matrix $\boldsymbol{\Sigma}_{AR}$. The simulation results with multivariate t data are shown in

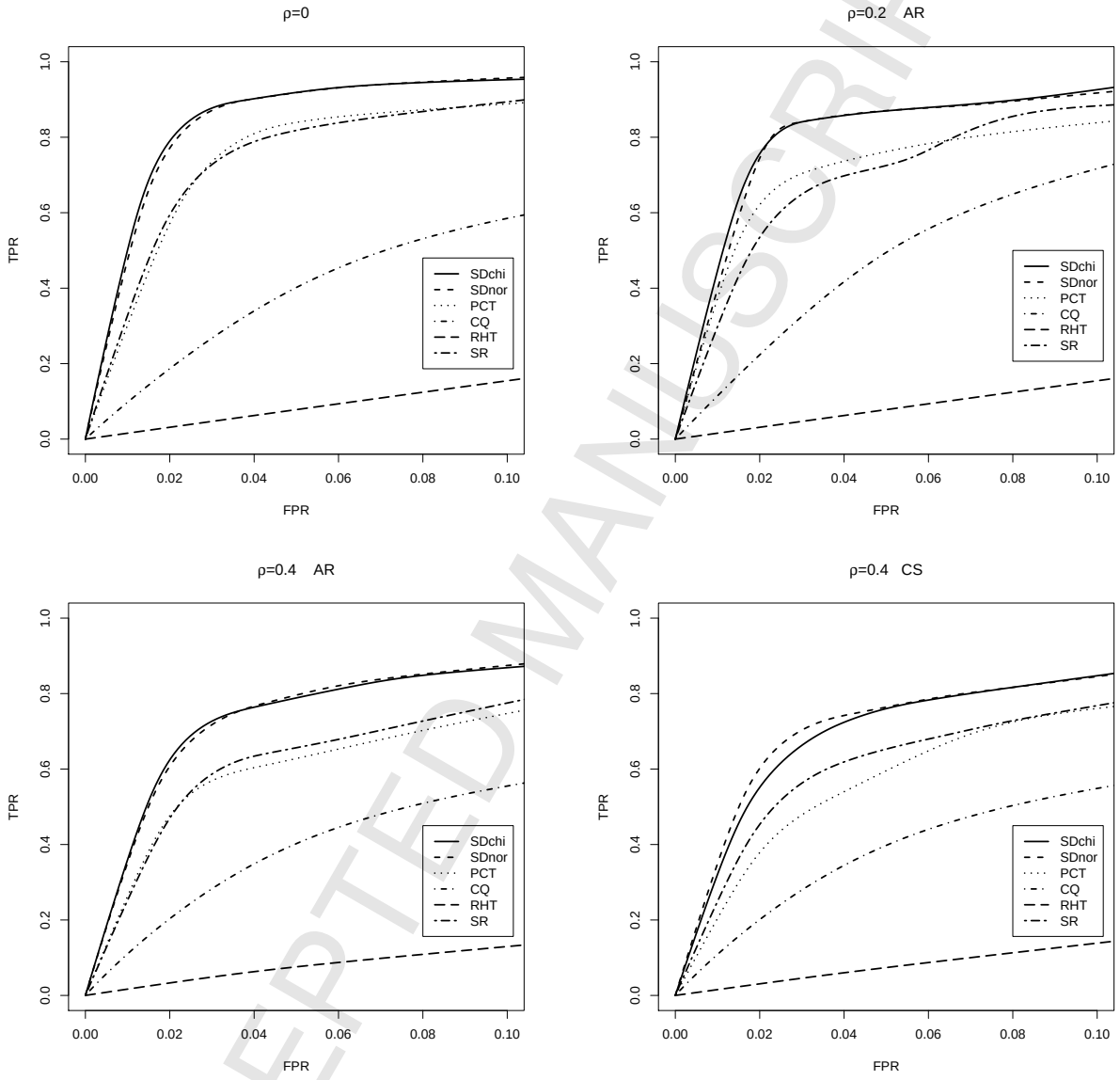


Figure 1: ROC curves for n=6 and p=50. "AR" represents $\mathbf{\Sigma}_{AR}$ and "CS" represents $\mathbf{\Sigma}_{CS}$.

Tables 4 and 5, respectively. We observe that our proposed tests still perform well when the data are generated from the multivariate t distribution. Our methods can control the type I error rate and have higher powers than the other four methods. Overall, our methods are robust in practice.

Table 4: Type I error rates under the null case with multivariate t data.

ρ	n	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	5	0.034	0.060	0.101	0.120	0.000	0.025
	6	0.027	0.048	0.034	0.066	0.000	0.017
	7	0.024	0.049	0.016	0.040	0.002	0.015
	8	0.032	0.042	0.009	0.026	0.001	0.013
	9	0.030	0.042	0.013	0.025	0.003	0.010
	10	0.039	0.055	0.010	0.025	0.001	0.012
0.2	5	0.036	0.058	0.093	0.123	0.000	0.014
	6	0.032	0.055	0.038	0.066	0.002	0.025
	7	0.033	0.056	0.018	0.042	0.001	0.019
	8	0.033	0.045	0.015	0.030	0.001	0.014
	9	0.035	0.042	0.016	0.028	0.001	0.008
	10	0.046	0.061	0.014	0.027	0.001	0.013
0.4	5	0.040	0.073	0.091	0.115	0.002	0.020
	6	0.042	0.068	0.039	0.068	0.000	0.019
	7	0.051	0.066	0.022	0.050	0.001	0.009
	8	0.039	0.057	0.013	0.031	0.002	0.014
	9	0.045	0.061	0.019	0.035	0.001	0.010
	10	0.051	0.067	0.014	0.029	0.000	0.007

4.3.2 With unequal covariance matrices

We generate data from multivariate normal distributions with unequal covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. To make $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, we add the random error ε_j to the component variance σ_{1j}^2 in $\mathbf{\Sigma}_1$. Specifically, we set $\sigma_{2j}^2 = \sigma_{1j}^2 + \varepsilon_j$, where σ_{2j}^2 is the component variance in $\mathbf{\Sigma}_2$, and σ_{1j}^2 is randomly drawn from the scaled chi-squared distribution $(1/5)\chi_5^2$. In our simulations, we follow the simulation settings in Section 4.1 and ε_j follows the uniform distribution U[0,0.2]. Accordingly, we only present the simulation results with the covariance matrix $\mathbf{\Sigma}_{AR}$. The simulation results with unequal covariance matrices

Table 5: Powers under the alternative case with multivariate t data.

ρ	n	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	5	0.848	0.889	0.792	0.799	0.019	0.039
	6	0.890	0.921	0.778	0.806	0.036	0.057
	7	0.926	0.946	0.793	0.816	0.086	0.044
	8	0.913	0.924	0.764	0.803	0.130	0.042
	9	0.910	0.928	0.728	0.770	0.161	0.041
	10	0.890	0.903	0.667	0.719	0.162	0.040
0.2	5	0.843	0.891	0.719	0.793	0.017	0.068
	6	0.895	0.919	0.765	0.803	0.036	0.058
	7	0.927	0.945	0.786	0.819	0.076	0.055
	8	0.907	0.928	0.755	0.800	0.115	0.061
	9	0.914	0.930	0.728	0.764	0.155	0.037
	10	0.888	0.901	0.654	0.705	0.146	0.037
0.4	5	0.839	0.883	0.787	0.789	0.014	0.061
	6	0.869	0.908	0.765	0.791	0.034	0.050
	7	0.915	0.935	0.769	0.806	0.070	0.055
	8	0.897	0.920	0.740	0.781	0.101	0.041
	9	0.895	0.915	0.718	0.744	0.141	0.031
	10	0.874	0.895	0.635	0.678	0.112	0.041

are shown in Tables 6 and 7, respectively. We can see that the simulation results are similar to those under the common covariance matrix assumption. Our methods can control the type I error rates and have a comparable power compared to the other four methods.

Table 6: Type I error rates under the null case with unequal covariance matrices.

ρ	n	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	5	0.041	0.073	0.220	0.327	0.000	0.047
	6	0.051	0.084	0.110	0.269	0.000	0.043
	7	0.046	0.070	0.076	0.197	0.000	0.037
	8	0.045	0.063	0.050	0.165	0.000	0.027
	9	0.043	0.057	0.043	0.136	0.000	0.033
	10	0.041	0.053	0.030	0.125	0.000	0.020
0.2	5	0.047	0.080	0.190	0.301	0.000	0.030
	6	0.052	0.081	0.137	0.272	0.000	0.060
	7	0.054	0.079	0.070	0.196	0.000	0.040
	8	0.050	0.063	0.063	0.167	0.000	0.036
	9	0.051	0.069	0.040	0.153	0.000	0.030
	10	0.047	0.067	0.050	0.115	0.000	0.030
0.4	5	0.062	0.102	0.256	0.297	0.000	0.036
	6	0.059	0.087	0.123	0.203	0.000	0.034
	7	0.065	0.088	0.080	0.180	0.000	0.044
	8	0.066	0.089	0.076	0.161	0.000	0.035
	9	0.066	0.087	0.053	0.131	0.000	0.031
	10	0.067	0.086	0.050	0.123	0.000	0.028

5 Case Studies

A real gene expression data set, whose sample size is smaller than 10, is not uncommon. In addition to the references in Section 1, the following data sets also demonstrate this point:

- Kuster et al. (2011). The data set includes two groups: the sham control group ($n_1 = 8$) and the myocardial infarction group ($n_2 = 8$). The total number of genes is 24,123.

Table 7: Powers under the alternative case with unequal covariance matrices.

ρ	n	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	5	0.827	0.871	0.913	0.959	0.000	0.076
	6	0.920	0.878	0.883	0.962	0.000	0.050
	7	0.930	0.946	0.880	0.983	0.000	0.053
	8	0.949	0.978	0.866	0.987	0.000	0.043
	9	0.910	0.975	0.780	0.962	0.004	0.047
	10	0.937	0.871	0.730	0.971	0.007	0.060
0.2	5	0.846	0.891	0.913	0.990	0.000	0.067
	6	0.875	0.934	0.886	0.964	0.000	0.064
	7	0.872	0.935	0.853	0.966	0.000	0.046
	8	0.911	0.970	0.856	0.971	0.000	0.044
	9	0.877	0.952	0.790	0.940	0.001	0.047
	10	0.857	0.912	0.733	0.955	0.003	0.040
0.4	5	0.833	0.923	0.923	0.955	0.000	0.079
	6	0.871	0.860	0.823	0.951	0.000	0.068
	7	0.912	0.910	0.803	0.952	0.000	0.066
	8	0.850	0.967	0.815	0.943	0.001	0.058
	9	0.849	0.827	0.744	0.924	0.002	0.045
	10	0.914	0.903	0.667	0.871	0.003	0.047

- Bchetnia et al. (2012). The data set includes two groups: the control group ($n_1 = 6$) and the epidermolysis bullosa simplex group ($n_2 = 3$). The total number of genes is 32,321.

- Kaur et al. (2012). The data set includes two groups: the control group ($n_1 = 3$) and the polycystic ovary syndrome group ($n_2 = 7$). The total number of genes is 54,675.

- Mokry et al. (2012). The data set includes two groups: the Ls174T-L8 group ($n_1 = 6$) and the Ls174T-pTER-β-catenin group ($n_2 = 8$). The total number of genes is 54,675.

- Searcy et al. (2012). The data set includes two groups: the control group ($n_1 = 8$) and the Pioglitazone group ($n_2 = 8$). The total number of genes is 45,101.

In this section, we assess the performances of the shrinkage-based diagonall Hotelling's tests when they are applied to real gene expression data sets. The following four gene expression data sets are used in our case studies.

I. *Nakayama data* (Nakayama et al. 2007).

In this data set, there are ten types of soft tissue tumors based on 105 samples and 22,283 probe sets. Without loss of generality, we use the first two types of soft tissue tumors: synovial sarcoma (SS) and myxoid/round cell liposarcoma (MRCL), and the sample sizes are 16 and 19, respectively. All samples are log-transformed as described by Nakayama et al. (2007). This data set has been analyzed by Witten and Tibshirani (2011), who discussed the classification problem, and it can be downloaded from Gene Expression Omnibus (GEO) Datasets using access number GDS2736.

II. *Myeloma data* (Zhan et al. 2007).

This data set includes two therapy groups who have multiple myeloma: Therapy 2 (TH2) with 351 samples and Therapy 3 (TH3) with 208 samples. The total

number of genes is 54,675. This data set has been analyzed by Pang, Tong and Zhao (2009), and it can be downloaded from GEO Datasets using series number GSE2658.

III. *Glioma data* (Sun et al. 2006).

There are four classes of data in this data set: one non-tumor class and three glioma classes. Totally, the data include 54,613 genes and 180 samples. We use the non-tumor (NON) class and the astrocytomas (AS) class, and the sample sizes are 23 and 26, respectively. The data set has also been analyzed by Witten and Tibshirani (2011), and it can be downloaded from GEO Datasets with access number GDS1962.

IV. *Leukemia data* (Golub et al. 1999).

There are two different groups in this data set: the acute lymphoblastic leukemia (ALL) patients group and the acute myeloid leukemia (AML) patients group. The data contain 7,129 genes and 72 samples. We follow the method of Dudoit, Fridlyand and Speed (2002) to threshold, filter, logarithmically (base 10) transform and standardize the data. Finally, we obtain leukemia data with 3,571 genes, 47 ALL patients and 25 AML patients, which are used in our analysis. The data set is available from the package "golubEsets" in Bioconductor.

To plot the ROC curves and calculate the AUC values, we first randomly select p genes from each data set for further analysis. Throughout this section, we consider $p = 50$ and $n = 5$. We then choose one class from each data set to calculate FPR. Specifically, they are SS, TH2, NON and ALL in our analysis. Now we use the first data set (SS and MRCL) for illustration to describe how the FPR and TPR are calculated. For the FPR, we randomly sample two distinct groups (each with size n) from SS and then use them to assess the type I errors. Instead, for the TPR, we sample one group (with size n) from SS and the other group (with size n) from MRCL and use them to assess the power. The FPR and TPR are calculated based on 1000 simulations.

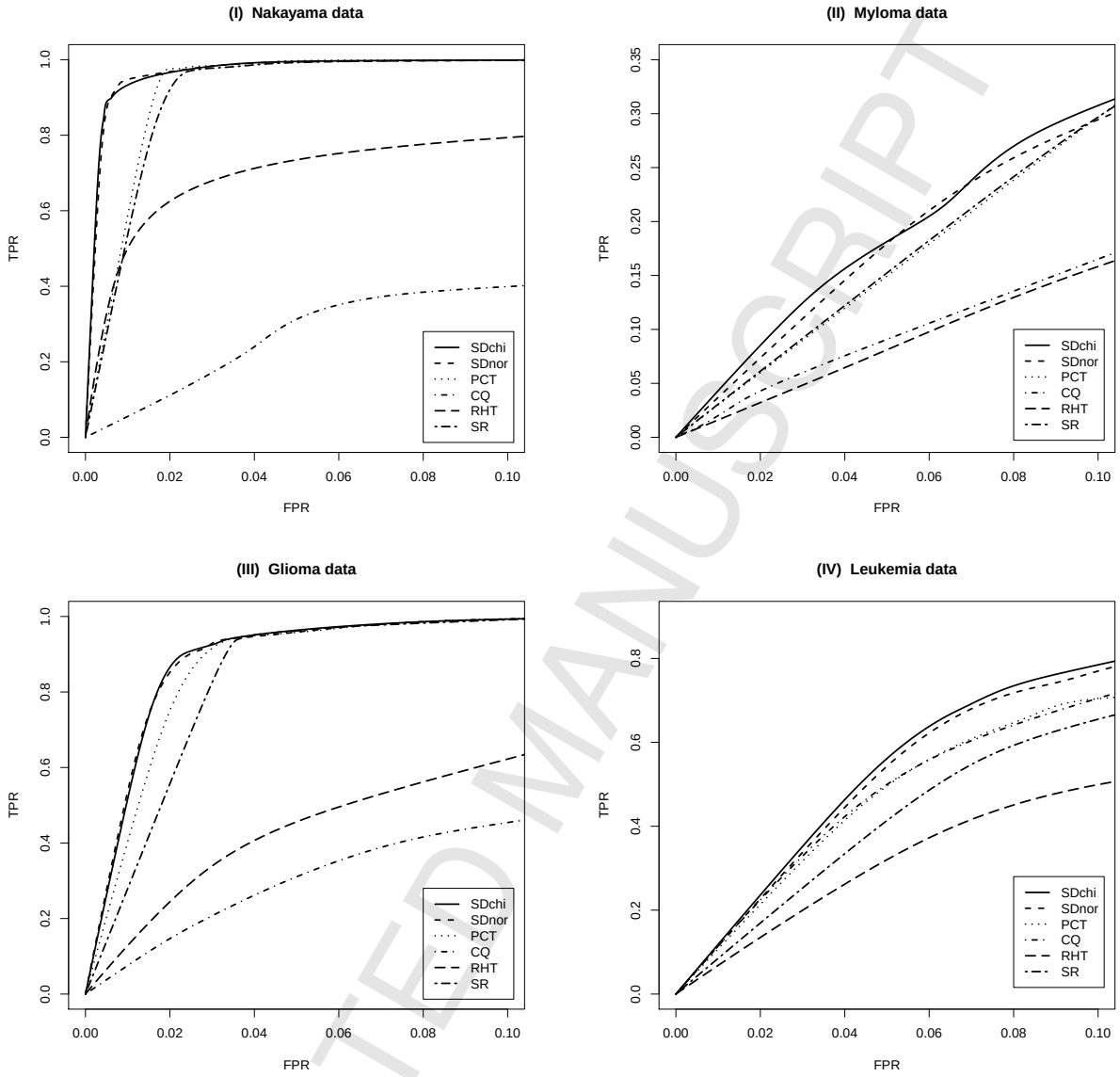


Figure 2: ROC curves for all data sets when p=50 and n=5

Table 8: AUC values for all data sets when p=50 and n=5.

DataSet	*SDchi*	*SDnor*	*PCT*	*SR*	*CQ*	*RHT*
Nakayama	0.0942	0.0941	0.0899	0.0880	0.0255	0.0668
Myeloma	0.0174	0.0165	0.0148	0.0148	0.0088	0.0780
Glioma	0.0868	0.0856	0.0831	0.0781	0.0276	0.0399
Leukemia	0.0484	0.0479	0.0435	0.0382	0.0434	0.0291

Figure 2 shows the ROC curves for all four data sets, and AUC values are provided in Table 8. Similar to Figure 1, the ROC curves in Figure 2 are also generated with a range of FPR values from 0 to 0.1, and the same for AUC values. In Figure 2 and Table 8, our proposed approaches have the highest ROC curves and largest AUC values. This illustrates the advantage of the shrinkage-based diagonal Hotelling's tests. Additionally, the same result as in Section 4 can also be obtained; that is, the diagonal Hotelling's tests perform better than the unscaled Hotelling's tests and the regularized Hotelling's tests when the sample size is small.

6 Discussion

The Hotelling's T^2 test is an important and useful tool for testing multivariate differences in means. However, its requirement that the sample size must be larger than the number of variables is violated in gene expression data analysis. Testing the significance of gene sets would be impossible with the Hotelling's T^2 test due to the difficulty of estimating $\mathbf{\Sigma^{-1}}$. It is therefore necessary to develop new methods to tackle "large p small n" multivariate testing problems. Currently, many statisticians have devoted themselves to solving this problem and some available approaches have been discovered, such as the unscaled Hotelling's tests, the regularized Hotelling's tests and the diagonal Hotelling's tests. However, because of cost or rarity of samples, a small sample size is a very common case. Current available approaches encounter difficulties while testing high-dimensional small sample size data. Our Monte Carlo simulation studies have demonstrated these issues.

In this paper, we proposed a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. For high-dimensional small sample size data, the diagonal Hotelling's tests are better than the unscaled Hotelling's tests and the regularized Hotelling's tests. However, sample variance is an unreliable variance estimator for limited observations. Therefore, we use optimal shrinkage variance estimations to improve the performance of the diagonal Hotelling's test. The improvements are shown

in our simulation studies. Consequently, we suggest using shrinkage-based diagonal Hotelling's tests to test the significance of gene sets with small sample sizes. Furthermore, if the number of genes in the gene sets is not large, the scaled chi-squared null distribution is recommended.

Nevertheless, from our simulation studies, we find that when the correlation becomes high, our methods have higher type I error rates, although higher ROC curves and larger AUC values than those of other methods can be obtained. This phenomenon is likely because the approximate null distributions in this paper are not accurate enough. In addition, real data might not come from a multivariate normal distribution. Some heavy-tailed distributions or even discrete distributions are possible in real data. For example, RNA-seq data, obtained by next-generation sequencing technologies, have better coverage than microarray data have and such data have already been applied in medical science. RNA-seq data from typical high-dimensional small sample size discrete data sets and thus the shrinkage-based diagonal Hotelling's tests, based on multivariate normal distributions, are not suitable for testing the significance of RNA-seq gene sets.

Acknowledgements

Tiejun Tong's research was supported in part by Hong Kong Research grant HKBU202711 and Hong Kong Baptist University FRG grants FRG1/10-11/031, FRG2/13-14/062, and FRG1/14-15/044. The authors thank the editor, the associate editor and the referees for their constructive comments that led to a substantial improvement of the paper.

References

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

Bchetnia, M., Tremblay, M.-L., Leclerc, G., Dupérée, A., Powell, J., McCuaig, C.,

Morin, C., Legendre-Guillemin, V. and Laprise, C. (2012). Expression signature of epidermolysis bullosa simplex, *Human Genetics* **131**: 393–406.

Bickel, P. J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, *Bernoulli* **10**: 989–1010.

Chen, L. S., Paul, D., Prentice, R. L. and Wang, P. (2011). A regularized Hotellings T^2 test for pathway analysis in proteomic studies, *Journal of the American Statistical Association* **106**: 1345–1360.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing, *The Annals of Statistics* **38**: 808–835.

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples, *Biometrics* **16**: 41–50.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A., Jhangri, G., Einecke, G., Famulski, K., Halloran, P., and Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS, *BMC Bioinformatics* **8**: 242.

Dong, S., Nutt, C. L., Betensky, R. A., Stemmer-Rachamimov, A. O., Denko, N. C., Ligon, K. L., Rowitch, D. H. and Louis, D. N. (2005). Histology-based expression profiling yields novel prognostic markers in human glioblastoma, *Journal of Neuropathology and Experimental Neurology* **64**: 948–955.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association* **97**: 77–87.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes, *The Annals of Applied Statistics* **1**: 107–129.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**: 531–537.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays, *Biostatistics* **8**: 86–100.

Hotelling, H. (1931). The generalization of Student's ratio, *Annals of Mathematical Statistics* **2**: 360–378.

Kaur, S., Archer, K. J., Devi, M. G., Kriplani, A., Strauss, J. F. and Singh, R. (2012). Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis, *Journal of Clinical Endocrinology and Metabolism* **97**: E2016–E2021.

Kuster, D. W., Merkus, D., Kremer, A., van IJcken, W. F., de Beer, V. J., Verhoeven, A. J. and Duncker, D. J. (2011). Left ventricular remodeling in swine after myocardial infarction: a transcriptional genomics approach, *Basic Research in Cardiology* **106**: 1269–1281.

Lee, J. W., Lee, J. B., Park, M. and Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics and Data Analysis* **48**: 869–885.

Maciejewski, H. (2013). Gene set analysis methods: statistical models and methodological differences, *Briefings in Bioinformatics, doi: 10.1093/bib/bbt002* .

Mokry, M., Hatzis, P., Schuijers, J., Lansu, N., Ruzius, F.-P., Clevers, H. and Cuppen, E. (2012). Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes, *Nucleic Acids Research* **40**: 148.

Nakayama, R., Nemoto, T., Takahashi, H., Ohta, T., Kawai, A., Seki, K., Yoshida, T., Toyama, Y., Ichikawa, H. and Hasegawa, T. (2007). Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma, *Modern Pathology* **20**: 749–759.

Newton, M. A., Quintana, F. A., Den Boon, J. A., Sengupta, S. and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis, *The Annals of Applied Statistics* **1**: 85–106.

Pang, H., Tong, T. and Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data, *Biometrics* **65**: 1021–1029.

Park, J. and Nag Ayyala, D. (2013). A test for the mean vector in large dimension and small samples, *Journal of Statistical Planning and Inference* **143**: 929–943.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S. and Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature* **415**: 436–442.

Searcy, J. L., Phelps, J. T., Pancani, T., Kadish, I., Popovic, J., Anderson, K. L., Beckett, T. L., Murphy, M. P., Chen, K.-C., Blalock, E. M. et al. (2012). Long-term pioglitazone treatment improves learning and attenuates pathological markers in a mouse model of alzheimer's disease, *Journal of Alzheimer's Disease* **30**: 943–961.

Shen, Y., Lin, Z. and Zhu, J. (2011). Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis, *Computational Statistics and Data Analysis* **55**: 2221–2233.

Si, Y. and Liu, P. (2013). An optimal test with maximum average power while controlling FDR with application to RNA-seq data, *Biometrics* **69**: 594–605.

Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality, *Journal of Multivariate Analysis* **100**: 518–532.

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension, *Journal of Multivariate Analysis* **99**: 386–402.

Srivastava, M. S., Katayama, S. and Kano, Y. (2013). A two sample test in high dimensional data, *Journal of Multivariate Analysis* **114**: 349–358.

Sun, L., Hui, A.-M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R. et al. (2006). Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain, *Cancer Cell* **9**: 287–300.

Tong, T., Chen, L. and Zhao, H. (2012). Improved mean estimation and its application to diagonal discriminant analysis, *Bioinformatics* **28**: 531–537.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis, *Journal of the American Statistical Association* **102**: 113–122.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings National Academic Science* **98**: 5116–5121.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant, *Journal of the Royal Statistical Society: Series B* **73**: 753–772.

Wu, Y., Genton, M. G. and Stefanski, L. A. (2006). A multivariate two-sample mean test for small sample size and missing data, *Biometrics* **62**: 877–885.

Zhan, F., Barlogie, B., Arzoumanian, V., Huang, Y., Williams, D. R., Hollmig, K., Pineda-Roman, M., Tricot, G., van Rhee, F., Zangari, M. et al. (2007). Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis, *Blood* **109**: 1692–1700.

Zhang, J. and Xu, J. (2009). On the k-sample Behrens-Fisher problem for high-dimensional data, *Science in China Series A: Mathematics* **52**: 1285–1304.

Appendix 1: Proof of Lemma 1

For any non-zero $t > -\nu/2$, by Lemma 1 of Tong and Wang (2007), we have

$$E(\hat{\sigma}_j^{2t}) = \sigma_j^{2t}/h_{\nu,1}(t), \qquad j = 1, \ldots, p.$$

This leads to

$$\begin{aligned}
E\{\tilde{\sigma}_j^{-2}(\alpha)\} &= E\left[\{h_{\nu,p}(-1)\hat{\sigma}_{pool}^{-2}\}^{\alpha}\{h_{\nu,1}(-1)\hat{\sigma}_j^{-2}\}^{1-\alpha}\right] \\
&= h_{\nu,p}^{\alpha}(-1)h_{\nu,1}^{1-\alpha}(-1)E\left(\hat{\sigma}_1^{-2\alpha/p}\cdots\hat{\sigma}_j^{-2\alpha/p-2(1-\alpha)}\cdots\hat{\sigma}_p^{-2\alpha/p}\right) \\
&= h_{\nu,p}^{\alpha}(-1)h_{\nu,1}^{1-\alpha}(-1)E(\hat{\sigma}_1^{-2\alpha/p})\cdots E(\hat{\sigma}_j^{-2\alpha/p-2(1-\alpha)})\cdots E(\hat{\sigma}_p^{-2\alpha/p}) \\
&= h_{\nu,p}^{\alpha}(-1)h_{\nu,1}^{1-\alpha}(-1)\frac{\sigma_1^{-2\alpha/p}}{h_{\nu,1}(-\alpha/p)}\cdots\frac{\sigma_j^{-2\alpha/p-2(1-\alpha)}}{h_{\nu,1}\{-\alpha/p-(1-\alpha)\}}\cdots\frac{\sigma_p^{-2\alpha/p}}{h_{\nu,1}(-\alpha/p)} \\
&= C_1\sigma_{pool}^{-2\alpha}\sigma_j^{-2(1-\alpha)}.
\end{aligned}$$

Further, noting that $\bar{X}_j$ and σ_j^2 are independent of each other, we have

$$\begin{aligned}
E\left\{T_{SD1}^2(\alpha)\right\} &= nE\left\{\sum_{j=1}^{p}(\bar{X}_j-\mu_{0j})^2\tilde{\sigma}_j^{-2}(\alpha)\right\} \\
&= n\sum_{j=1}^{p}\frac{\sigma_j^2}{n}C_1\sigma_{pool}^{-2\alpha}\sigma_j^{-2(1-\alpha)} \\
&= C_1\sigma_{pool}^{-2\alpha}\sum_{j=1}^{p}\sigma_j^{2\alpha}.
\end{aligned}$$

To find the variance of $T_{SD1}^2(\alpha)$, it suffices to compute the second moment of $T_{SD1}^2(\alpha)$. For any $j \neq k$, by similar algebra as above, we have

$$E\{\tilde{\sigma}_j^{-2}(\alpha)\tilde{\sigma}_k^{-2}(\alpha)\} = C_3\sigma_{pool}^{-4\alpha}\sigma_j^{-2(1-\alpha)}\sigma_k^{-2(1-\alpha)}.$$

In addition, by the fact that $E(\bar{X}_j-\mu_{oj})^4 = 3\sigma_j^4/n^2$, it gives

$$E(\tilde{\sigma}_j^{-4}(\alpha)) = 3C_2\sigma_{pool}^{-4\alpha}\sigma_j^{-4(1-\alpha)}.$$

Therefore,

$$\begin{aligned} E\left\{T_{SD1}^2(\alpha)\right\}^2 &= n^2 E\left\{\sum_{j=1}^{p}\sum_{k=1}^{p}(\bar{X}_j-\mu_{0j})^2(\bar{X}_k-\mu_{0k})^2\tilde{\sigma}_j^{-2}(\alpha)\tilde{\sigma}_k^{-2}(\alpha)\right\} \\ &= n^2\sum_{j=1}^{p}E(\bar{X}_j-\mu_{0j})^4\tilde{\sigma}_j^{-4}(\alpha) \\ &\quad +n^2\sum_{j\neq k}E(\bar{X}_j-\mu_{0j})^2E(\bar{X}_k-\mu_{0k})^2E\{\tilde{\sigma}_j^{-2}(\alpha)\tilde{\sigma}_k^{-2}(\alpha)\} \\ &= 3C_2\sigma_{pool}^{-4\alpha}\sum_{j=1}^{p}\sigma_j^{4\alpha}+C_3\sigma_{pool}^{-4\alpha}\sum_{j\neq k}\sigma_j^{2\alpha}\sigma_k^{2\alpha}. \end{aligned}$$

Finally, we have

$$\begin{aligned} \mathrm{Var}\{T_{SD1}^2(\alpha)\} &= E\left\{T_{SD1}^2(\alpha)\right\}^2-\left[E\{T_{SD1}^2(\alpha)\}\right]^2 \\ &= (3C_2-C_3)\sigma_{pool}^{-4\alpha}\sum_{j=1}^{p}\sigma_j^{4\alpha}+(C_3-C_1^2)\sigma_{pool}^{-4\alpha}(\sum_{j=1}^{p}\sigma_j^{2\alpha})^2. \quad \square \end{aligned}$$

Appendix 2: Derivation of formula (10)

Note that

$$\ln(\hat{\sigma}_{pool}^{-2})=-\frac{1}{p}\sum_{j=1}^{p}\ln(\sigma_j^2)-\frac{1}{p}\sum_{j=1}^{p}\ln\left(\frac{\nu\hat{\sigma}_j^2}{\sigma_j^2}\right)+\ln(\nu).$$

Given that σ_j^2 are i.i.d. random variables with $E\{\ln(\sigma_1^2)\}<\infty$, by the strong law of large numbers,

$$\frac{1}{p}\sum_{j=1}^{p}\ln(\sigma_j^2)\xrightarrow{a.s.}E\left\{\ln(\sigma_1^2)\right\} \quad \text{as } p\to\infty.$$

In addition, noting that $\nu\hat{\sigma}_j^2/\sigma_j^2$ are i.i.d. chi-squared distributed with ν degrees of freedom, we have

$$\frac{1}{p}\sum_{j=1}^{p}\ln\left(\frac{\nu\hat{\sigma}_j^2}{\sigma_j^2}\right)\xrightarrow{a.s.}E\left\{\ln(\frac{\nu\hat{\sigma}_1^2}{\sigma_1^2})\right\}=\Psi\left(\frac{\nu}{2}\right)+\ln(2) \quad \text{as } p\to\infty.$$

Then, by Slutsky's theorem,

$$\ln(\hat{\sigma}_{pool}^{-2})\xrightarrow{a.s.}-E\left\{\ln(\sigma_1^2)\right\}+\ln\left(\frac{\nu}{2}\right)-\Psi\left(\frac{\nu}{2}\right) \quad \text{as } p\to\infty,$$

which leads to (10). $\square$