<table>
<thead>
<tr>
<th>Statistica Sinica Preprint No: SS-2022-0048</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
</tr>
<tr>
<td><strong>Corresponding Authors</strong></td>
</tr>
<tr>
<td><strong>E-mails</strong></td>
</tr>
</tbody>
</table>
Homogeneity Tests for High-dimensional Mean Vectors and Covariance Matrices

Wenwen Guo\textsuperscript{1}, Xinyuan Song\textsuperscript{2*} and Hengjian Cui\textsuperscript{1}

\textit{Capital Normal University\textsuperscript{1} and The Chinese University of Hong Kong\textsuperscript{2}}

Abstract: This study aims to develop homogeneity tests for high-dimensional mean vectors and covariance matrices, in which the number of features may be greater than the sample size. We introduce two categorically weighted statistics to test the equality of means and of covariance matrices. We establish the asymptotic distributions of the proposed test statistics under certain mild conditions, and develop simplified algorithms to facilitate the implementation and application. Simulation studies demonstrate the satisfactory performance of the proposed tests in terms of the empirical size and power. We also apply the proposed test procedures to two microarray data sets.

Key words and phrases: Homogeneity, K-sample problem, High-dimension, Location and scale, MANOVA.

1. Introduction

Despite numerous studies on homogeneity tests for distributions or distribution features (mean vectors or covariance matrices) in different populations,
a crucial remaining problem is establishing whether gene expression levels differ among predefined patient populations in order to identify a disease's capital causal gene. However, in modern biological and financial studies, the data dimension is often much larger than the sample size. This “large \( p \), small \( n \)” paradigm poses a considerable challenge to classical homogeneity tests, which were originally designed for fixed-dimensional problems.

This study focuses on homogeneity tests for high-dimensional mean vectors and covariance matrices. Assume that homogeneity tests for means, consider \( R \) groups. When \( R = 2 \), the traditional Hotelling \( T^2 \) test is optimal for normally distributed data when \( p \) is fixed. Several extensions of the Hotelling \( T^2 \) test have been proposed to accommodate high dimensionality; examples include those of Bai and Saranadasa (1996), Srivastava and Du (2008), Chen and Qin (2010), Cai et al. (2013), Feng et al. (2016), and Chang et al. (2017). When \( R > 2 \), researchers often use a multivariate analysis of variance (MANOVA) to investigate whether the population mean vectors are the same under the “large \( n \), small \( p \)” paradigm. Cai and Xia (2014) test the equality of multiple high-dimensional sparse mean vectors under dependency. Recently, Hu et al. (2017) proposed a test for the equality of high-dimensional mean vectors based on the work of Chen and Qin (2010).
Several studies also test covariances based mostly on entropy or a quadratic loss function. Studies that examine the case of $R = 2$ include Wolf (2002), Bai et al. (2009), Chen et al. (2010), Li and Chen (2012), Cai and Ma (2013), Jiang and Yang (2013), Cai and Liu (2016), and Chang et al. (2017). For $R > 2$, Zhang et al. (2018) extend the two-sample test for covariances presented by Li and Chen (2012), and obtain the asymptotic distribution of the statistic in a high-dimension case. Zheng et al. (2020) propose a homogeneity test for high-dimensional covariances, and enhance its power by comparing covariance matrices. Liu et al. (2017) also propose a two-sample homogeneity test for means and covariances.

In this study, we consider this kind of homogeneity test from a different perspective. Assume that $Y$ is a categorical variable with $R$ categories, and $X$ is a $p$-dimensional random vector. Cui et al. (2015) propose a mean-variance index defined by $MV(X|Y) = \text{Ex}[\text{var}_Y F(x|Y)]$, where $F(x|Y)$ stands for the conditional distribution function of $X$ given $Y$. $MV(X|Y)$ indicates that $X$ and $Y$ are independent if and only if the conditional distributions $F_r = F(x|Y = r)$, for $r = 1, \cdots, R$, are homogenous. Then, the homogeneity test for distributions can be regarded as an independence test between a categorical variable and a multivariate random vector. The mean-variance index takes advantage of the probabilities of the categorical
variable, which motivates us to introduce a categorically weighted index to measure the differences between the mean vectors and covariance matrices of different groups.

To accommodate high dimensionality, we correct the bias by adjusting the weights, and propose two statistics for testing the mean vectors and covariance matrices. Moreover, we obtain the asymptotic distributions of the proposed statistics under certain mild conditions. The proposed tests have four advantages. (1) They accommodate the high-dimensional setting. (2) No explicit distribution is imposed on the $p$-dimensional vectors; hence, our tests have high theoretical and practical value. (3) The proposed categorically weighted tests optimize the information of the categorical variable to improve the performance. (4) Simplified algorithms are proposed to calculate the associated statistics, thereby facilitating implementation and application.

The remainder of this paper is organized as follows. Sections 2 and 3 describe the methodology and asymptotic distributions of testing means and covariances, respectively. Section 4 introduces simplified algorithms to calculate the test statistics. Section 5 presents Monte Carlo simulations for assessing the performance of the proposed tests. Applications to gene expression data analysis are given in Section 6. Technical proofs are provided
2. Homogeneity Test for Mean Vectors

We consider the homogeneity test for mean vectors, that is,

\[ H_{10} : \mu_1 = \mu_2 = \cdots = \mu_R = \mu, \]  

(2.1)

versus the composite alternative \( H_{11} : \mu_r \neq \mu_s \), for \( 1 \leq r < s \leq R \), where 
\[ \mu_r = E(X|Y = r), \quad \mu = E(X) = \sum_{r=1}^{R} p_r \mu_r, \] 
and \( p_r \) is the probability that \( X \) comes from the \( r \)th population.

2.1 Measuring the difference between mean vectors

Similarly to the analysis of Cui et al. (2015), we use the variance of the conditional means of \( X \) given \( Y \), \( \text{var}_Y \{E(X|Y)\} \), to measure the difference between the mean vectors, as expressed in Definition 1.

**Definition 1.**  The variance of the conditional expectations of \( X \) given \( Y = r \), for \( r = 1, \ldots, R \), can be defined by

\[ U(X|Y) = E(X_1^T X_2) \left\{ \sum_{r=1}^{R} \frac{I(Y_1 = r)I(Y_2 = r)}{p_r} - 1 \right\}, \]

where \((X_1, Y_1)\) and \((X_2, Y_2)\) are independent copies of \((X, Y)\), and \( I(\cdot) \) is the indicator function.
2.1 Measuring the difference between mean vectors

The following lemma shows that Definition 1 is reasonable.

**Lemma 2.1.** If $X$ has a finite first moment, then $\mathcal{U}(X|Y) = \text{var} \{E(X|Y)\} \geq 0$, and the equality holds if and only if the null hypothesis (2.1) is true.

Section S1 of the Supplementary Material shows the proof of Lemma 2.1. For observed random samples $\{(X_k, Y_k) : k = 1, 2, \cdots, n\}$, we define

$$M_{n,p} = \sum_{(i,j)}^{*} X_i^T X_j \left\{ \sum_{r=1}^{R} I(Y_i = r) I(Y_j = r) \frac{\hat{p}_r}{\hat{p}_r - 1} - 1 \right\},$$

where $\sum^{*}_{(i,j)}$ denotes summations over distinct indices, and $\hat{p}_r = (N_r - 1)/(n - 1)$, with $N_r = \sum_{i=1}^{n} I(Y_i = r)$. Notably, $\hat{p}_r$ is a consistent estimator of $p_r$, and more importantly, it enables $\sum_{i \neq j} c_i \left\{ \sum_{r=1}^{R} I(Y_i = r) I(Y_j = r) / \hat{p}_r - 1 \right\} = 0$, where $c_i$ is any function of the $i$th sample. The good properties of the estimator $\hat{p}_r$ make our test applicable to high-dimensional data.

**Remark 1.** Using an elementary calculation, we obtain

$$M_{n,p} = \sum_{r>s}^{R} N_r N_s \left\{ \frac{\sum_{i \neq j} N_r X_{ri}^T X_{rj}}{N_r (N_r - 1)} + \frac{\sum_{i \neq j} N_s X_{si}^T X_{sj}}{N_s (N_s - 1)} - 2 \frac{\sum_{i=1}^{N_r} \sum_{j=1}^{N_s} X_{ri}^T X_{sj}}{N_r N_s} \right\}, \tag{2.2}$$

where $X_{ri}$ denotes the $i$th sample of the $r$th group, that is, $Y_i = r$. We show the proof in Section S2 of the Supplementary Material. When $R = 2$, Equation (2.2) indicates that $M_{n,p}$ is proportional to the statistic proposed by Chen and Qin (2010), which they use to measure the distance between
2.2 Main results for the homogeneity test for means

two sample means, that is, $\|\mathbf{\mu}_1 - \mathbf{\mu}_2\|^2$. Therefore, our proposed statistic can be regarded as a weighted summation of the distances between the means in two different categories.

2.2 Main results for the homogeneity test for means

To establish the limiting distribution of $M_{n,p}$, we assume the following conditions.

**Condition 1.** Suppose that $R$ is fixed, and there exist two positive constants $c_1$ and $c_2$, such that $c_1/R \leq \min_{1 \leq r \leq R} p_r \leq \max_{1 \leq r \leq R} p_r \leq c_2/R$.

**Condition 2.** Suppose that the random expression of $X_i$ given $Y_i = r$ is $X_i|Y_i = r = \mathbf{\mu}_r + \Gamma_r Z_i$, where $\mathbf{\mu}_r$ is the conditional mean vector, $\Gamma_r$ is a $p \times p$ matrix, $Z_i$ is independent of $Y_i$, and the coordinates of $Z_i$ are assumed to be independent and identically distributed (i.i.d.); the first coordinate, denoted as $Z_{i1}$, satisfies $E(Z_{i1}) = 0$, $E(Z_{i1}^2) = 1$ and $E(Z_{i1}^4) = 3 + \Delta < \infty$.

**Condition 3.** $p = p(n) \to \infty$ as $n \to \infty$; $\text{tr}(\Sigma_r \Sigma_s \Sigma_k \Sigma_t) = o\{\text{tr}(\Sigma^2)\}$, for $r, s, k, t \in \{1, 2, \cdots, R\}$.

**Condition 4.** $(\mathbf{\mu}_r - \mathbf{\mu}_s)^t \Sigma_k (\mathbf{\mu}_r - \mathbf{\mu}_s) = o\{n^{-1}\text{tr}(\Sigma^2)\}$, for $r, s, k \in \{1, 2, \cdots, R\}$.

Condition 1 imposes that $p_r$, for $r = 1, 2, \cdots, R$, must not degenerate; a similar condition appears in the study of Cui et al. (2015). Instead of im-
2.2 Main results for the homogeneity test for means

Posing a specific parametric distribution of $X|Y$, the pseudo-independence assumption is required in Condition 2. The pseudo-independence model was proposed by Bai and Saranadasa (1996), and is widely used in high-dimensional theoretical models; see Chen and Qin (2010), Li and Chen (2012), and Zhang et al. (2018). The eigenvalues of the conditional variance of $(X|Y)$ are assumed to satisfy Condition 3, which holds naturally when the conditional covariances are bounded away from above and zero. We explore the asymptotic properties of the statistic $M_{n,p}$ under high dimensionality and local alternatives in Condition 4. This work does not impose any explicit relationships between $p$ and $n$, and our test applies to high-dimensional data.

**Theorem 2.1.** Under Conditions 1, 2, 3, and either $H_{10}$ or Condition 4, we have

$$
M_{n,p} - \sum_{r>s}^{R} N_r N_s \| \mu_r - \mu_s \|^2 \sqrt{\frac{d_{n,p}}{n(n-1)}} \xrightarrow{d} \mathcal{N}(0, 1)
$$

as $n, p \to \infty$, where $d_{n,p} = 2n(n-1)\{\sum_{r=1}^{R}(1-p_r)^2\text{tr}(\Sigma_r^2) + \sum_{(r,s)} p_r p_s \text{tr}(\Sigma_r \Sigma_s)\}$, where $\xrightarrow{d}$ denotes convergence in distribution.

Theorem 2.1 establishes the asymptotic normality of $M_{n,p}$, without imposing explicit conditions on the relationship between $n$ and $p$. Under Condition 3, $d_{n,p} = O(n^2p)$. Furthermore, if the conditional covariances of $(X|Y = r)$ are equal, that is, $\Sigma_1 = \cdots = \Sigma_R = \Sigma$, then...
2.2 Main results for the homogeneity test for means

\[ d_{n,p} = 2n(n-1)(R-1)\text{tr}(\Sigma^2). \]  
Under \( H_{10} \) in (2.1),

\[ \frac{M_{n,p}}{\sqrt{d_{n,p}}} \overset{d}{\to} \mathcal{N}(0, 1). \]  
(2.3)

We use (2.3) to formulate a test procedure based on Theorem 2.1; thus, estimating \( d_{n,p} \) is required. Here, we choose the estimators of \( \text{tr}(\Sigma^2_r) \) and \( \text{tr}(\Sigma_r, \Sigma_s) \) proposed by Li and Chen (2012), and use \( \hat{p}_r = (N_r - 1)/(n - 1) \) to estimate \( p_r \). As \( n \to \infty \), \( \hat{p}_r \) is consistent, by the law of large numbers, and \( \hat{\text{tr}}(\Sigma^2_r) \) and \( \hat{\text{tr}}(\Sigma_r, \Sigma_s) \) are consistent under Conditions 1, 2, and 3 by Theorem 2 in Li and Chen (2012). Additional details about the algorithm for calculating \( \hat{\text{tr}}(\Sigma^2_r) \) and \( \hat{\text{tr}}(\Sigma_r, \Sigma_s) \) are discussed in Section 4.

The proposed test rejects \( H_{10} \) at significance level \( \alpha \) if

\[ M_{n,p} \geq \frac{\hat{d}_{n,p}}{\sqrt{d_{n,p}}} \frac{1}{2} z_\alpha, \]

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

Theorem 2.1 also implies that the proposed test has the asymptotic local power

\[ \Psi_{1,n}^{\text{New}}(\mu_1, \cdots, \mu_R; \alpha) = \Phi\left( -z_\alpha + \frac{n \sum_{r>s} p_r p_s \| \mu_r - \mu_s \|^2}{\sqrt{d_{n,p}/n^2}} \right). \]

When \( \sum_{r>s} p_r p_s \| \mu_r - \mu_s \|^2 \) has a higher order of \( \sqrt{p}/n \), the power converges to one.
3. Homogeneity Test for Covariance Matrices

In this section, we consider the homogeneity test for covariance matrices, that is,

\[ H_{20} : \Sigma_1 = \cdots = \Sigma_R = \Sigma, \] (3.1)

versus the composite alternative \( H_{21} : \Sigma_r \neq \Sigma_s \), for \( 1 \leq r < s \leq R \). Here, \( \Sigma_r = \text{var}(X|Y = r) \) and \( \Sigma = \sum_{r=1}^{R} p_r \Sigma_r \).

3.1 Measuring the difference between covariance matrices

Similarly to the analysis in Section 2, we propose an index to measure the difference between \( \Sigma_r \), for \( r = 1, 2, \cdots, R \). The expression of this index is relatively complex compared with that of \( U(X|Y) \).

**Definition 2.** The distance between the covariances of \( R \) categories is defined by

\[ V(X|Y) = \frac{1}{4} \mathbb{E} \left\{ (X_1 - X_2)^T (X_3 - X_4) \right\}^2 f_{1234}, \]

where \((X_i, Y_i)\), for \( i = 1, \cdots, 4 \), are independent copies of \((X, Y)\), and

\[ f_{1234} = \sum_{r=1}^{R} I(Y_1 = Y_2 = Y_3 = Y_4 = r) \frac{(1 - p_r)}{p_r^3} - \sum_{(r,s)}^* I(Y_1 = Y_2 = r) I(Y_3 = Y_4 = s) \frac{p_r p_s}{p_r p_s}. \]

The following lemma ensures that Definition 2 is reasonable.
3.2 Main results for the homogeneity test for covariance matrices

**Lemma 3.1.** If $X$ has a finite second moment, then $\mathcal{V}(X|Y) \geq 0$, and the equality holds if and only if the null hypothesis (3.1) is true.

Similarly to the analysis for testing means, we define

$$T_{n,p} = \sum_{(i_1,i_2,i_3,i_4)}^* \frac{1}{4} \{(X_{i_1} - X_{i_2})^T(X_{i_3} - X_{i_4})\}^2 \hat{f}_{i_1i_2i_3i_4},$$

where $\sum^*$ denotes summations over distinct indices, and

$$\hat{f}_{i_1i_2i_3i_4} = \sum_{r=1}^R I(Y_{i_1} = Y_{i_2} = Y_{i_3} = Y_{i_4} = r) \frac{(1 - \hat{p}_r)}{\hat{p}_r^3} - \sum_{(r,s)}^* \frac{I(Y_{i_1} = Y_{i_2} = r)I(Y_{i_3} = Y_{i_4} = s)}{\hat{p}_r \hat{p}_s},$$

with $\hat{p}_r = (N_r - 1)/(n - 1)$ and $\hat{p}_r^3 = (N_r - 3)(N_r - 2)(N_r - 1)/(n - 1)^3$.

3.2 Main results for the homogeneity test for covariance matrices

**Theorem 3.1.** Suppose that Conditions 1, 2, and 3 hold. Then, we have

$$T_{n,p} - (n - 1)^2 \sum_{r>s}^R N_rN_s \text{tr}\{(\Sigma_r - \Sigma_s)^2\} \overset{d}{\rightarrow} \mathcal{N}(0, 1)$$

as $p \to \infty$ and $n \to \infty$, where

$$\delta_{n,p} = 4n^6 \left\{ \sum_{r=1}^R (1 - p_r)^2 \text{tr}^2(\Sigma_r^2) + \sum_{(r,s)}^* p_r p_s \text{tr}^2(\Sigma_r \Sigma_s) \right\} + 8n^7 \sum_{r=1}^R p_r \text{tr}\{(\Sigma_r^2 - \Sigma_r \Sigma)^2\} + 4\Delta n^7 \sum_{r=1}^R p_r \text{tr}\{\Gamma_r^T(\Sigma_r - \Sigma)\Gamma_r \circ \Gamma_r^T(\Sigma_r - \Sigma)\Gamma_r}\.$$
Theorem 3.1 establishes the asymptotic normality of $T_{n,p}$. Under $H_{20}$ and Condition 3, \[ \delta_{n,p} = 4n^6(R - 1)^2 \{ \text{tr}(\Sigma^2) \}^2 = O(n^6 p^2). \] We define
\[ \delta_{0,n,p} = 4n^6 \left\{ \sum_{r=1}^{R} (1 - p_r)^2 \text{tr}^2(\Sigma_r^2) + \sum_{r \neq s} p_r p_s \text{tr}^2(\Sigma_r \Sigma_s) \right\}. \]
From Theorem 3.1, we obtain
\[ \frac{T_{n,p}}{\sqrt{\delta_{0,n,p}}} \xrightarrow{d} \mathcal{N}(0, 1), \]
under $H_{20}$. To formulate a test procedure based on Theorem 3.1, we choose
\[ \hat{\delta}_{0,n,p} = 4n^6 \left[ \sum_{r=1}^{R} (1 - \hat{p}_r)^2 \left\{ \text{tr}(\hat{\Sigma}_r^2) \right\}^2 + \sum_{r \neq s} \hat{p}_r \hat{p}_s \left\{ \text{tr}(\hat{\Sigma}_r \hat{\Sigma}_s) \right\}^2 \right]. \]
The proposed test rejects $H_{20}$ at significance level $\alpha$ if \[ T_{n,p} \geq \frac{\delta_{0,n,p}^{1/2} \epsilon \sqrt{\delta_{n,p}}}{1/n^8} z_\alpha. \]
Theorem 3.1 also implies that the proposed test has asymptotic power \[ \Psi_{2,n}^{\text{New}}(\mu_1, \cdots, \mu_R; \alpha) = \Phi \left[ -\sqrt{\frac{\delta_{0,n,p}}{\delta_{n,p}}} z_\alpha + \frac{\sum_{r > s} p_r p_s \text{tr} \{ (\Sigma_r - \Sigma_s)^2 \}}{\sqrt{\delta_{n,p}/n^8}} \right]. \]
When $\sum_{r > s} p_r p_s \text{tr} \{ (\Sigma_r - \Sigma_s)^2 \}$ is of order $p/n$, the power converges to one.

4. Implementation

In this section, we introduce two efficient algorithms for our proposed tests for mean vectors and covariance matrices.
4.1 Testing for mean vectors

When calculating the statistics of two tests, we need to introduce an efficient algorithm to estimate $\text{tr}(\Sigma^2_r)$ and $\text{tr}(\Sigma_r \Sigma_s)$. We use the estimators of $\text{tr}(\Sigma^2_r)$ and $\text{tr}(\Sigma_r \Sigma_s)$ proposed by Li and Chen (2012). That is,

$$\hat{\text{tr}}(\Sigma^2_r) = \frac{1}{N_r(N_r-1)} \sum_{i,j}^* (X^T_{ri}X_{rj})^2 - \frac{2}{N_r(N_r-1)(N_r-2)} \sum_{i,j,k}^* X^T_{ri}X_{rj}X^T_{rk}$$

$$+ \frac{1}{N_r(N_r-1)(N_r-2)(N_r-3)} \sum_{i,j,k,l}^* X^T_{ri}X_{rj}X^T_{rk}X_{sl},$$

$$\hat{\text{tr}}(\Sigma_r \Sigma_s) = \frac{1}{N_rN_s} \sum_i \sum_j (X^T_{ri}X_{sj})^2 - \frac{1}{N_rN_s(N_s-1)} \sum_i \sum_{j,k}^* X^T_{ri}X_{sj}X^T_{s}X_{sk}$$

$$- \frac{1}{N_rN_s(N_s-1)} \sum_i \sum_{j,k,l}^* X^T_{ri}X_{sj}X^T_{rk}X_{sl} + \frac{1}{N_r(N_r-1)N_s(N_s-1)} \sum_{i,j,k,l}^* X^T_{ri}X_{sj}X^T_{rk}X_{sl}.$$ 

Then, we obtain

$$\hat{\text{tr}}(\Sigma^2_r) = \frac{1}{N_r(N_r-3)} \sum_{i \neq j}^* A^r_{ij} A^r_{ij},$$

where $A^r_{ij} = a^r_{ij} - a^r_j/(N_r - 2) - a^r_i/(N_r - 2) + a^r/(N_r - 1)/(N_r - 2)$, with

$$a^r_{ij} = \|X_{ri} - X_{rj}\|^2/2, \quad a^r_i = \sum_{k=1}^{N_r} a^r_{ik}, \text{ and } a^r = \sum_{k=1}^{N_r} \sum_{l=1}^{N_r} a^r_{kl}.$$ 

Similarly,

$$\hat{\text{tr}}(\Sigma_r \Sigma_s) = \frac{1}{(N_r-1)(N_s-1)} \sum_i \sum_{j=1}^{N_s} (X^T_{ri} - \bar{X}_r)^T(X_{sj} - \bar{X}_s)^2,$$
4.2 Testing for covariance matrices

where \( \overline{X}_t = \sum_{i=1}^{N_t} X_{ti}/N_t \), for \( t = 1, 2, \cdots, R \). Because the proofs of Equations (4.1) and (4.2) require complicated calculations, we omit them here. Interested readers can derive them through numerical calculations.

4.2 Testing for covariance matrices

As indicated in Subsection 4.1, \( \hat{\delta}_{0,n,p} \) can be calculated straightforwardly. Hence, we discuss only the calculation of \( T_{n,p} \) in the following. We write

\[
D_r = \frac{1}{4} \sum_{(i,j,k,l)}^* \left\{ (X_{ri} - X_{rj})^T (X_{rk} - X_{rl}) \right\}^2 ,
\]

\[
D_{rs} = \frac{1}{4} \sum_{(i,j)}^* \sum_{(k,l)}^* \left\{ (X_{ri} - X_{rj})^T (X_{sk} - X_{sl}) \right\}^2 .
\]

Then, \( T_{n,p} = \sum_{r=1}^{R} D_r (1 - \hat{p}_r)/\hat{p}_r^3 - \sum_{(r,s)}^* D_{rs}/(\hat{p}_r \hat{p}_s) \). Similarly to the analysis for Equations (4.1) and (4.2), we obtain

\[
D_r = (N_r - 1)(N_r - 2) \sum_{(i,j)}^* A_{ij}^r A_{ij}^r ,
\]

\[
D_{rs} = N_r N_s \sum_{i=1}^{N_r} \sum_{j=1}^{N_s} \left\{ (X_{ri} - \overline{X}_r)^T (X_{sj} - \overline{X}_s) \right\}^2 .
\]

Using the derivations, the two statistics and the associated parameters are expressed in the form of order two. Hence, these statistics are easy to calculate.
5. Simulation Study

We design several simulation experiments to evaluate the performance of the two proposed tests by comparing them with other tests. Here, \( R \) is designed to be three or four, with probabilities \( P_1 = (0.4, 0.4, 0.2) \) or \( P_2 = (0.3, 0.3, 0.2, 0.2) \), respectively. We choose \( n = 100 \) or \( 200 \), and \( p \) ranges from 50 to 400.

Example 1  Test for means

We compare the proposed test for means (NEW.mean) with the distance covariance (dCov) test developed by Székely et al. (2007), the rank of distance test (HHG) proposed by Heller et al. (2013), and the HBWW test suggested by Hu et al. (2017). The distances of \( Y_i \) and \( Y_j \) when applying the dCov and HHG tests are defined as one if they are different, and zero otherwise. We randomly generate a categorical random variable \( Y \) from \( R \) classes. Then, for each given \( Y_i = r \), the \( i \)th predictor \( X_i \) is generated by letting \( X_i = \mu_r + \xi_i \), where \( \xi_i \), for \( i = 1, \ldots, n \), are random errors following \( \mathcal{N}(0, I_p) \) or \( \mathcal{N}(0, \Sigma) \), with \( \Sigma = (0.5|i-j|) \). We set \( \mu_1 = \text{signal} \ast (1, 2, 3, 0, \cdots, 0)^T/\sqrt{14}, \mu_2 = \text{signal} \ast (1, \cdots, 1, 0, \cdots, 0)^T/\sqrt{p/2}, \) and \( \mu_r = 0 \), for \( r \neq 1, 2 \). The tests are repeated 1000 times to simulate the power.
Table 1 shows the empirical sizes of the proposed test (NEW.mean) and the related tests (dCov, HHG, and HBWW). As shown in Table 1, the empirical sizes in all tests maintain the 5% nominal level. Figures 1 and 2 depict the empirical power of the tests when $\varepsilon \sim N(0, I_p)$. As the “signal” increases, the proposed test outperforms the three other tests, and dCov and HBWW tests exhibit similar performance. The HHG test is the least effective in terms of detecting difference between the means of the $R$ groups, implying that considering only the rank of a distance leads to a severe loss of information on distance. For example, when $(n, p) = (100, 200)$, $R = 3$, and signal = 1.4, the empirical power of the proposed test reaches as high as 67.3%. In contrast, the dCov and HBWW tests have power of 56.2% and 51.0%, respectively, and the HHG test has power of only 7.0%. Figure 3 displays the empirical power as $p$ increases. The proposed test consistently outperforms the other tests.

Example 2 Test for covariance matrices

We compare our proposed test for covariances (NEW.cov) with the distance covariance (dCov) test developed by Székely et al. (2004) and Székely et al. (2007), the rank of distance test (HHG) proposed by Heller et al. (2013), the ZBHW test suggested by Zhang et al. (2018), and the ZLGY
Table 1: Empirical sizes of the NEW.mean, dCov, HHG, and HBWW tests for means at a significance level of 5% in Example 1.

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>( R = 3 )</th>
<th>( R = 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NEW.mean</td>
<td>dCov</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>0.043</td>
<td>0.033</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>0.056</td>
<td>0.049</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>0.047</td>
<td>0.043</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.062</td>
<td>0.051</td>
</tr>
<tr>
<td>200</td>
<td>50</td>
<td>0.065</td>
<td>0.041</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>0.058</td>
<td>0.048</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>0.056</td>
<td>0.042</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.058</td>
<td>0.047</td>
</tr>
</tbody>
</table>

Case 1: \( \epsilon \sim \mathcal{N}(\mathbf{0}, I_p) \)

Case 2: \( \epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma) \)
Figure 1: Performance of tests for means with different \((n, p)\) and \(R = 3\).

We randomly generate a categorical random variable \(Y\) from \(R\) classes. Then, for each given \(Y_i = r\), the \(i\)th predictor \(X_i\) is generated by letting \(X_i = \Sigma_i^{1/2} Z_i\), where \(Z_i\), for \(i = 1, \cdots, n\), are random errors following \(N(0, I_p)\). Set \(\Sigma_1 = 3I_p + \text{signal} \ast \eta_1 \eta_1^T\), \(\Sigma_2 = 3I_p + \text{signal} \ast \text{diag}(w_1, \cdots, w_p)\), and \(\Sigma_r = 3I_p\), for \(r \neq 1, 2\), where \(\eta_1 = (3, 3, 3, 0, \cdots, 0)^T\) and \(w_i \sim \text{Unif}(-3, 3)\).
Figure 2: Performance of tests for means with different \((n, p)\) and \(R = 4\).

Table 2 presents the empirical sizes of the tests. As \(n\) and \(p\) approach infinity, the sizes of the five tests are close to the 5\% nominal level. Figures 4 and 5 show the empirical power of the tests. As the “signal” increases, the proposed test outperforms the four other tests. Unlike in the test for means, the HHG test for covariances performs much better than the dCov test, which has power of around 5\%. For example, when \((n, p) = (100, 200)\),
Figure 3: Performance of tests for means with different $p$ values.

$R = 3$, and signal = 0.7, the empirical power of the proposed test reaches as high as 80.1%. In contrast, the ZBHW test has power of around 67.4%, the ZLGY test has power of around 62.8%, and the HHG test has power of only around 36.8%. Figure 6 displays the empirical power as $p$ increases. Again, the proposed test consistently outperforms the other tests.
Table 2: Empirical sizes of the NEW.cov, dCov, HHG, ZBHW, and ZLGY tests for covariances at a significance level of 5% in Example 2.

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>R = 3</th>
<th></th>
<th></th>
<th></th>
<th>R = 4</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>NEW.cov</td>
<td>dCov</td>
<td>HHG</td>
<td>ZBHW</td>
<td>ZLGY</td>
<td>NEW.cov</td>
<td>dCov</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>0.050</td>
<td>0.065</td>
<td>0.062</td>
<td>0.041</td>
<td>0.039</td>
<td>0.058</td>
<td>0.055</td>
<td>0.056</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>0.064</td>
<td>0.054</td>
<td>0.041</td>
<td>0.065</td>
<td>0.047</td>
<td>0.055</td>
<td>0.054</td>
<td>0.047</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>0.051</td>
<td>0.053</td>
<td>0.049</td>
<td>0.039</td>
<td>0.031</td>
<td>0.040</td>
<td>0.047</td>
<td>0.062</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.062</td>
<td>0.038</td>
<td>0.062</td>
<td>0.043</td>
<td>0.033</td>
<td>0.054</td>
<td>0.059</td>
<td>0.053</td>
</tr>
<tr>
<td>200</td>
<td>50</td>
<td>0.060</td>
<td>0.054</td>
<td>0.047</td>
<td>0.046</td>
<td>0.043</td>
<td>0.060</td>
<td>0.068</td>
<td>0.057</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>0.051</td>
<td>0.062</td>
<td>0.043</td>
<td>0.055</td>
<td>0.046</td>
<td>0.061</td>
<td>0.056</td>
<td>0.027</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>0.060</td>
<td>0.049</td>
<td>0.039</td>
<td>0.048</td>
<td>0.039</td>
<td>0.056</td>
<td>0.045</td>
<td>0.054</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.032</td>
<td>0.069</td>
<td>0.035</td>
<td>0.027</td>
<td>0.027</td>
<td>0.051</td>
<td>0.050</td>
<td>0.045</td>
</tr>
</tbody>
</table>

6. Application

6.1 Application 1

We apply the proposed test to a gene expression data set collected by Ko-...
Figure 4: Performance of tests for covariances with different \((n, p)\) and \(R = 3\) at the four stages. Based on their biological functions, the genes were defined using gene ontology (GO), yielding 3,910 GO terms. The data set is obtained from \url{http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5088}. For each GO term, our aim is to test whether the mean vectors and covariance matrices of the gene expression data are the same during the four stages. Table 3 shows the GO terms detected as significant
Figure 5: Performance of tests for covariances with different \((n, p)\) and \(R = 4\). by the NEW.mean, dCov, HHG, HBWW, NEW.cov, ZBHW, and ZLGY tests. The gene set GO:0008499 is detected as significant only by the proposed NEW.mean test, and GO:0070513 and GO:0043008 are detected as significant only by the dCov test. A possible reason for this finding is that the proposed NEW.mean test is designed to detect the difference between mean vectors, whereas the dCov test focuses on identifying the variation
6.1 Application 1

Figure 6: Performance of tests for covariances with different $p$-values.

of distribution functions. The NEW.mean and HBWW tests identify 12 GO terms as significant. Of these, GO:0050786 and GO:0005212 are also detected by the dCov and HHG tests, respectively, and GO:0005179 is identified as significant by the dCov and HHG tests.

In addition, our proposed NEW.cov test and the ZBHW test identify 12 other GO terms as significant gene sets for covariance matrices. However,
the dCov and HHG tests fail to identify any of them. Note that the ZLGY test identifies 219 significant GO terms, of which the trace-based term identifies nine, and the maximum norms detect the rest. This finding implies that using dimension reduction or feature screening methods can further enhance the power for detecting significant gene sets under sparse alternatives. It also motivates a promising extension of our tests to incorporate dimension reduction or feature screening.

Table 3: Significant GO terms obtained by the different tests at a significance level of 5%.

<table>
<thead>
<tr>
<th>GO term</th>
<th>No. of genes</th>
<th>Satisfied test(s)</th>
<th>GO term</th>
<th>No. of genes</th>
<th>Satisfied test(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0004899</td>
<td>36</td>
<td>NEW.mean</td>
<td>GO:0008200</td>
<td>11</td>
<td>NEW.cov/ZBHW/ZLGY</td>
</tr>
<tr>
<td>GO:0070513</td>
<td>17</td>
<td>dCov</td>
<td>GO:0008378</td>
<td>11</td>
<td>NEW.cov/ZBHW/ZLGY</td>
</tr>
<tr>
<td>GO:0043008</td>
<td>10</td>
<td>dCov</td>
<td>GO:0047617</td>
<td>14</td>
<td>NEW.cov/ZBHW</td>
</tr>
<tr>
<td>GO:0008499</td>
<td>15</td>
<td>NEW.mean/HBWW</td>
<td>GO:0015267</td>
<td>12</td>
<td>NEW.cov/ZBHW</td>
</tr>
<tr>
<td>GO:0008083</td>
<td>171</td>
<td>NEW.mean/HBWW</td>
<td>GO:0019870</td>
<td>10</td>
<td>NEW.cov/ZBHW</td>
</tr>
<tr>
<td>GO:0019864</td>
<td>13</td>
<td>NEW.mean/HBWW</td>
<td>GO:000412</td>
<td>15</td>
<td>NEW.cov/ZBHW/ZLGY</td>
</tr>
<tr>
<td>GO:0015254</td>
<td>15</td>
<td>NEW.mean/HBWW</td>
<td>GO:0019870</td>
<td>10</td>
<td>NEW.cov/ZBHW</td>
</tr>
<tr>
<td>GO:0015204</td>
<td>10</td>
<td>NEW.mean/HBWW</td>
<td>GO:0070410</td>
<td>17</td>
<td>NEW.cov/ZBHW</td>
</tr>
<tr>
<td>GO:0015250</td>
<td>16</td>
<td>NEW.mean/HBWW</td>
<td>GO:0016712</td>
<td>10</td>
<td>NEW.cov/ZBHW/ZLGY</td>
</tr>
<tr>
<td>GO:0048037</td>
<td>18</td>
<td>NEW.mean/HBWW</td>
<td>GO:0033038</td>
<td>19</td>
<td>NEW.cov/ZBHW/ZLGY</td>
</tr>
<tr>
<td>GO:0005524</td>
<td>13</td>
<td>NEW.mean/HBWW</td>
<td>GO:0039275</td>
<td>10</td>
<td>NEW.cov/ZBHW/ZLGY</td>
</tr>
<tr>
<td>GO:0016594</td>
<td>14</td>
<td>NEW.mean/HBWW</td>
<td>GO:0030109</td>
<td>16</td>
<td>NEW.cov/ZBHW/ZLGY</td>
</tr>
<tr>
<td>GO:0050786</td>
<td>11</td>
<td>NEW.mean/dCov/HBWW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005212</td>
<td>20</td>
<td>NEW.mean/HHG/HBWW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005179</td>
<td>92</td>
<td>NEW.mean/dCov/HHG/HBWW</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.2 Application 2

Here, we apply the proposed tests to a gene expression data set collected by Taylor et al. (2007) in a study to identify gene sets with significant differences in mean vectors and covariances over time. In this study, 69 patients with the hepatitis C virus were treated for up to 48 weeks using a specific clinical protocol. Their peripheral blood mononuclear cells were collected before treatment (day 0), and on days 1, 2, 7, 14, and 28 during treatment. The original data set is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7123. The microarray gene expression data were measured using 22,283 genes for each patient repeatedly at six stages, defined using GO, based on the biological function of these genes. There are 1,218 GO terms, and a given gene can be a member of multiple GO terms. Further details about the data can be found in Taylor et al. (2007).

Before applying our tests, we preprocess the data by removing 11 individuals with an absent Microarray Suite 5.0 signal transcript, and keep 58 individuals with gene expression arrays at all six stages. We apply the NEW.mean test, dCov test, HHG test, HBWW test, NEW.cov test, ZBHW test, and ZLGY test to the 585 GO terms, with minimums of 10 genes. Let $X_{ri}^{(g)}|Y_i = r$ ($i = 1, 2, \cdots, 58$, $r = 1, 2, \cdots, 6$, $g = 1, 2, \cdots, 585$) be the gene expression data for the $g$th GO term of the $i$th individual at the $r$th period,
where \( r = 1, 2, \cdots, 6 \) represents day 0, 1, 2, 7, 14, and 28, respectively. For each GO term, we test whether the means \( \mu_{r}^{(g)} \) and covariance matrices \( \Sigma_{r}^{(g)} \) are the same across \( r = 1, 2, \cdots, 6 \). Table 4 shows the various numbers of GO terms detected as significant by tests.

In all six stages, the NEW.mean and HBWW tests identify 525 and 524 GO terms, respectively, as significant, where the New.mean test detects GO:0005721, but the HBWW test does not. The dCov and HHG tests simultaneously identify only 459 of the 524 GO terms as significant. For the covariance matrices, the NEW.cov, ZBHW, and ZLGY tests identify 264, 263, and 297 GO terms, respectively, as significant, where the NEW.cov and ZLGY tests detect GO:0000792, but the ZBHW test fails to do so.

Table 4: Number of significant GO terms detected by different tests at a significance level of 5%

<table>
<thead>
<tr>
<th></th>
<th>NEW.mean</th>
<th>HBWW</th>
<th>dCov</th>
<th>HHG</th>
<th>NEW.cov</th>
<th>ZBHW</th>
<th>ZLGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 0, 1, 2, 7, 14, and 28</td>
<td>525</td>
<td>524</td>
<td>543</td>
<td>475</td>
<td>264</td>
<td>263</td>
<td>297</td>
</tr>
<tr>
<td>Day 0 and 1</td>
<td>525</td>
<td>525</td>
<td>535</td>
<td>447</td>
<td>297</td>
<td>296</td>
<td>310</td>
</tr>
<tr>
<td>Day 1 and 2</td>
<td>138</td>
<td>137</td>
<td>149</td>
<td>78</td>
<td>42</td>
<td>42</td>
<td>39</td>
</tr>
<tr>
<td>Day 2 and 7</td>
<td>315</td>
<td>311</td>
<td>395</td>
<td>248</td>
<td>126</td>
<td>126</td>
<td>123</td>
</tr>
<tr>
<td>Day 7 and 14</td>
<td>41</td>
<td>41</td>
<td>48</td>
<td>21</td>
<td>157</td>
<td>157</td>
<td>145</td>
</tr>
<tr>
<td>Day 14 and 28</td>
<td>55</td>
<td>54</td>
<td>40</td>
<td>26</td>
<td>122</td>
<td>122</td>
<td>122</td>
</tr>
</tbody>
</table>

After identifying the significant GO terms, we apply the tests on binary...
segmentation to identify the changes over time. As shown in Table 4, most of the identified changes in the mean vectors and the covariance matrices occurred within days zero and one. However, during the treatment, more GO terms are detected as having significant changes in means between days two and seven. In contrast, more significant changes are identified in the covariance matrices between days 7 and 14. These findings complement the results of Taylor et al. (2007), who observed that the majority of genes altered expression.

7. Conclusion

This study develops two categorically weighted tests for means and covariance matrices in high dimensions. Simulation studies and applications demonstrate the satisfactory performance of our tests. However, the present study has limitations, providing opportunities of future work in this area. While our proposed tests accommodate the high-dimensional setting, they are affected adversely by an increasing dimension, as shown in Figures 3 and 6. Therefore, they cannot deal with ultrahigh-dimensional problems. Moreover, the two tests are less powerful in detecting sparse signals of means and covariance matrices, which may be corrected using dimension reduction or feature screening.
8. Supplementary Material

All technical proofs are provided in the Supplementary Material.

Acknowledgments

Guo’s research was supported by grants from the State Key Program of the National Natural Science Foundation of China (12031016), the National Natural Science Foundation of China (11901406), and the Beijing Natural Science Foundation (Z210003). Song’s research was supported by GRF (14302519, 14302220) from the Research Grant Council of Hong Kong Special Administration Region. Cui’s research was supported by a grant from the National Natural Science Foundation of China (11971324). This work was also supported by the R&D Program of Beijing Municipal Education Commission (KM202110028017), and the Interdisciplinary Construction of Bioinformatics and Statistics.

References


Wenwen Guo

School of Mathematical Sciences, Capital Normal University, Beijing 100048, China. E-mail: wwguo@cnu.edu.cn

Xinyuan Song

Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China. E-mail: xysong@cuhk.edu.hk
REFERENCES

Hengjian Cui

School of Mathematical Sciences, Capital Normal University, Beijing 100048, China. E-mail:

hcui@bnu.edu.cn