Applied Science and Innovative Research
ISSN 2474-4972 (Print) ISSN 2474-4980 (Online)
Vol. 1, No. 2, 2017
www.scholink.org/ojs/index.php/asir

A Two-Sample Test for Mean Vectors in High-Dimensional Data

Knavoot Jiamwattanapong[1*] & Samruam Chongcharoen[1]

[1] School of Applied Statistics, National Institute of Development Administration (NIDA), Bang Kapi, Bangkok, Thailand

[*] Knavoot Jiamwattanapong, E-mail: tuistat10@hotmail.com

Received: October 30, 2017 Accepted: November 6, 2017 Online Published: November 10, 2017
doi:10.22158/asir.v1n2p118 URL: http://dx.doi.org/10.22158/asir.v1n2p118

Abstract

Modern measurement technology has enabled the capture of high-dimensional data by researchers and statisticians and classical statistical inferences, such as the renowned Hotelling's T^2 test, are no longer valid when the dimension of the data equals or exceeds the sample size. Importantly, when correlations among variables in a dataset exist, taking them into account in the analysis method would provide more accurate conclusions. In this article, we consider the hypothesis testing problem for two mean vectors in high-dimensional data with an underlying normality assumption. A new test is proposed based on the idea of keeping more information from the sample covariances. The asymptotic null distribution of the test statistic is derived. The simulation results show that the proposed test performs well comparing with other competing tests and becomes more powerful when the dimension increases for a given sample size. The proposed test is also illustrated with an analysis of DNA microarray data.

Keywords

high-dimensional data, hypothesis testing, two-sample mean vectors, block diagonal structure

1. Introduction

With modern measurement technology, high-dimensional data are nowadays growing in fields as diverse as genetic microarrays, medical imaging, econometrics, geophysics, text or document classification, etc. Such a data creates a wide variety of challenges for quantitative researchers, particularly for statisticians. New strategies are required to analyze and extract useful information from this kind of data. When the data dimension p is larger than the sample size n, many classical statistical methods cannot be applied. For example, the classical Hotelling's T^2 test is not applicable in the case of high-dimensional data (Rencher, 2001; Seber, 2009; Zhang & Xu, 2009) even this test is uniformly most powerful when the dimension is less than the sample size or $p < n$. The reason is that the Hotelling's T^2 statistic requires the sample covariance matrix invertible; but when the data dimension equals or exceeds the sample size, the sample covariance matrix loses its full rank and becomes singular (Eaton & Perlman, 1973).

To make inference on mean vectors is one of the fundamental techniques which can be generalized to other topics in high-dimensional data analysis such as discriminant analysis and regression analysis. Extensive studies have been published for the mean testing problem when both p and n go to infinity with the ratio p / n is bounded. Amongst these works, some have addressed the problem of testing concerning mean vector(s) in high-dimensional data by overcoming the need of the inverse of sample covariance matrix. These developments include Dempster (1958), Bai and Saranadasa (1996), Srivastava and Du (2008), Srivastava (2009), Zhang and Xu (2009), Chen and Qin (2010) and Park and Ayyala (2013). Most of the tests perform well when n large or $p/n \rightarrow (0,1)$, but in practice we often found the datasets with fixed sample size but much larger dimension or $p \rightarrow \infty$ (Park & Ayyala, 2013). The test using for one-sample analysis in high-dimensional data can be found in Jiamwattanapong and Chongcharoen (2015). However, the main attempt of this study is to construct a new test for two-sample problems in such a data. Most importantly, the idea of gaining more information from the sample covariance matrix originally came from Srivastava and Du (2008), and considered the situation that the data are multivariate normal.

Now let $\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i}$ represent a random sample of $p-$dimensional multivariate normal random vectors from the ith group, $i=1,2$, each of which has mean vector $\boldsymbol{\mu}_i$, and unknown common positive definite covariance matrix $\boldsymbol{\Sigma}$, or $\mathbf{x}_{ij} \sim N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$. The sample mean vectors ($\bar{\mathbf{x}}_i$) and the pooled sample covariance matrix ($\mathbf{S}_p$) are used as the estimates of the population counterparts. Let the sample mean vectors, the sample covariance matrices, and the pooled sample covariance matrix be defined, respectively, by

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}, \quad i=1,2 , \tag{1}$$

$$\mathbf{S}_i = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)' \ , \tag{2}$$

and

$$\mathbf{S}_p = \frac{1}{\nu}\sum_{i=1}^{2}(n_i-1)\mathbf{S}_i \ , \tag{3}$$

where $\nu = n_1+n_2-2$.

Consider the hypothesis testing problem:

$$H:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \ \text{ vs. } \ K:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2 \ , \tag{4}$$

and the data come from multivariate normal distributions. When the dimension $p \le \nu,\ \nu = n_1+n_2-2$, the Hotelling's T^2 test is defined as

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \ . \tag{5}$$

The Hotelling's T^2 test is uniformly most powerful and the statistic T^2 can be converted to a central

$F-$distribution as $(\nu-p+1)T^2/\nu p \sim F_{p,\ \nu-p+1}$,$\nu=n_1+n_2-2$ (Davis, 2002; Anderson, 2003). However, it can be easily seen from (5) that the Hotelling's T^2 statistic requires the pooled sample covariance matrix $\mathbf{S}_p$ invertible, so it cannot be applied for high-dimensional data because of the singularity of the pooled sample covariance matrix.

The approach of this paper is based on the idea of keeping the information of the sample covariance matrix as much as possible (Srivastava & Du, 2008; Srivastava, 2009) and also based on the idea of using the submatrices on the diagonal of the sample covariance matrix (Jiamwattanapong & Chongcharoen, 2015). The hypothesis testing problem is in (4) where the data $\mathbf{x}_{ij}$, $i=1,2, j=1,...,n_i$, are independent $p-$dimensional multivariate normal random vectors with mean vectors $\boldsymbol{\mu}_i$, $i=1,2$, and unknown common positive definite covariance matrix $\boldsymbol{\Sigma}$. The situation we consider here, say high dimension problem, is that $p>\nu$, $\nu=n_1+n_2-2$. The organization of this paper is as follows. Description of the new test statistic and its asymptotic null distribution are presented in the next section. The influential works are discussed as the competing tests followed by a report on the performance of the proposed test via simulation study. Then the proposed test is demonstrated by using a dataset of DNA microarray. The last section goes to the conclusion of the study.

2. Test Statistic and Its Asymptotic Distribution

In this section, we describe the proposed test for the testing problem (4) in the case that the dimension of the data exceeds the rank of the pooled sample covariance matrix, or $p>\nu$ where $\nu=n_1+n_2-2$. The unknown common positive definite covariance matrix $\boldsymbol{\Sigma}$ can be written is blocks as

$$\boldsymbol{\Sigma}_{p\times p}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1m}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{\Sigma}_{m1} & \boldsymbol{\Sigma}_{m2} & \cdots & \boldsymbol{\Sigma}_{mm}\end{pmatrix},$$

where $\boldsymbol{\Sigma}_{kk}$, $j=1,...,m$ are $q_k\times q_k$ blocks or submatrices on the diagonal of $\boldsymbol{\Sigma}$ where $q_k\le\nu-6$ and $\sum_{k=1}^{m} q_k = p$. The population correlation matrix R is defined as

$$\mathfrak{R}=\mathbf{D}_\sigma^{-1/2}\boldsymbol{\Sigma}\mathbf{D}_\sigma^{-1/2}=(\mathfrak{R}_{kl}),$$

where $\mathbf{D}_\sigma=\mathrm{diag}(\sigma_{11},...,\sigma_{pp})$, $\sigma_{ii}, i=1,...,p$ are the diagonal elements of $\boldsymbol{\Sigma}$ and $\mathfrak{R}_{kk}$, $k=1,...,m$ is a $q_k\times q_k$ submatrix.

In order to obtain the asymptotic null distribution, an assumption on the population correlation matrix is made as follows:

$$\text{As } p\to\infty \text{ and } n<+\infty,\ \mathfrak{R}_{kl}\to\mathbf{0},\ k\ne l,\ k,l=1,...,m. \tag{6}$$

The pooled sample covariance matrix $\mathbf{S}_p$, defined in (3), is partitioned as for $\boldsymbol{\Sigma}$ and define a block diagonal matrix $\mathbf{D}_q$ as $\mathbf{D}_q=\mathrm{diag}(\mathbf{S}_{11},\mathbf{S}_{22},...,\mathbf{S}_{mm})$, where $\mathbf{S}_{kk}, k=1,...,m$ are submatrices obtained from the pooled sample covariance matrix $\mathbf{S}_p$, giving

$$\mathbf{D}_q = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{S}_{mm} \end{pmatrix}. \tag{7}$$

In the case where the first $m-1$ blocks $\mathbf{S}_{11}, \mathbf{S}_{22},..., \mathbf{S}_{(m-1)(m-1)}$ are size of *q*, i.e., $q_1 = q_2 = ... = q_{m-1} = q$ and $p = \sum_{k=1}^{m-1} q_k + q_m$, the block size *q* is called the "***common block***" size of $\mathbf{D}_q$ (Jiamwattanapong & Chongcharoen, 2015). When $q_k < \nu - 3$, $\mathbf{S}_{kk}, k = 1,...,m$ are invertible (Eaton & Perlman, 1973); thus, the inverse of $\mathbf{D}_q$ can be obtained as $\mathbf{D}_q^{-1} = \text{diag}(\mathbf{S}_{11}^{-1},...,\mathbf{S}_{mm}^{-1})$.

Let the statistic T_n be as

$$T_n = \left(\frac{n_1 n_2}{n_1 + n_2}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{D}_q^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{8}$$

where $\bar{\mathbf{x}}_i$, $i = 1, 2$ defined in (1) and $\mathbf{D}_q$ in (7). The following theorem gives the expectation and variance of the statistic T_n.

Theorem 1. *Let* $\mathbf{x}_{ij}$, $i = 1, 2, j = 1,...,n_i$, *be independent* $p-$*dimensional multivariate normal random vectors with unknown mean vectors* $\boldsymbol{\mu}_i$, $i = 1, 2$, *and unknown common positive definite covariance matrix* $\boldsymbol{\Sigma}$. *When* $p > n - 1$, *let the sample mean vectors, the sample covariance matrices, and the pooled sample covariance matrix be defined by (1) to (3) respectively. Then, under assumption (6), the expectation and variance of* T_n *in (8) are, respectively,*

$$\text{(i)} \quad E(T_n) = \sum_{k=1}^{m}\left(\frac{\nu q_k}{\nu - q_k - 1}\right),$$

$$\text{(ii)} \quad Var(T_n) = \sum_{k=1}^{m}\frac{2\nu^2(\nu-1)q_k}{(\nu - q_k - 1)^2(\nu - q_k - 3)}.$$

Proof. Partition the sample mean vectors $\bar{\mathbf{x}}_i$, $i = 1, 2,$ corresponding to the block sizes in $\mathbf{D}_q$, i.e.,

$$\bar{\mathbf{x}}_i = \begin{pmatrix} \bar{\mathbf{x}}_{i1} \\ \bar{\mathbf{x}}_{i2} \\ \vdots \\ \bar{\mathbf{x}}_{im} \end{pmatrix}, \quad i = 1, 2,$$

where $\bar{\mathbf{x}}_{ik}$ is of dimension $q_k \times 1$, $q_k \le \nu - 6$ and $\sum_{k=1}^{m} q_k = p$.

The statistic T_n can be written in terms of Y_k as

$$T_n = \left(\frac{n_1 n_2}{n_1 + n_2}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{D}_q^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \sum_{k=1}^{m} Y_k, \quad Y_k = \left(\frac{n_1 n_2}{n_1 + n_2}\right)(\bar{\mathbf{x}}_{1k} - \bar{\mathbf{x}}_{2k})'\mathbf{S}_{kk}^{-1}(\bar{\mathbf{x}}_{1k} - \bar{\mathbf{x}}_{2k}).$$

As the statistic Y_k has Hotelling's T^2 distribution with q_k and $\nu - q_k + 1$ degrees of freedom, it can also be converted to a statistic of the $F-$distribution with q_k and $\nu - q_k + 1$ degrees of freedom statistic as follows:

$$Y_k = T^2_{q_k,n-q_k+1} = \frac{\nu q_k}{\nu - q_k + 1} F_{q_k,\nu-q_k+1} \ .$$

By applying the first moment of the statistic F with q_k and $\nu - q_k + 1$ degrees of freedom, we have

$$E(T_n) = E\left(\sum_{k=1}^{m} Y_k\right) = \sum_{k=1}^{m} \frac{\nu q_k}{\nu - q_k - 1} \ .$$

By assumption (6), which Y_j and Y_k are independent when $j \neq k,\ j,k = 1,...,m$, the covariances between Y_j and Y_j are zero, so when applying the second central moment of the statistic F, we have the variance of T_n as

$$Var(T_n) = \sum_{k=1}^{m} \frac{2\nu^2(\nu-1)q_k}{(\nu - q_k + 1)^2(\nu - q_k - 3)} \ .$$

The proof is complete.

We propose a test, for the hypothesis problem $H: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $K: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, which is based on the statistic T_n as

$$T_q = \frac{T_n - \sum_{k=1}^{m} \frac{\nu q_k}{\nu - q_k - 1}}{\left[\sum_{k=1}^{m} \frac{2\nu^2(\nu-1)q_k}{(\nu - q_k + 1)^2(\nu - q_k - 3)}\right]^{1/2}} \ . \tag{9}$$

The following theorem gives the asymptotic null distribution of the test statistic T_n.

Theorem 2. *Let T_q be defined in (9). Under assumption (6) and under the null hypothesis* $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Then $T_q \xrightarrow{d} N(0,1)$,

where " $\xrightarrow{d}$ *" denotes the convergence in distribution.*

Proof. Under assumption (6), when n is fixed, the condition $p \to \infty$ is equivalent to $m \to \infty$. Following from Theorem 1, as $q_k \leq \nu - 6$, this satisfies Lyapunov's condition of the Central Limit Theorem for non-identically distributed random variables. Thus, we have

$$T_q = \frac{T_n - E(T_n)}{\sqrt{Var(T_n)}} \xrightarrow{d} N(0,1) \ .$$

This completes the proof.

It should be noted here that the test statistic T_q is invariant under a group of scalar transformation $\mathbf{x} \to \mathbf{Dx}$, where $\mathbf{D} = \text{diag}(d_1,...,d_p)$ and $d_i \neq 0,$ for all $i, i = 1,..., p$. For this case of two-sample problems, ones may have little trouble to decide on how large of block size should be. We can give some guidance when there is no prior information to arrange variables as blocks that ones should keep maximum block size of $q_k = \nu - 6,\ \nu = n_1 + n_2 - 2$.

3. Other Competing Tests

In this section, three important tests for the hypothesis problem (4) when $p > \nu$ where $\nu = n_1 + n_2 - 2$ are discussed: Dempster (1958) non-exact test, T_D, Bai and Saranadasa (1996) Test, T_{BS}, and Srivastava and Du (2008) test, T_{SD}. All of these tests can be considered as they are based on the same form of a statistic *d* as

$$d = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{A}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{10}$$

where $\mathbf{A}$ is a $p \times p$ matrix used to develop such a test. For example, when $\mathbf{A} = \mathbf{I}_p$, then $d = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ which this term is appeared in the initial test, proposed by Dempster (1958). The Dempster test statistic is defined as

$$T_D = \frac{\left(\dfrac{n_1 n_2}{n_1 + n_2}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\mathrm{tr}(\mathbf{S}_p)} \tag{11}$$

where $\bar{\mathbf{x}}_i, i = 1,2$ are defined in (1) and S_P is the pooled sample covariance matrix defined in (3). Under the null hypothesis, the statistic T_D can be approximated by the $F-$ distribution with $[\hat{r}]$ and $[\nu\hat{r}]$ degrees of freedom, where $[a]$ denotes the largest integer less than or equal to a.

Let $\hat{a}_1 = \dfrac{\mathrm{tr}(\mathbf{S}_p)}{p}$ and $\hat{a}_2 = \dfrac{\nu^2}{p(\nu-1)(\nu+2)}\left[\mathrm{tr}(\mathbf{S}_p^2) - \dfrac{(\mathrm{tr}(\mathbf{S}_p))^2}{\nu}\right]$ where $\nu = n_1 + n_2 - 2$, the approximate degrees of freedom of the numerator $\hat{r}$ can be obtained from $\hat{r} = p\hat{a}_1^2 / \hat{a}_2$. It is known that under particular conditions, Dempster's test is the uniformly most powerful test when the population covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ where $\sigma^2 > 0$. Under null hypothesis, Dempster's test is approximate $F-$ distribution whereas the following two tests, Bai and Saranadasa's test, T_{BS}, and Srivastava and Du's test, T_{SD}, have asymptotic normality. The test proposed by Bai and Saranadasa (1996) is given by

$$T_{BS} = \frac{\left(\dfrac{n_1 n_2}{n_1 + n_2}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \mathrm{tr}(\mathbf{S}_p)}{\left[\dfrac{2\nu(\nu+1)}{(\nu-1)(\nu+2)}\left(\mathrm{tr}(\mathbf{S}_p^2) - \dfrac{\mathrm{tr}^2(\mathbf{S}_p)}{\nu}\right)\right]^{1/2}} \tag{12}$$

where $\nu = n_1 + n_2 - 2$. When considering $d = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{A}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, in (10), it can be seen that T_{BS} is rather similar to Dempster's T_D test in such a way that both of them also use $\mathbf{A} = \mathbf{I}_p$ in their statistics but Bai and Saranadasa derived their statistic T_{BS} to a standard normal distribution instead of approximating it by an $F-$ distribution, like Dempster non-exact test. It is also noted here that both Dempster's test and Bai and Saranadasa's test are invariant under orthogonal transformation $\mathbf{x} \to c\mathbf{Q}\mathbf{x}$, $c \neq 0$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$. The other influential test, presented by Srivastava and Du (2008), is based on the statistic T_{SD} as

$$T_{SD} = \frac{\left(\frac{n_1 n_2}{n_1+n_2}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{D}_{SD}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \frac{\nu p}{\nu - 2}}{\left[2\left(\operatorname{tr}(\mathbf{R}^2) - \frac{p^2}{\nu}\right)c_{p,n}\right]^{1/2}}, \qquad (13)$$

where $\nu = n_1 + n_2 - 2$, $c_{p,n} = 1 + \operatorname{tr}(\mathbf{R}^2)/p^{3/2}$, $\mathbf{D}_{SD}^{-1}$ = $\operatorname{diag}(1/s_{11},...,1/s_{pp})$, S_{ii} are diagonal elements of. S_p Like T_{BS}, the test statistic T_{SD} has asymptotic normality under the null hypothesis. It can be seen that the test T_{SD} is developed by applying $\mathbf{A} = \mathbf{D}_{SD}^{-1}$ in (10) to keep the information more from the diagonal elements of the pooled sample covariance matrix. The test T_{SD} performs well when $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p, \sigma^2 > 0$, particularly for dimension and sample size large.

4. Simulation Study

The performance of the proposed test was evaluated through a simulation study and also it was compared with those of the three tests in the literature: T_D (Dempster, 1958), T_{BS} (Bai & Saranadasa, 1996) and T_{SD} (Srivastava & Du, 2008). Attained significance level and the empirical power of the four test statistics were evaluated and four forms of population covariance matrices were studied. The attained significance level and the empirical power are defined first as follows.

4.1 Attained Significance Level (ASL) and Empirical Power

Let $z_{1-\alpha}$ be the $100(1-\alpha)\%$ quantile of the asymptotic null distribution of the test statistic *T*, e.g., *T* = T_q, then $z_{1-\alpha}$ is the $100(1-\alpha)\%$ quantile of the standard normal distribution $N(0,1)$ with *m* iterations of the datasets simulated under the null hypothesis. Also let $F_{[\hat{r}],[\nu\hat{r}],1-\alpha}$ be the $100(1-\alpha)\%$ quantile of the approximate null distribution of Dempster's test statistic T_D. The ASL for the three tests T_{BS}, T_{SD} and T_q is computed as ASL $= \frac{\text{number of } t_H > z_{1-\alpha}}{rep}$ and that for T_D is as ASL $= \frac{\text{number of } t_H > F_{[\hat{r}],[\nu\hat{r}],1-\alpha}}{rep}$, where t_H represents the values of the test statistic *T* based on the datasets generated under the null hypothesis and *rep* is the number of iterations in the simulation. The nominal significance level α = 0.05 was chosen throughout the simulation. As the ASL is approximately distributed as the binomial distribution b(10000, 0,05), so the standard deviation is estimated by $se(\text{ASL}) = \sqrt{0.05(0.95)/10{,}000} \simeq 0.00218$.

The empirical power was obtained by generating datasets under the alternative hypothesis with *rep* =10,000 replications, followed by computing the empirical power as

$$\text{Empirical power } = \frac{\text{number of } t_K > c}{rep},$$

where t_K represents the values of the test statistic *T* based on the datasets generated under the alternative hypothesis and c equals $z_{1-\alpha}$ or $F_{[\hat{r}],[\nu\hat{r}],1-\alpha}$, as defined previously.

4.2 Parameter Selection

The mean vectors for the null hypothesis are $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (0,...,0)'$ and those for the alternative hypothesis are $\boldsymbol{\mu}_1 = (0,...,0)'$ and $\boldsymbol{\mu}_2 = (\nu_1, \nu_2, ..., \nu_p)'$,$\nu_{2k-1} = 0$, $\nu_{2k} \overset{iid}{\sim} U(-1.0, -0.5)$, $k = 1, 2, ..., p/2$. Four forms of population covariance matrix are included in these simulations: 1) $\mathbf{\Sigma}_1 = \mathbf{I}_p$, 2) $\mathbf{\Sigma}_2 = \text{diag}(\mathbf{\Sigma}_{11}, ..., \mathbf{\Sigma}_{mm})$, where $\mathbf{\Sigma}_{kk} = c\mathbf{I} + (1-c)\mathbf{11}'$, $c = 0.8$ and **1** is a vector of 1's, $\mathbf{\Sigma}_{kk}, k = 1, ..., m-1$ are of dimension q, and the last block is q_m, where $p = (m-1)q + q_m$, 3) $\mathbf{\Sigma}_3 = \mathbf{D}_\sigma^{1/2} \mathfrak{R} \mathbf{D}_\sigma^{1/2}$, where $\mathbf{D}_\sigma^{1/2} = \text{diag}(\sigma_1, ..., \sigma_p)$, $\sigma_i = 2 + (-1)^{i-1}(p-i+1)/p$ and $\mathfrak{R} = \text{diag}(\mathfrak{R}_{11}, ..., \mathfrak{R}_{mm})$, where $\mathfrak{R}_{kk} = (\rho_{ij})$, $\rho_{ij} = (-1)^{i+j}(c^{|i-j|})$, $i, j = 1, ..., q_i$ and $c = 0.9$, and 4) $\mathbf{\Sigma}_4 = \mathbf{D}_\sigma^{1/2} \mathfrak{R} \mathbf{D}_\sigma^{1/2}$, where $\mathbf{\Sigma}_4$ are formed as in the third case except that the blocks in R are of five different sizes and these blocks are randomly located on the diagonal.

The simulations were studied at (p, n_i) = (60,20), (100,20), (100,40), (200,20), (200,40), (200,60), (400,20), (400,40), and (400,60). i = 1,2. For each combination of the data dimension (p) and sample size (n_i), the proposed statistic T_q was computed for the chosen common block size (q) of $\nu - 6$,$\nu = n_1 + n_2 - 2$. The first form covariance matrix, which is an identity matrix, was studied under these simulations so that we can compare with the best one, the Dempster's test. Also, the results from the simulations of the important works, Bai and Saranadasa T_{BS} and Srivastava and Du T_{SD} tests were provided. Under each setting, n_i, i = 1,2, multivariate normal vectors with the chosen mean vector and covariance matrix were generated, then the ASL and the empirical power were recorded.

4.3 Simulation Results

The performance of the proposed test was evaluated through simulations with four different forms of covariance matrix. Both the attained significance levels and the empirical powers for each form of covariance matrix are reported in Tables 1 to 4.

The performance of the proposed test statistic T_q when $\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \mathbf{I}_p$ was investigated, as shown in Table 1. The proposed statistic T_q was computed for both cases where the common block sizes in matrix D_q were q = 1 and $q = \nu - 6$, $\nu = n_1 + n_2 - 2$. In this form of covariance matrix, $\mathbf{\Sigma} = \mathbf{I}_p$, all of the tests perform well which, as is known for this case, the Dempster's test is uniformly most powerful test. The ASLs of the proposed test T_q and the T_{BS} test are quite similar. i.e., they are slightly higher than the nominal level 0.05. For some cases of $p < 200$, even the empirical powers of the proposed test T_q were in acceptable level but the powers were slightly lower than the other three tests. In addition, when consider the empirical power, it is obvious that the proposed test T_q performed better when the dimension (p) increased.

Table 1. ASLs and Empirical Powers When $\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \mathbf{I}$ at Nominal Significance Level $\alpha = 0.05$

p	n_i	ASL				Empirical Power			
		T_D	T_{BS}	T_{SD}	T_q	T_D	T_{BS}	T_{SD}	T_q
$q = 1$									
60	20	0.051	0.059	0.053	0.058	1.000	1.000	1.000	1.000
100	20	0.050	0.059	0.050	0.056	1.000	1.000	1.000	1.000
	40	0.049	0.055	0.048	0.055	1.000	1.000	1.000	1.000
200	20	0.050	0.056	0.045	0.057	1.000	1.000	1.000	1.000
	40	0.051	0.056	0.046	0.056	1.000	1.000	1.000	1.000
	60	0.047	0.053	0.045	0.053	1.000	1.000	1.000	1.000
400	20	0.050	0.054	0.038	0.054	1.000	1.000	1.000	1.000
	40	0.053	0.058	0.045	0.056	1.000	1.000	1.000	1.000
	60	0.053	0.056	0.043	0.053	1.000	1.000	1.000	1.000
$q = v - 6$									
60	20	0.051	0.059	0.053	0.051	1.000	1.000	1.000	0.869
100	20	0.050	0.059	0.050	0.053	1.000	1.000	1.000	0.960
	40	0.049	0.055	0.048	0.052	1.000	1.000	1.000	0.958
200	20	0.050	0.056	0.045	0.058	1.000	1.000	1.000	0.997
	40	0.051	0.056	0.046	0.056	1.000	1.000	1.000	1.000
	60	0.047	0.053	0.045	0.046	1.000	1.000	1.000	1.000
400	20	0.050	0.054	0.038	0.056	1.000	1.000	1.000	1.000
	40	0.053	0.058	0.045	0.057	1.000	1.000	1.000	1.000
	60	0.053	0.056	0.043	0.057	1.000	1.000	1.000	1.000

For the covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_2$ and the common block size in $\boldsymbol{D}_q$ chosen corresponded to $\mathbf{\Sigma}$ with the common block size being $q_1 = ... = q_{m-1} = v - 6$, the results are shown in Table 2. In this form of covariance matrix, the proposed statistic T_q performed well and was superior to all other statistics. The ASLs of T_{SD} and T_{BS} were not close to the nominal level 0.05. The ASL of T_{BS} was too high whereas that of T_{SD} was too low. The results when the covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_3$ with the common block size being $q_1 = ... = q_{m-1} = v - 6$, as shown in Table 3, are similar to those in Table 2 even when the elements in the blocks of $\mathbf{\Sigma}_3$ were changed. In other words, varying the entries of the blocks in the covariance matrix but still keeping the same block size, did not have much impact on the proposed statistic T_q; it still performed well. Additionally, when the ASLs of the test statistics T_{SD} and T_{BS} were not close to the nominal value 0.05, their empirical powers, whether they were high or not, were less reliable.

Table 2. ASLs and Empirical Powers When $\Sigma = \Sigma_2(q = v - 6)$ at Nominal Significance Level α = 0.05

p	***n_i***	**ASL**				**Empirical Power**			
		T_D	T_{BS}	T_{SD}	T_q	T_D	T_{BS}	T_{SD}	T_q
60	20	0.050	0.080	0.026	0.051	0.636	0.765	0.465	1.000
100	20	0.046	0.074	0.024	0.053	0.852	0.933	0.716	1.000
	40	0.041	0.070	0.017	0.052	0.983	1.000	0.839	1.000
200	20	0.055	0.076	0.028	0.058	0.997	0.999	0.983	1.000
	40	0.046	0.072	0.015	0.052	1.000	1.000	0.997	1.000
	60	0.048	0.075	0.009	0.046	1.000	1.000	0.999	1.000
400	20	0.054	0.071	0.023	0.057	1.000	1.000	1.000	1.000
	40	0.052	0.074	0.015	0.057	1.000	1.000	1.000	1.000
	60	0.050	0.075	0.011	0.057	1.000	1.000	1.000	1.000

Table 3. ASLs and Empirical Powers When $\Sigma = \Sigma_3(q = v - 6)$ at Nominal Significance Level α = 0.05

p	***n_i***	**ASL**				**Empirical Power**			
		T_D	T_{BS}	T_{SD}	T_q	T_D	T_{BS}	T_{SD}	T_q
60	20	0.051	0.071	0.040	0.051	0.212	0.287	0.468	1.000
100	20	0.051	0.071	0.041	0.053	0.319	0.387	0.702	1.000
	40	0.052	0.071	0.035	0.052	0.747	0.825	0.994	1.000
200	20	0.052	0.067	0.037	0.058	0.589	0.650	0.962	1.000
	40	0.050	0.066	0.036	0.052	0.984	0.991	1.000	1.000
	60	0.050	0.065	0.034	0.046	1.000	1.000	1.000	1.000
400	20	0.058	0.068	0.037	0.057	0.884	0.907	0.999	1.000
	40	0.057	0.067	0.039	0.057	1.000	1.000	1.000	1.000
	60	0.053	0.065	0.037	0.057	1.000	1.000	1.000	1.000

For the last form of covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_4$, which contained blocks of five different sizes (the smallest block size is set to smaller than 10 and the largest size is $v - 6$) on the diagonal, it can be concluded that the proposed statistic T_q outperformed both statistics T_{SD} and T_{BS} used for comparison, as shown in Table 4. Once again, when the ASLs of the test statistics T_{SD} and T_{BS} were not close to the nominal value 0.05, their empirical powers were less reliable than T_q. From the results, the ASLs of T_{SD} are under the nominal value of 0.05 while those of T_{BS} are over; this indicates the unfavorable performance of the two test statistics.

Table 4. ASLs and Empirical Powers When $\Sigma = \Sigma_4$ (with Five Different Block Sizes) at Nominal Significance Level $\alpha = 0.05$

p	n_i	ASL				Empirical Power			
		T_D	T_{BS}	T_{SD}	T_q	T_D	T_{BS}	T_{SD}	T_q
60	20	0.050	0.068	0.041	0.048	0.249	0.315	0.568	1.000
100	20	0.052	0.071	0.041	0.052	0.328	0.399	0.723	1.000
	40	0.052	0.067	0.039	0.049	0.809	0.871	0.999	1.000
200	20	0.052	0.065	0.037	0.052	0.601	0.666	0.965	1.000
	40	0.052	0.066	0.039	0.050	0.988	0.994	1.000	1.000
	60	0.052	0.065	0.035	0.049	1.000	1.000	1.000	1.000
400	20	0.054	0.065	0.036	0.052	0.900	0.922	1.000	1.000
	40	0.056	0.064	0.039	0.053	1.000	1.000	1.000	1.000
	60	0.053	0.065	0.036	0.053	1.000	1.000	1.000	1.000

5. Demonstration of the Proposed Test

An example to demonstrate the new test is an analysis of DNA microarray data from an oncology study. The data, published by Notterman et al. (2001) were retrieved on April 20, 2015 from the Princeton University Gene Expression Project website (http://www.genomics-pubs.princeton.edu/oncology). A selection of 200 genes (p) was used to test the mean vectors of two independent groups, tumor tissue and normal tissue. Each group is the sample size of 10, i.e., n_1 = n_2 = 10, provided that $v = n_1 + n_2 - 2 = 18$.

Before calculating the test statistics for mean vectors, the data were tested for the equality of covariance matrices, using the method presented by Chaipitak and Chongcharoen (2013), and led to the conclusion of equal covariance matrices.

Table 5. Testing the Equality of the Gene Expression Levels between Tumor and Normal Tissue

	T_D	T_{BS}	T_{SD}	T_q
Test Statistic	3.922	7.028	5.574	12.147
$p-$value	< 0.001	< 0.001	< 0.001	< 0.001

To compute the proposed test statistic T_q, the variables in blocks were arranged in such a way that the correlation coefficient of any two adjacent variables in the same block was greater than or equal to 0.5 and the largest block size was 12. The test values T_D, by Dempster (1958) in (11), T_{BS}, by Bai and Saranadasa (1996) in (12), and T_{SD}, by Srivastava and Du (2008) in (13) were also reported. The test results from all of the tests are shown in Table 5; all of them lead to the rejection of the null hypothesis of no difference between the two mean vectors, i.e., the gene expression levels of tumor tissue are significantly different from those of normal tissue at the 0.05 level of significance.

6. Conclusion

In this study, we developed a new test for two-sample problems in high-dimensional data when the data are multivariate normal. The development of the test is based on the idea of keeping more information from the pooled sample covariance matrix. When the data dimension goes to infinity, or $p \to \infty$, the proposed test statistic has been shown to follow a standard normal distribution under the null hypothesis. One of the advantages of the new test and is that it is invariant under a group of scalar transformation. Simulation results show that the proposed test performs well and becomes more powerful when the dimension increases for a given sample size.

Acknowledgment

The authors would like to express their gratitude to the Ministry of Science and Technology, Thailand, for their financial support.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed., p. 176). Wiley, New Jersey, U.S.A.

Bai, Z., & Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, 6, 311-329.

Chaipitak, S., & Chongcharoen, S. (2013). A test for testing the equality of two covariance matrices for high-dimensional data. *Journal of Applied Sciences*, *13*(2), 270-277. https://doi.org/10.3923/jas.2013.270.277

Chen, S. X., & Qin, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, *38*(2), 808-835. https://doi.org/10.1214/09-AOS716

Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements* (1st ed., p. 47). Springer New York, U.S.A.

Dempster, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, *29*(4), 995-1010. https://doi.org/10.1214/aoms/1177706437

Eaton, M. L., & Perlman, M. D. (1973). The Non-Singularity of Generalized Sample Covariance Matrices. *The Annals of Statistics*, *1*(4), 710-717. https://doi.org/10.1214/aos/1176342465

Jiamwattanapong, K., & Chongcharoen, S. (2015). A new test for the mean vector in high-dimensional data. *Songklanakarin J. Sci. Technol.*, *37*(4), 477-484.

Notterman, D. A., Alon, U., Sierk, A. J., & Levine, A. J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research*, *61*, 3124-3130.

Park, J., & Ayyala, D. N. (2013). A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning and Inference*, *143*(5), 929-943. https://doi.org/10.1016/j.jspi.2012.11.001

Rencher, A. C. (2001). *Methods of Multivariate Analysis* (2nd ed., pp. 118-120) Wiley, New York, U.S.A.

Seber, G. A. (2009). *Multivariate observations* (Vol. 252, p. 28). John Wiley & Sons, New York, U.S.A.

Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis*, *100*(3), 518-532. https://doi.org/10.1016/j.jmva.2008.06.006

Srivastava, M. S., & Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, *99*(3), 386-402. https://doi.org/10.1016/j.jmva.2006.11.002

Tracy, C. A., & Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, *177*, 727-754. https://doi.org/10.1007/BF02099545

Zhang, J., & Xu, J. (2009). On the k-sample Behrens-Fisher problem for high-dimensional data. *Science in China, Series A: Mathematics*, *52*(6), 1285-1304. https://doi.org/10.1007/s11425-009-0091-x