Songklanakarin J. Sci. Technol.
37 (4), 477-484, Jul. - Aug. 2015

http://www.sjst.psu.ac.th

Original Article

A new test for the mean vector in high-dimensional data

Knavoot Jiamwattanapong* and Samruam Chongcharoen

School of Applied Statistics, National Institute of Development Administration (NIDA), Bang Kapi, Bangkok, 10240 Thailand.

Received: 25 February 2015; Accepted: 6 May 2015

Abstract

For the testing of the mean vector where the data are drawn from a multivariate normal population, the renowned Hotelling's T^2 test is no longer valid when the dimension of the data equals or exceeds the sample size. In this study, we consider the problem of testing the hypothesis $H:\boldsymbol{\mu}=\mathbf{0}$ and propose a new test based on the idea of keeping more information from the sample covariance matrix. The development of the statistic is based on Hotelling's T^2 distribution and the new test has invariance property under a group of scalar transformation. The asymptotic distribution is derived under the null hypothesis. The simulation results show that the proposed test performs well and is more powerful when the data dimension increases for a given sample size. An analysis of DNA microarray data with the new test is demonstrated.

Keywords: high-dimensional data, hypothesis testing, mean vector, block diagonal structure

1. Introduction

The rapid increase in the occurrence of high-dimensional data has been found in many areas of interest thus making it essential to acquire new statistical methods to deal with them. Examples of data can be found in genetic microarrays, bioinformatics, economics, engineering, industry and meteorology, amongst other sources. It is often not possible to address these datasets by classical statistical methods due to their high dimensionality and complexity. When making inference on this type of data, testing the mean vector for one sample is a fundamental technique which is not only beneficial for its main purpose, but also provides an important step when developing other statistical techniques, e.g. regression analysis. Among the test statistics on the mean vector used in multivariate analysis, the most well-known one is Hotelling's T^2, $n\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}$, where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean vector and sample covariance matrix respectively, based on a sample of size *n*. The test statistic T^2 takes into account the correlation in the data, which is the concept conforming to the Mahalanobis distance (Mahalanobis, 1936) $D_M(\mathbf{x})=\sqrt{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$, so it is not surprising that many tests on the mean vector are also based on this distance. The advantages of the statistic T^2 are: it is powerful when the dimension of data is sufficiently small, it is invariant under linear transformation, and particularly, its exact distribution is known under the null hypothesis. However, one serious limitation of Hotelling's T^2 is that the sample covariance matrix needs to be invertible. Theoretically, when $p>n-1$, where p is the data dimension and n is the sample size, the sample covariance matrix loses its full rank and becomes singular. Consequently, the classical test statistic T^2 is no longer valid (Eaton and Perlman, 1973; Rencher, 2001; Zhang and Xu, 2009).

To overcome the problem of the need for the inverse of a sample covariance matrix, extensive work has been carried out by many researchers (see Dempster, 1958; Bai and Saranadasa, 1996; Srivastava and Du, 2008; Srivastava, 2009; Chen and Qin, 2010; Park and Ayyala, 2013). The initial one, proposed by Dempster (1958), henceforth T_D, is given by

* Corresponding author.
Email address: tuistat10@hotmail.com

$$T_D = \frac{n\overline{\mathbf{x}}'\overline{\mathbf{x}}}{\mathrm{tr}(\mathbf{S})}\,, \tag{1}$$

where $\overline{\mathbf{x}}$, the sample mean vector, and $\mathbf{S}$, the sample covariance matrix, are defined, respectively by

$$\overline{\mathbf{x}} = \left(\sum_{i=1}^{n} \mathbf{x}_i\right)/n \tag{2}$$

and $$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = (s_{ij})\,. \tag{3}$$

The Dempster statistic T_D has been shown to have approximately an F-distribution with r and n-r degrees of freedom under the null hypothesis $H:\boldsymbol{\mu} = \mathbf{0}$. One of the advantages of Dempster's test is that, when the dimension is proportionally close to the sample size, this test is more powerful than Hotelling's test even in situations where Hotelling's T^2 is well defined (Bai and Saranadasa, 1996). However, one of the hardships of performing Dempster's test is that the test statistic involves a complicated estimation of the parameter r, the degrees of freedom of the F-distribution, whose explicit form is unknown. The other two fundamental test statistics include one presented by Bai and Saranadasa (1996), T_{BS}, and the other by Srivastava and Du (2008), T_{SD}. Bai and Saranadasa presented their statistic for testing the mean vectors of the two-sample case, but in our study we focus on the one-sample case, so it needs to be subsequently adjusted for the one-sample case before use in this study. The test statistics T_{BS} and T_{SD} are as follows:

$$T_{BS} = \frac{n\overline{\mathbf{x}}'\overline{\mathbf{x}} - \mathrm{tr}(\mathbf{S})}{\sqrt{\dfrac{2n(n-1)}{(n-2)(n+1)}\left(\mathrm{tr}(\mathbf{S}^2) - \dfrac{\mathrm{tr}^2(\mathbf{S})}{n-1}\right)}} \tag{4}$$

and $$T_{SD} = \frac{n\overline{\mathbf{x}}'\mathbf{D}_{SD}^{-1}\overline{\mathbf{x}} - \dfrac{(n-1)p}{n-3}}{\sqrt{2\left[\mathrm{tr}(\mathbf{R}^2) - \dfrac{p^2}{n-1}\right]c_{p,n}}}\,, \tag{5}$$

where $\overline{\mathbf{x}}$ and $\mathbf{S}$ are defined as in (2)–(3), $\mathbf{D}_{SD}^{-1} = \mathrm{diag}(1/s_{11}, \ldots, 1/s_{pp})$, where s_{ii} are the diagonal elements of $\mathbf{S}$, $\mathbf{R}$ is the sample correlation matrix and $c_{p,n} = 1 + \mathrm{tr}(\mathbf{R}^2)/p^{3/2}$. Both of the test statistics have asymptotic normality. The T_{BS} test is invariant under an orthogonal transformation $\mathbf{x} \to c\boldsymbol{\Gamma}\mathbf{x}$, where is a nonzero constant and $\boldsymbol{\Gamma}$ is a $p \times p$ matrix such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \mathbf{I}$ while the T_{SD} test is invariant under a group of scalar transformation $\mathbf{x} \to \mathbf{D}\mathbf{x}$, where $\mathbf{D} = \mathrm{diag}(c_1, \ldots, c_p)$ and $c_i \neq 0$, for all $i, i = 1, \ldots, p$. Other tests in the literature, such as that of Chen and Qin (2010) and Park and Ayyala (2013), were studied under different conditions involving the trace of the correlation matrix and they provided good results in certain situations.

Although existing tests proposed by Dempster (1958), Bai and Saranadasa (1996) and Srivastava and Du (2008) do not need the inverse of the sample covariance matrix, there are still some limitations in the sense that they are based on the assumption that the data dimension (p) increases at the same rate as the sample size (n), i.e. $p/n \to c \in (0, \infty)$, but, in practice, there are so many current datasets which have a dimension much larger than the sample size, $p >> n$ (Park and Ayyala, 2013). Motivated by this kind of data and also the previous literature, we propose a test for the mean vector based on the idea of keeping more or as much as possible of the information from the sample covariance matrix.

In this article, we consider a set of independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ drawn from a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where both the mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$ are unknown. The problem of interest is testing $H:\boldsymbol{\mu} = \mathbf{0}$ vs. $K:\boldsymbol{\mu} \neq \mathbf{0}$, and we propose a new test which is applicable to high-dimensional data. The organization of this paper is as follows. The new test statistic and its asymptotic null distribution as $p \to \infty$ are presented in the next section, followed by a report on the performance of the proposed test which is investigated through a simulation study. Next, the proposed test is applied to a DNA microarray dataset, and lastly, the conclusion of this study is provided.

2. Description of the Proposed Test

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n independent observations from a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where both the mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite covariance matrix $\boldsymbol{\Sigma}$ are unknown. Assuming $p \geq n$, we are interested in testing

$$H:\boldsymbol{\mu} = \mathbf{0} \text{ vs. } K:\boldsymbol{\mu} \neq \mathbf{0}\,. \tag{6}$$

Define the positive definite covariance matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\,, \tag{7}$$

and write $\boldsymbol{\Sigma}$ in blocks as

$$\boldsymbol{\Sigma}_{p\times p} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1m} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{m1} & \boldsymbol{\Sigma}_{m2} & \cdots & \boldsymbol{\Sigma}_{mm} \end{pmatrix} = (\boldsymbol{\Sigma}_{jk})\,, \tag{8}$$

where $\boldsymbol{\Sigma}_{jj}$, $j = 1, \ldots, m$ are blocks, or submatrices, on the diagonal of $\boldsymbol{\Sigma}$ and the dimension of $\boldsymbol{\Sigma}_{jj}$ is q_j, $q_j < n-4$ and $\sum_{j=1}^{m} q_j = p$. The population correlation matrix $\mathfrak{R}$ is defined as

$$\mathfrak{R} = \mathbf{D}_{\sigma}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}_{\sigma}^{-1/2} = (\mathfrak{R}_{jk})\,, \tag{9}$$

where $\mathbf{D}_\sigma = \text{diag}(\sigma_{11},...,\sigma_{pp})$, $\sigma_{ii}, i = 1,...,p$ are the diagonal elements of $\mathbf{\Sigma}$ and $\mathfrak{R}_{jj}$, $j = 1,...,m$ is a $q_j \times q_j$ submatrix, $q_j < n-4$ and $\sum_{j=1}^{m} q_j = p$.

We make an assumption on the population correlation matrix as follows:

$$\text{As } p \to \infty \text{ and } n < +\infty, \mathfrak{R}_{jk} \to \mathbf{0},\ j \neq k,$$
$$j,k = 1,...,m\ . \tag{10}$$

From the sample covariance matrixS, we partition as for $\mathbf{\Sigma}$. We define a block diagonal matrix $\mathbf{D}_q$ as $\mathbf{D}_q$= diag(S$_{11}$,S$_{22}$,... S$_{mm}$), where $\mathbf{S}_{jj}$, $j = 1,...,m$ are submatrices obtained from the sample covariance matrix **S**, giving

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

In the case where the first $m-1$ blocks $\mathbf{S}_{11}, \mathbf{S}_{22},..., \mathbf{S}_{(m-1)(m-1)}$ are of equal size $q_1 = q_2 = ... = q_{m-1} = q$, i.e. $p = (m-1)q + q_m$, where q is referred to as the "***common block size***" of $\mathbf{D}_q$. Since $\mathbf{S}_{jj}$ can be considered as the sample covariance matrix of dimension q_j with sample size n where $q_j < n-4$, then $\mathbf{S}_{jj}$, $j = 1,...,m$ are all invertible. Consequently, the block diagonal matrix $\mathbf{D}_q$ is also invertible—its inverse is $\mathbf{D}_q^{-1}$ = diag$(\mathbf{S}_{11}^{-1},...,\mathbf{S}_{mm}^{-1})$.

Now, we consider the statistic

$$T_n = n\overline{\mathbf{x}}'\mathbf{D}_q^{-1}\overline{\mathbf{x}}\ , \tag{12}$$

where $\overline{\mathbf{x}}$ is defined in (2). The following theorem gives the expectation and variance of T_n.

Theorem 1. Let $\mathbf{x}_1,...,\mathbf{x}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is positive definite. Then, under assumption (10), the expectation and variance of T_n are, respectively,

$$\text{(i)}\quad E\left(T_n\right) = \sum_{j=1}^{m} \frac{(n-1)q_j}{(n-q_j-2)}\ ,$$

$$\text{(ii)}\quad Var\left(T_n\right) = \sum_{j=1}^{m} \frac{2(n-1)^2(n-2)q_j}{(n-q_j-2)^2(n-q_j-4)}\ .$$

Proof. Partition the sample mean vector $\overline{\mathbf{x}}$ corresponding to the block sizes in $\mathbf{D}_q$, i.e.

$$\overline{\mathbf{x}} = \begin{pmatrix} \overline{\mathbf{x}}_1 \\ \overline{\mathbf{x}}_2 \\ \vdots \\ \overline{\mathbf{x}}_m \end{pmatrix},\ \text{where } \overline{\mathbf{x}}_j \text{ is of dimension } q_j \times 1,$$

$$q_j < n-4 \text{ and } \sum_{j=1}^{m} q_j = p\ .$$

Write T_n as $T_n = n\overline{\mathbf{x}}'\mathbf{D}_q^{-1}\overline{\mathbf{x}} = \sum_{j=1}^{m} n\overline{\mathbf{x}}_j'\mathbf{S}_{jj}^{-1}\overline{\mathbf{x}}_j$

$$= \sum_{j=1}^{m} Y_j\ ,\ \text{where } Y_j = n\overline{\mathbf{x}}_j'\mathbf{S}_{jj}^{-1}\overline{\mathbf{x}}_j\ .$$

As Y_j has Hotelling's T^2 distribution with q_j and $n-1$ degrees of freedom, it can also be converted to an F-statistic as follows:

$$\frac{n-q_j}{(n-1)q_j} Y_j \sim F_{q_j, n-q_j}\ ,$$

where $F_{q_j, n-q_j}$ is the F-distribution with q_j and $n-q_j$ degrees of freedom.

Since $E\left(F\right) = \dfrac{n-q_j}{n-q_j-2}$ and

$$Var\left(F\right) = \frac{2(q_j-2)^2(n-2)}{q_j(n-q_j-2)^2(n-q_j-4)},$$

where $F \sim F_{q_j,n-q_j}$, then we have

$$E(T_n) = \sum_{j=1}^{m} E(Y_j) = \sum_{j=1}^{m} \frac{(n-1)q_j}{(n-q_j-2)}\ .$$

Under assumption (10), Y_j and Y_k are independent when $j \neq k,\ j,k = 1,...,m$, then $Cov\left(Y_j, Y_k\right) = 0$ and

$$Var\left(T_n\right) = \sum_{j=1}^{m} \frac{2(n-1)^2(n-2)q_j}{(n-q_j-2)^2(n-q_j-4)}\ .$$

This completes the proof.

We propose a test, for the problem $H: \boldsymbol{\mu} = \mathbf{0}$ vs. $K: \boldsymbol{\mu} \neq \mathbf{0}$, which is based on the statistic T_n as

$$T_q = \frac{T_n - \sum_{j=1}^{m} \frac{(n-1)q_j}{(n-q_j-2)}}{\sqrt{\sum_{j=1}^{m} \frac{2(n-1)^2(n-2)q_j}{(n-q_j-2)^2(n-q_j-4)}}}\ . \tag{13}$$

It is noted that the T_q test is invariant under a group of scalar transformation $\mathbf{x} \to \mathbf{D}\mathbf{x}$, where $\mathbf{D} = \text{diag}(c_1,...,c_p)$ and $c_i \neq 0$, for all i, $i = 1,...,p$; this is a desirable property of the test. The following theorem gives the asymptotic distribution of the test statistic T_q under the null hypothesis.

Theorem 2. Under assumption (10) and under the null hypothesis $\boldsymbol{\mu} = \mathbf{0}$. Then

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

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

Proof. This follows from Theorem 1 and by applying the Central Limit Theorem for non-identically distributed random variables. For a given *n*, as $p \to \infty$, which is equivalent to $m \to \infty$, 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 proposed test statistic T_q has the statistic $T_n = n\overline{\mathbf{x}}'\mathbf{D}_q^{-1}\overline{\mathbf{x}}$ in its numerator and the term T_n uses more information from the off-diagonal elements of the sample covariance matrix than the terms $n\overline{\mathbf{x}}'\overline{\mathbf{x}}$ in T_{BS} and $n\overline{\mathbf{x}}'\mathbf{D}_{SD}^{-1}\overline{\mathbf{x}}$ in T_{SD}.

One point of interest here is deciding how large to make the block sizes because, theoretically, the test statistic, based on the *F*-distribution, only requires block sizes $q_j < n-4$. The reasoning is as follows. When block sizes are close to $n-4$, it is more likely that the smallest eigenvalue of $\mathbf{S}_{jj}$ is close to zero; this can make the matrix $\mathbf{D}_q$ singular (Tracy and Widom, 1996). In addition, based on the idea of obtaining as much information from the sample covariance matrix as possible, it is recommended from our simulation results, not presented here, that the appropriate block sizes are $n-6$ (for $0 < n/p < 0.5$) and $n-7$ (for $0.5 \le n/p \le 1$).

3. Simulation Study

In this section, the performance of the proposed test was evaluated using a simulation study with a variety of parameter settings of the population covariance matrices. For comparison, the other two important statistics T_{SD} and T_{BS} were also included in the study, and the attained significance level and the empirical power of the three test statistics were evaluated. Firstly, the attained significance level and the empirical power are defined.

3.1 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. The ASL is computed as

$$\text{ASL} = \frac{\text{number of } t_H > z_{1-\alpha}}{m} ,$$

where t_H represents the values of the test statistic T based on the datasets generated under the null hypothesis. Throughout the simulation study, $m = 10,000$ iterations was chosen and the nominal significance level (α) was fixed at 0.05. The ASL is approximately distributed as the binomial distribution $b(10000, 0.05)$ and has a standard deviation estimated by $\widehat{se}(\text{ASL}) = \sqrt{0.05(0.95)/10,000} \simeq 0.00218$.

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

$$\text{Empirical power} = \frac{\text{number of } t_K > z_{1-\alpha}}{m} ,$$

where t_K represents the values of the test statistic T based on the datasets generated under the alternative hypothesis.

3.2 Parameter selection

The mean vector for the alternative hypothesis is $\boldsymbol{\mu} = \boldsymbol{\mu}_1 = (v_1, v_2, ..., v_p)'$, $v_{2k-1} = 0$ and $v_{2k} \overset{iid}{\sim} U(-0.5, 0.5)$ $k = 1, 2, ..., p/2$. The four following forms of covariance matrix were considered: (1) the identity matrix $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$, (2) a covariance matrix with a common block size, (3) a covariance matrix with a common block size but different elements, and (4) a covariance matrix with various block sizes. The second form of covariance matrix is $\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{\Sigma}_{11},...,\boldsymbol{\Sigma}_{mm})$, where $\boldsymbol{\Sigma}_{jj} = c\mathbf{I} + (1-c)\mathbf{1}\mathbf{1}'$, $c = 0.8$ and $\mathbf{1}$ is a vector where all entries are 1's, $\boldsymbol{\Sigma}_{jj}, j = 1,...,m-1$ are of dimension *q*, and the last block is q_m, where $p = (m-1)q + q_m$. The third form of covariance matrix is $\boldsymbol{\Sigma}_3 = \mathbf{D}_\sigma^{1/2}\mathfrak{R}\mathbf{D}_\sigma^{1/2}$, 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 $\mathfrak{R}_{kk} = (\rho_{ij})$, $\rho_{ij} = (-1)^{i+j}(c^{|i-j|})$, $i, j = 1,...,q_j$, $c = 0.9$ and $k = 1,...,m$. The block sizes of the third form of covariance matrix are similar to those of the second one, i.e. the common block size is q and the last block size is q_m, where $p = (m-1)q + q_m$. The last form of covariance matrix is $\boldsymbol{\Sigma}_4 = \mathbf{D}_\sigma^{1/2}\mathfrak{R}\mathbf{D}_\sigma^{1/2}$, where $\boldsymbol{\Sigma}_4$ is formed as in the third case except that the blocks in $\mathfrak{R}$ are of five different sizes and these blocks are randomly located on the diagonal.

The simulations were conducted at $p = 60$ with $n \in \{30, 60\}$ and at $p \in \{100, 150, 200, 400\}$ with $n \in \{40, 60, 80\}$. For each combination of the data dimension (*p*) and sample size (*n*), the proposed statistic T_q was computed for the chosen common block size (*q*) of $n-6$ (for $0 < n/p < 0.5$) or $n-7$ (for $0.5 \le n/p \le 1$). The common block size of 1

($q = 1$) was chosen in order to compare with the other two statistics from the literature when the covariance matrix is an identity matrix. Under each condition, n multivariate normal vectors with the chosen mean vector and covariance matrix were generated, then the ASL and the empirical power were recorded.

3.3 Simulation results

The performance of the proposed test was evaluated through simulations using 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 $\mathbf{D}_q$ were $q = 1$ and $q = n - 6$ (for $0 < n/p < 0.5$) or $q = n - 7$ (for $0.5 \le n/p \le 1$). The case of $q = 1$ was studied here in order to compare with the existing test statistic T_{SD}, presented by Srivastava and Du (2008). As is known, when the population covariance matrix is the identity matrix, the sample counterpart need not be the identity, so we are interested in cases where the common block size in matrix $\mathbf{D}_q$ is $q = n - 6$ (or $q = n - 7$). The results, from Table 1, show that with the common block size $q = 1$, the proposed statistic T_q performed well overall, and quite similarly to Bai and

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

p	n	ASL			Empirical Power		
		T_{SD}	T_{BS}	T_q	T_{SD}	T_{BS}	T_q
$q = 1$							
60	30	0.057	0.061	0.059	0.998	0.999	0.998
	60	0.050	0.057	0.057	1.000	1.000	1.000
100	40	0.053	0.060	0.060	1.000	1.000	1.000
	60	0.051	0.058	0.058	1.000	1.000	1.000
	80	0.052	0.059	0.060	1.000	1.000	1.000
150	40	0.046	0.054	0.054	1.000	1.000	1.000
	60	0.047	0.054	0.055	1.000	1.000	1.000
	80	0.048	0.055	0.056	1.000	1.000	1.000
200	40	0.043	0.054	0.053	1.000	1.000	1.000
	60	0.046	0.055	0.056	1.000	1.000	1.000
	80	0.042	0.052	0.051	1.000	1.000	1.000
400	40	0.039	0.054	0.055	1.000	1.000	1.000
	60	0.041	0.055	0.055	1.000	1.000	1.000
$q = n - 6$ (for $0 < n/p < 0.5$) or $n - 7$ (for $0.5 \le n/p \le 1$)							
60	30	0.057	0.061	0.049	0.998	0.999	0.512
	60	0.050	0.057	0.047	1.000	1.000	0.581
100	40	0.053	0.060	0.053	1.000	1.000	0.653
	60	0.051	0.058	0.050	1.000	1.000	0.848
	80	0.052	0.059	0.047	1.000	1.000	0.761
150	40	0.046	0.054	0.050	1.000	1.000	0.744
	60	0.047	0.054	0.044	1.000	1.000	0.797
	80	0.048	0.055	0.048	1.000	1.000	0.932
200	40	0.043	0.054	0.049	1.000	1.000	0.841
	60	0.046	0.055	0.053	1.000	1.000	1.000
	80	0.042	0.052	0.048	1.000	1.000	1.000
400	40	0.039	0.054	0.053	1.000	1.000	0.991
	60	0.041	0.055	0.052	1.000	1.000	0.996
	80	0.043	0.055	0.051	1.000	1.000	0.998

Table 2. ASLs and Empirical Powers when $\mathbf{\Sigma} = \mathbf{\Sigma}_2$ at Nominal Significance Level $\alpha = 0.05$

p	n	ASL			Empirical Power		
		T_{SD}	T_{BS}	T_q	T_{SD}	T_{BS}	T_q
60	30	0.030	0.076	0.049	0.230	0.499	1.000
	60	0.018	0.074	0.047	0.160	0.810	1.000
100	40	0.022	0.077	0.053	0.336	0.799	1.000
	60	0.019	0.079	0.050	0.362	0.974	1.000
	80	0.015	0.073	0.047	0.503	1.000	1.000
150	40	0.030	0.079	0.050	0.558	0.925	1.000
	60	0.018	0.079	0.044	0.613	0.998	0.994
	80	0.014	0.078	0.048	0.626	1.000	1.000
200	40	0.028	0.072	0.049	0.705	1.000	0.946
	60	0.019	0.079	0.053	0.839	1.000	1.000
	80	0.012	0.071	0.048	0.926	1.000	1.000
400	40	0.023	0.066	0.053	0.999	1.000	1.000
	60	0.018	0.070	0.052	1.000	1.000	1.000
	80	0.014	0.073	0.051	1.000	1.000	1.000

Table 3. ASLs and Empirical Powers when $\mathbf{\Sigma} = \mathbf{\Sigma}_3$ at Nominal Significance Level $\alpha = 0.05$

p	n	ASL			Empirical Power		
		T_{SD}	T_{BS}	T_q	T_{SD}	T_{BS}	T_q
60	30	0.048	0.080	0.049	0.195	0.153	1.000
	60	0.039	0.075	0.047	0.461	0.245	1.000
100	40	0.039	0.071	0.053	0.405	0.221	1.000
	60	0.039	0.075	0.050	0.730	0.337	1.000
	80	0.038	0.071	0.047	0.966	0.506	1.000
150	40	0.038	0.069	0.050	0.575	0.290	1.000
	60	0.038	0.066	0.044	0.913	0.416	1.000
	80	0.038	0.070	0.048	0.998	0.633	1.000
200	40	0.041	0.070	0.049	0.681	0.303	1.000
	60	0.039	0.070	0.053	0.978	0.508	1.000
	80	0.036	0.066	0.048	1.000	0.740	1.000
400	40	0.036	0.067	0.053	0.970	0.545	1.000
	60	0.035	0.066	0.052	1.000	0.852	1.000
	80	0.038	0.065	0.051	1.000	0.986	1.000

Saranadasa's statistic T_{BS}, while the ASL of the statistic T_{SD} was lower than the nominal significance level 0.05 when the data dimension increased. When the common block size in matrix $\mathbf{D}_q$ was $q = n-6$ (or $q = n-7$), the ASL of the proposed statistic T_q was closer to the nominal level than the other two tests. Although the empirical power of the proposed test was rather low for a small p, or $p < 200$, it became acceptably high when the dimension increased ($p \geq 200$). Moreover, when comparing the maximum difference of the ASL from the nominal value, $\max\left|\text{ASL} - 0.05\right|$, of the three test statistics when $p \geq 200$, Table 1 shows that the statistic T_{SD} performed the worst; $\max\left|\text{ASL of } T_{SD} - 0.05\right| = 0.011$ while $\max\left|\text{ASL of } T_{BS} - 0.05\right| = 0.005$ and $\max\left|\text{ASL of } T_q - 0.05\right| = 0.006$.

For the covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_2$ and the common block size in $\mathbf{D}_q$ chosen corresponded to $\mathbf{\Sigma}$, the results are

Table 4. ASLs and Empirical Powers when $\mathbf{\Sigma} = \mathbf{\Sigma}_4$ at Nominal Significance Level $\alpha = 0.05$

p	n	ASL			Empirical Power		
		T_{SD}	T_{BS}	T_q	T_{SD}	T_{BS}	T_q
0	30	0.048	0.077	0.045	0.219	0.153	1.000
	60	0.039	0.073	0.047	0.489	0.252	1.000
100	40	0.039	0.067	0.048	0.450	0.225	1.000
	60	0.040	0.073	0.051	0.806	0.366	1.000
	80	0.038	0.070	0.049	0.982	0.558	1.000
150	40	0.039	0.068	0.048	0.611	0.270	1.000
	60	0.037	0.066	0.048	0.936	0.437	1.000
	80	0.037	0.069	0.047	1.000	0.667	1.000
200	40	0.041	0.069	0.046	0.732	0.316	1.000
	60	0.040	0.069	0.046	0.982	0.523	1.000
	80	0.036	0.066	0.046	1.000	0.760	1.000
400	40	0.036	0.066	0.054	0.975	0.555	1.000
	60	0.037	0.068	0.049	1.000	0.863	1.000
	80	0.035	0.065	0.048	1.000	0.989	1.000

shown in Table 2. In this form of covariance matrix, the proposed statistic T_q performed well and was superior to both of the statistics T_{SD} and T_{BS}. The results when the covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_3$, 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.

For the last form of covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_4$, which contained blocks of five different sizes 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.

4. A Real Example

In this section, the proposed test with a 5% significance level was applied to a sample of DNA microarray data from an oncology study. The data, published by Notterman *et al.* (2001), were retrieved on Nov 23, 2014 from the Princeton University Gene Expression Project website (http://genomics-pubs.princeton.edu/oncology). A selection of 200 genes (p) with the sample size (n) 18 were analyzed with using the differences between tumor tissue and normal tissue of gene expression levels as variables. To compute the proposed test statistic T_q, the variables in blocks were arranged so that the correlation coefficient of any two adjacent variables in the same block was greater than or equal to 0.5. The test values T_{BS} of Bai and Saranadasa, in (4), and T_{SD} of Srivastava and Du, in (5), were also computed. The calculated test value of the proposed statistic was $T_q = 19.737$ with corresponding p-value < 0.001. The other two test values were $T_{BS} = 21.299$ and $T_{SD} = 21.773$, both of which had corresponding p-value lower than 0.001. Thus, all three tests led to the rejection of the null hypothesis of a zero mean vector, i.e. the gene expression levels of tumor tissue are significantly different from those of normal tissue.

5. Conclusions

In this study, we proposed a test for the mean vector when the data dimension is larger than the sample size and the data are multivariate normal. The development of the test is based on the idea of keeping more information from the sample covariance matrix than similar previously published tests and follows the concept of the Mahalanobis distance. One of the desirable properties of the proposed test is that it is invariant under a group of scalar transformation. Among the tests reviewed in the literature, two important tests, one proposed by Bai and Saranadasa (1996) and the other by Srivastava and Du (2008), are highlighted and also used for comparison in a simulation study. Under the null hypothesis, the proposed statistic has been shown to converge in distribution to a standard normal distribution when the data dimension $p \to \infty$.

Acknowledgments

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

References

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

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(2), 808–835.

Dempster, A.P. 1958. A high dimensional two sample significance test. The Annals of Mathematical Statistics. 29 (4), 995–1010.

Eaton, M.L. and Perlman, M.D. 1973. The Non-Singularity of Generalized Sample Covariance Matrices. The Annals of Statistics. 1(4), 710–717.

Mahalanobis, P.C. 1936. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India, Calcutta, India, 1936, 49–55.

Notterman, D.A., Alon, U., Sierk, A.J. and 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. and 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.

Rencher, A.C. 2001. Methods of Multivariate Analysis, John Wiley and Sons, New York, U.S.A., pp. 118–120.

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.

Srivastava, M.S. and Du, M. 2008. A test for the mean vector with fewer observations than the dimension. Journal of Multivariate Analysis. 99(3), 386–402.

Tracy, C.A. and Widom, H. 1996. On orthogonal and symplectic matrix ensembles. Communications in Mathematical Physics. 177, 727–754.

Zhang, J. and Xu, J. 2009. On the k-sample Behrens-Fisher problem for high-dimensional data. Science in China, Series A: Mathematics. 52(6), 1285–1304.