Journal of Mathematics and Statistics

Original Research Paper

A Test on the Multivariate Behrens–Fisher Problem in High–Dimensional Data by Block Covariance Estimation

Paranut Sukcharoen and Samruam Chongcharoen

Graduate School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand

Article history
Received: 22-11-2018
Revised: 14-03-2019
Accepted: 01-04-2019

Corresponding Author:
Paranut Sukcharoen
Graduate School of Applied Statistic, National Institute of Development Administration, Bangkok, Thailand
Email: louis.paranut@gmail.com

Abstract: In this paper, we proposed a new testing statistic for testing the equality of mean vectors from two multivariate normal populations when the covariance matrices are unknown and unequal in high–dimensional data. A new test is proposed based on the idea of keeping more information from the sample covariance matrices as much as possible. A proposed test is invariant under scalar transformations and location shifts. We showed that the asymptotic distribution of proposed statistic is standard normal distribution when number of random variables approach infinity. We also compared the performance of the proposed test with other three existing tests by the simulation study. The simulation results showed that the attained significance level of proposed test close to setting nominal significance level satisfactorily. The attained power of proposed test outperforms as the other comparative tests under form of covariance matrices considered which can be arranged to block diagonal matrix structure. The attained power becomes more powerful when the dimension increases for a given sample size or vice versa, or relationship level between random variables in each sample increases. Finally, the proposed test is also illustrated with an analysis of DNA microarray data.

Keywords: Hypothesis Testing, Two–Sample Mean Vectors, Multivariate Behrens–Fisher Problem, High–Dimensional Data, Block Diagonal Matrix Structure

Introduction

Currently data collecting technology is rapidly evolving. Its evolution makes the statistical methods going to two directions. When the sample is being collected more and more, the first direction of the statistical methods will focus about asymptotic optimality of statistical methods. In the other direction, when variables or dimensions of data are being considered increasingly, the focus of statistical analysis shifted from the univariate to multivariate (Zhou, 2016). However, in many practical applications of modern multivariate statistical methods often found a data sets which are much larger number of measurements than the sample size. In this case, it will be referred to high–dimensional data, which referring to a large number of measurements are taken on comparably many or relatively few subjects. High–dimensional data appears in various fields, such as online data from markets around the world are accumulated on a Giga–octet basis every day in financial studies, gene expression data that collects from DNA microarray technology in genetic experiments (Yao *et al*., 2015). In such high–dimensional data, classical multivariate statistical methods is not often applicable because they involved with the inversion of sample covariance matrix which does not exist.

Now suppose $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}$ represent a random sample with size n_1 and n_2 from p–dimensional multivariate normal random vectors from the ith group, i = 1,2, each of which has $p \times 1$ mean vector $\boldsymbol{\mu}_i$ and $p \times p$ unknown positive definite covariance matrix $\boldsymbol{\Sigma}_i$ or $\mathbf{x}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. The problem of testing for the equality of means vectors from two multivariate normal population when the covariance matrices are unknown and unequal or when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ is referred to as the multivariate Behrens–Fisher problem. That is, we are considering the testing hypothesis as:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{1}$$

where, p denotes the dimension or the number of variables with $p \leq n_1 + n_2 - 2$. A natural invariant test statistic for testing hypothesis in (1) is:

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \left(\frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2} \right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \quad (2)$$

where the $p \times 1$ sample mean vectors $(\bar{\mathbf{x}}_i)$ and the $p \times p$ sample covariance matrix ($\mathbf{S}_i$) are defined, respectively by:

$$\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \quad (3)$$

and:

$$\mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)', \quad i = 1,2. \quad (4)$$

The exact distribution of T^2 under the null hypothesis in (1) was obtained by Nel and van der Merwe (1990). However, the exact distribution is very complicated and which are extremely difficult to compute in practice, it is of no use for practical applications. But the approximation of this testing statistic is used when both sample size n_1 and n_2 approach infinity, the distribution of T^2 converges to the chi–square distribution with p degrees of freedom. Its approximation is very simple and easier to compute, but this approximation is suffering from either the sample size n_1 or n_2 is small. So, this approximation is more accurate when $\min(n_1, n_2) \to \infty$, (Srivastava, 2002; Yanagihara and Yuan, 2005; Richard and Dean, 2014). Unfortunately, in practice the sample size is not very large, so this approximation is not recommended for application in practice

There is a vast literature devoted to the solution of this problem, many researchers tried to approximate the distribution of T^2 by a constant times F–distribution with numerator degrees of freedom p and the approximate denominator degrees of freedom estimate from sample size, sample mean vector and sample covariance matrix. Among the approximate solutions based on T^2, some approximate solutions suggested by James (1954), Yao (1965), Johansen (1980) and Yanagihara and Yuan (2005) are invariant, whereas the solution due to Nel and van der Merwe (1986) is not invariant. Afterward, Krishnamoorthy and Yu (2004) modified the solution of Nel and van der Merwe (1986) by providing an invariant test statistic and Kawasaki and Seo (2015) improved the solution of Yanagihara and Yuan (2005) by asymptotic expansions.

From literary review, we found that solution due to Krishnamoorthy and Yu (2004) has the attained significance level close to the nominal significance level satisfactorily, Krishnamoorthy and Xia (2006) among others showed via intensive simulation studies that this test performs best among the approximation solutions to the multivariate Behrens–Fisher problem (Zhou, 2016). Krishnamoorthy and Yu (2004) has been shown to have approximately distribution of T^2 as F–distribution is given by:

$$T^2 \sim \frac{vp}{v - p + 1} F_{p,v-p+1}, \text{ approximately,} \quad (5)$$

where, $F_{p,v-p+1}$ denotes a random variable with an F–distribution with p and $v - p + 1$ degrees of freedom and the degrees of freedom v are estimated from the sample covariance matrices using the relation:

$$v = \frac{p + p^2}{\sum_{i=1}^{2} \frac{\mathrm{tr}\left[\left(\mathbf{S}_i \tilde{\mathbf{S}}^{-1} \right)^2 \right] + \left[\mathrm{tr}\left(\mathbf{S}_i \tilde{\mathbf{S}}^{-1} \right) \right]^2}{n_i^2 (n_i - 1)}}. \quad (6)$$

where, the sample covariance matrix $\tilde{\mathbf{S}}$ defined by:

$$\tilde{\mathbf{S}} = \frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2} \quad (7)$$

and $\min(n_1 - 1, n_2 - 1) \le v \le n_1 + n_2 - 2$, this approximation reduces to the usual Welch's approximate degrees of freedom to the Behrens–Fisher problem in the univariate (p = 1) case (Richard and Dean, 2014).

In high–dimensional data, for one population when the data has the number of variable exceed sample size (minus 1), $p > n_i - 1$, for example the data that collects from DNA microarrays technology where a large number of gene expression levels may be in the hundreds or thousands, are measured on relatively few subjects (Zhou *et al*., 2017), then the sample covariance matrix $\mathbf{S}_i$ lose its full rank and will be singular, which makes $\mathbf{S}_i$ does not have an inverse (Chongcharoen, 2011). Furthermore, for two populations when the data has the number of variable is larger than the sum of the sample sizes (minus 2), $p > n_1 + n_2 - 2$, then the sample covariance matrix $\tilde{\mathbf{S}}$ in (7) does not have an inverse. Hence, any statistic value involving inversion of $\tilde{\mathbf{S}}$ does not exist. However, test statistic T^2 in (2) requires the matrix $\tilde{\mathbf{S}}$ invertible, so it cannot be applied for high–dimensional data.

To overcome the problem of the need for the inverse of a sample covariance matrix in high–dimensional data, many efforts recently have been devoted to construct new test solutions for multivariate Behrens–Fisher problem in high–dimensional data. Most test statistics try to avoid the use of $\tilde{\mathbf{S}}^{-1}$. This problem has been considered by Bai and Saranadasa (1996) who proposed a test statistic is develop by using only information from the diagonal elements of $\mathbf{S}_i$, i = 1,2, as T_{BS} given by:

$$T_{BS} = Q_n/\hat{\sigma}_Q, \quad (8)$$

where, Q_n are defined by:

$$Q_n = \left[(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)-\frac{\mathrm{tr}(\mathbf{S}_1)}{n_1}-\frac{\mathrm{tr}(\mathbf{S}_2)}{n_2}\right]\Big/\sqrt{p}, \quad (9)$$

which the variance of this statistics σ_Q^2 is given by:

$$\sigma_Q^2 = \frac{2}{p}\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)}{n_1^2}+\frac{\mathrm{tr}(\boldsymbol{\Sigma}_2^2)}{n_2^2}+\frac{2\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}{n_1n_2}\right\} \quad (10)$$

Srivastava (2009) proposed a consistent estimator of σ_Q^2 in (10) as:

$$\hat{\sigma}_Q^2 = \frac{2}{p}\left\{\frac{\hat{a}_{21}}{n_1^2}+\frac{\hat{a}_{22}}{n_2^2}+\frac{2\mathrm{tr}(\mathbf{S}_1\mathbf{S}_2)}{n_1n_2}\right\} \quad (11)$$

Where:

$$\hat{a}_{2i} = \frac{(n_i-1)^2}{n_i(n_i-2)}\left\{\mathrm{tr}(\mathbf{S}_i^2)-\frac{\left[\mathrm{tr}(\mathbf{S}_i)\right]^2}{n_i-1}\right\},\ i=1,2. \quad (12)$$

Chen and Qin (2010) proposed a test statistic based on sidesteps covariance matrix estimation (Gregory *et al*., 2015) as T_{CQ} given by:

$$T_{CQ} = Q_n/\tilde{\sigma}_Q, \quad (13)$$

where, Q_n are defined as (9) and $\tilde{\sigma}_Q^2$ is given by:

$$\tilde{\sigma}_Q^2 = \frac{2}{p}\left\{\frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_1^2})}{n_1(n_1-1)}+\frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_2^2})}{n_2(n_2-1)}+\frac{2\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2})}{n_1n_2}\right\} \quad (14)$$

which as:

$$\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_i^2}) = \frac{\mathrm{tr}\left\{\sum_{j\neq k}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_{i(j,k)})\mathbf{x}_{ij}'(\mathbf{x}_{ik}-\bar{\mathbf{x}}_{i(j,k)})\mathbf{x}_{ik}'\right\}}{n_i(n_i-1)},\ i=1,2. \quad (15)$$

$$\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2}) = \frac{\mathrm{tr}\left\{\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}(\mathbf{x}_{1j}-\bar{\mathbf{x}}_{1(j)})\mathbf{x}_{1j}'(\mathbf{x}_{2k}-\bar{\mathbf{x}}_{2(k)})\mathbf{x}_{2k}'\right\}}{n_1n_2} \quad (16)$$

where, $\bar{\mathbf{x}}_{i(j,k)}$ is the ith sample mean vector after excluding $\mathbf{x}_{ij}$ and $\mathbf{x}_{ik}$, for $i = 1,2$; $j, k = 1,2,,\ldots,n_i$, given by:

$$\bar{\mathbf{x}}_{i(j,k)} = \frac{1}{n_i-2}(n_i\bar{\mathbf{x}}_i-\mathbf{x}_{ij}-\mathbf{x}_{ik}) \quad (17)$$

and $\bar{\mathbf{x}}_{i(k)}$ is the ith sample mean vector without $\mathbf{x}_{ik}$ for $i = 1,2$; $k = 1,2,,\ldots,n_i$, given by:

$$\bar{\mathbf{x}}_{i(k)} = \frac{1}{n_i-1}(n_i\bar{\mathbf{x}}_i-\mathbf{x}_{ik}) \quad (18)$$

Srivastava *et al*. (2013) proposed a test statistic which uses the diagonal matrix of the sample covariance matrix $\tilde{\mathbf{S}}$ and the trace of the sample correlation matrix as T_{SKK} given by:

$$T_{SKK} = \frac{(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\tilde{\mathbf{S}}_{diagonal}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)-p}{\sqrt{p\widehat{Var}(\hat{q}_n)\left(1+\frac{\mathrm{tr}(\mathbf{R}^2)}{p^{3/2}}\right)}}, \quad (19)$$

where, $\tilde{\mathbf{S}}_{diagonal}$ is the diagonal matrices of the diagonal elements of matrix $\tilde{\mathbf{S}}$ in (7) and $\mathbf{R}$ is defined by:

$$\mathbf{R} = \tilde{\mathbf{S}}_{diagonal}^{-1/2}\tilde{\mathbf{S}}\tilde{\mathbf{S}}_{diagonal}^{-1/2} \quad (20)$$

and $\widehat{Var}(\hat{q}_n)$ is given by:

$$\widehat{Var}(\hat{q}_n) = \frac{2}{p}\left\{\mathrm{tr}(\mathbf{R}^2)-\sum_{i=1}^{2}\frac{\left[\mathrm{tr}(\tilde{\mathbf{S}}_{diagonal}^{-1}\mathbf{S}_i)\right]^2}{n_i^2(n_i-1)}\right\} \quad (21)$$

All three test statistics T_{BS}, T_{CQ} and T_{SKK} have asymptotic standard normal distribution under null hypothesis in (1). Both T_{BS} and T_{CQ} tests are invariant under an orthogonal transformation, $\mathbf{x}_{ij} \rightarrow \mathbf{P}\mathbf{x}_{ij}$, $i = 1,2$, $j = 1,2,\ldots,n_i$, where $\mathbf{P}$ is an orthogonal $p\times p$ matrix such that $\mathbf{P}'\mathbf{P} = \mathbf{I}$. In contrast, the T_{SKK} test is invariant under location shifts and scalar transformations, $\mathbf{x}_{ij} \rightarrow \mathbf{D}\mathbf{x}_{ij}+\mathbf{c}$, $i = 1,2$, $j = 1,2,\ldots,n_i$, where $\mathbf{D}$ is nonsingular $p\times p$ diagonal matrice and $\mathbf{c}$ is a constant vector.

The performance of these three tests will be compared with the proposed tests. Other tests in the literature, such as that of Katayama and Kano (2014); Gregory *et al*., (2015), were studied a test on high–dimensional mean vector under without any assumption on population covariance matrix and not assume normally distributed. Zhang and Xu (2009); Yamada and Himeno (2015); Hu *et al*. (2017) proposed a testing the equality of several high–dimensional mean vectors with unequal covariance matrices, that is: The heteroscedastic one–way multivariate analysis of variance (MANOVA). Nishiyama *et al*. (2013); Zhou *et al*. (2017) proposed a high–dimensional general linear hypothesis testing problem on mean vectors of several populations under heteroscedasticity.

In this paper, we interested to use block diagonal structures of $\tilde{\mathbf{S}}$ in (7) to solve problem that the inverse of $\tilde{\mathbf{S}}$ does not exist. The test is very simple and provide more accurate new approximate test statistic for testing in the multivariate Behrens–Fisher problem in high–dimensional data. Based on the idea of keeping the information of $\mathbf{S}_i$ as much as possible (Jiamwattanapong and Chongcharoen, 2015; 2017), the asymptotic null distribution of proposed testing statistic is presented in section 2. The performance of the proposed testing statistic along with three existing tests will be investigated through a simulation study in section 3. Applying the proposed test by using a real DNA microarray data will be demonstrated in section 4. Finally, some conclusions are given.

A Proposed Test Statistic and Its Asymptotic Distribution

In this section, we proposed a test statistics for testing hypothesis in (1) in high–dimensional data case, that is, when $p > n_1 + n_2 - 2$. Consider the population covariance matrix $\tilde{\mathbf{\Sigma}}$ as:

$$\tilde{\mathbf{\Sigma}} = \frac{\mathbf{\Sigma}_1}{n_1} + \frac{\mathbf{\Sigma}_2}{n_2}, \tag{22}$$

which can be write $\tilde{\mathbf{\Sigma}}$ in blocks diagonal structures as:

$$\tilde{\mathbf{\Sigma}} = \begin{bmatrix} \tilde{\mathbf{\Sigma}}_{11} & \tilde{\mathbf{\Sigma}}_{12} & \cdots & \tilde{\mathbf{\Sigma}}_{1m} \\ \tilde{\mathbf{\Sigma}}_{21} & \tilde{\mathbf{\Sigma}}_{22} & \cdots & \tilde{\mathbf{\Sigma}}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{\Sigma}}_{m1} & \tilde{\mathbf{\Sigma}}_{m2} & \cdots & \tilde{\mathbf{\Sigma}}_{mm} \end{bmatrix}_{p\times p} = \left(\tilde{\mathbf{\Sigma}}_{kl}\right)$$

where, $\tilde{\mathbf{\Sigma}}_{kk}$ are $q_k \times q_k$ blocks matrices or submatrices on the diagonal of $\tilde{\mathbf{\Sigma}}$ with $k = 1,2,\ldots,m$, $m \le p$, and $\sum_{k=1}^{m} q_k = p$; m is the number of block on the diagonal of $\tilde{\mathbf{\Sigma}}$ and $q_k \times q_k$ is called the ***"block size"*** of kth block. The population correlation matrix $\mathfrak{R}$ is defined as:

$$\mathfrak{R} = \mathbf{D}^{-1/2}\tilde{\mathbf{\Sigma}}\mathbf{D}^{-1/2} \tag{23}$$

where, $\mathbf{D}$ is the matrix of the diagonal elements of $\tilde{\mathbf{\Sigma}}$. We can write $\mathfrak{R}$ in blocks diagonal structures as:

$$\mathfrak{R} = \begin{bmatrix} \mathfrak{R}_{11} & \mathfrak{R}_{12} & \cdots & \mathfrak{R}_{1m} \\ \mathfrak{R}_{21} & \mathfrak{R}_{22} & \cdots & \mathfrak{R}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathfrak{R}_{m1} & \mathfrak{R}_{m2} & \cdots & \mathfrak{R}_{mm} \end{bmatrix}_{p\times p} = \left(\mathfrak{R}_{ij}\right)$$

where, $\mathfrak{R}_{kk}$, $k = 1,2,\ldots,m$, $m \le p$, are size $q_k \times q_k$ block matrices or submatrices on the diagonal of $\mathfrak{R}$ with $\sum_{k=1}^{m} q_k = p$. In order to obtain the asymptotic null distribution, we make an assumption on the population correlation matrix as $p \to \infty, n_i < \infty, i = 1,2$, and $\mathfrak{R}_{kl} \to \mathbf{0}$, $k \ne l$, k, $l = 1,2,\ldots,m$.

From the assumption, the population covariance matrix $\tilde{\mathbf{\Sigma}}$ will be partitioned as block matrix structures. Thus the proposed test statistic based on constructing the sample covariance matrix $\tilde{\mathbf{S}}$ in (7) will be partitioned the same pattern as $\tilde{\mathbf{\Sigma}}$. To make it invertible, we made it as block diagonal matrix as:

$$\tilde{\mathbf{S}}_{block} = \begin{bmatrix} \tilde{\mathbf{S}}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{S}}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{S}}_{mm} \end{bmatrix}_{p\times p}, \tag{24}$$

where, $\tilde{\mathbf{S}}_{kk}$, $k = 1,2,\ldots,m$, $m \le p$, are size $q_k \times q_k$ block matrices or submatrices on the diagonal of $\tilde{\mathbf{S}}$ with $q_k < n_1 + n_2 - 2$ and $\sum_{k=1}^{m} q_k = p$. Since $q_k < n_1 + n_2 - 2$, then $\tilde{\mathbf{S}}_{kk}$, $k = 1,2,\ldots,m$ are all invertible. As a result, $\tilde{\mathbf{S}}_{block}$ is also invertible and the inverse of $\tilde{\mathbf{S}}_{block}$ can be obtained as:

$$\tilde{\mathbf{S}}_{block}^{-1} = \begin{bmatrix} \tilde{\mathbf{S}}_{11}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{S}}_{22}^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{S}}_{mm}^{-1} \end{bmatrix}_{p\times p}. \tag{25}$$

We substituted $\tilde{\mathbf{S}}_{block}^{-1}$ in place of $\tilde{\mathbf{S}}^{-1}$ in T^2 in (2) because $\tilde{\mathbf{S}}^{-1}$ does not exist for high–dimensional data.

Let a statistic T_n as:

$$T_n = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' \tilde{\mathbf{S}}_{block}^{-1} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right), \tag{26}$$

where, $\bar{\mathbf{x}}_i$, $i = 1,2$ defined in (3) and $\tilde{\mathbf{S}}_{block}^{-1}$ in (25). The following theorem gives the expectation and variance of the statistic T_n.

Theorem 1

Suppose $\mathbf{x}_{ij}$, be a random vectors from $N_p\left(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i\right)$, $i = 1,2$, $j = 1,2,\ldots,n_i$. Under assumption that the population correlation matrix as $p \to \infty, n_i < \infty, i = 1,2$, and $\mathfrak{R}_{kl} \to \mathbf{0}$, $k \ne l$, k, $l = 1,2,\ldots,m$. The expectation and variance of T_n in (26) are respectively:

$$E(T_n) = \sum_{k=1}^{m} \frac{v_k q_k}{v_k - q_k - 1}, \quad q_k < v_k - 1, \tag{27}$$

$$Var(T_n) = \sum_{k=1}^{m} \frac{2q_k (v_k - 1) v_k^2}{(v_k - q_k - 1)^2 (v_k - q_k - 3)}, \quad q_k < v_k - 3. \tag{28}$$

Proof

Partition the sample mean vectors $\bar{\mathbf{x}}_i$ and the sample covariance matrix $\mathbf{S}_i$, i = 1,2 in (3) and (4), corresponding to the block size as $\tilde{\mathbf{S}}_{block}$, i.e.:

$$\bar{\mathbf{x}}_i = \begin{bmatrix} \bar{\mathbf{x}}_{i1} \\ \bar{\mathbf{x}}_{i2} \\ \vdots \\ \bar{\mathbf{x}}_{im} \end{bmatrix}_{p\times 1}, \quad \text{and:} \quad \mathbf{S}_i = \begin{bmatrix} \mathbf{S}_{i11} & \mathbf{S}_{i12} & \cdots & \mathbf{S}_{i1m} \\ \mathbf{S}_{i21} & \mathbf{S}_{i22} & \cdots & \mathbf{S}_{i2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{im1} & \mathbf{S}_{im2} & \cdots & \mathbf{S}_{imm} \end{bmatrix}_{p\times p}.$$

where, $\bar{\mathbf{x}}_{ik}$ and $\mathbf{S}_{ikk}$, is of dimension $q_k \times 1$, $q_k \times q_k$, respectively. $q_k \le n_1 + n_2 - 2$, $\forall k$, k = 1,2,…,m, $m \le p$ and $\sum_{k=1}^{m} q_k = p$. So, we express T_n in (26) as:

$$\begin{aligned} T_n &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \tilde{\mathbf{S}}_{block}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \\ &= \underbrace{\begin{bmatrix} \bar{\mathbf{x}}_{11} - \bar{\mathbf{x}}_{21} \\ \bar{\mathbf{x}}_{12} - \bar{\mathbf{x}}_{22} \\ \vdots \\ \bar{\mathbf{x}}_{1m} - \bar{\mathbf{x}}_{2m} \end{bmatrix}'}_{1\times p} \underbrace{\begin{bmatrix} \tilde{\mathbf{S}}_{11}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{S}}_{22}^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{S}}_{mm}^{-1} \end{bmatrix}}_{p\times p} \underbrace{\begin{bmatrix} \bar{\mathbf{x}}_{11} - \bar{\mathbf{x}}_{21} \\ \bar{\mathbf{x}}_{12} - \bar{\mathbf{x}}_{22} \\ \vdots \\ \bar{\mathbf{x}}_{1m} - \bar{\mathbf{x}}_{2m} \end{bmatrix}}_{p\times 1}, \\ &= \sum_{k=1}^{m} Y_k, \quad \text{where, } Y_k = (\bar{\mathbf{x}}_{1k} - \bar{\mathbf{x}}_{2k})' \tilde{\mathbf{S}}_{kk}^{-1} (\bar{\mathbf{x}}_{1k} - \bar{\mathbf{x}}_{2k}). \end{aligned}$$

As the statistic Y_k corresponding to T^2 in (2). By Krishnamoorthy and Yu (2004), it can also be converted to a statistic of the F–distribution with numerator degrees of freedom q_k and the denominator degrees of freedom $v_k - q_k + 1$ as:

$$Y_k \sim \frac{v_k q_k}{v_k - q_k + 1} F_{q_k, v_k - q_k + 1}, \text{ approximately,} \tag{29}$$

where, v_k is approximate degrees of freedom in (6) of kth block which can be obtained by:

$$v_k = \frac{q_k + q_k^2}{\sum_{i=1}^{2} \frac{\mathrm{tr}\left[\left(\mathbf{S}_{ikk}\tilde{\mathbf{S}}_{kk}^{-1}\right)^2\right] + \left[\mathrm{tr}\left(\mathbf{S}_{ikk}\tilde{\mathbf{S}}_{kk}^{-1}\right)\right]^2}{n_i^2 (n_i - 1)}}. \tag{30}$$

We computed the expectation and variance of the statistics Y_k by applying the first moment and the second central moment of F–distribution with q_k and $v_k - q_k + 1$ degrees of freedom, respectively. Thus, we obtained:

$$\begin{aligned} E(Y_k) &= E\left[\frac{v_k q_k}{v_k - q_k + 1} F_{q_k, v_k - q_k + 1} \middle| v_k\right], \\ &= \frac{v_k q_k}{v_k - q_k - 1}, \quad q_k < v_k - 1, \\ Var(Y_k) &= Var\left[\frac{v_k q_k}{v_k - q_k + 1} F_{q_k, v_k - q_k + 1} \middle| v_k\right], \\ &= \frac{2q_k v_k^2 (v_k - 1)}{(v_k - q_k - 1)^2 (v_k - q_k - 3)}, \quad q_k < v_k - 3. \end{aligned}$$

Thus, the expectation and variance of the statistics T_n, respectively, can be obtained as:

$$\begin{aligned} E(T_n) &= \sum_{k=1}^{m} \frac{v_k q_k}{v_k - q_k - 1}, \quad q_k < v_k - 1, \\ Var(T_n) &= \sum_{k=1}^{m} Var(Y_k) + \sum_{k\ne l}^{m} Cov(Y_k, Y_l), \end{aligned}$$

Under uncorrelated assumption, Y_k and Y_l are uncorrelated when $k \ne l$, k,l = 1,2,…,m. Therefore, the covariance between Y_k and Y_l are zero or $Cov(Y_k, Y_l) = 0$. Thus:

$$Var(T_n) = \sum_{k=1}^{m} \frac{2q_k (v_k - 1) v_k^2}{(v_k - q_k - 1)^2 (v_k - q_k - 3)}, \quad q_k < v_k - 3.$$

The proof is completed.

We proposed a test statistic for testing the hypothesis in (1) based on the statistic T_n as:

$$T = \frac{T_n - \sum_{k=1}^{m} \frac{v_k q_k}{v_k - q_k - 1}}{\sqrt{\sum_{k=1}^{m} \frac{2q_k v_k^2 (v_k - 1)}{(v_k - q_k - 1)^2 (v_k - q_k - 3)}}} \tag{31}$$

Applying Lyapunov's Central Limit Theorem, we obtained:

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

This statistic make us reject H_0 in (1) at significance level α if the observed $T \ge z_{1-\alpha}$ where $z_{1-\alpha}$ denote the upper 100(1–α)% point of the standard normal distribution. It is noted that the proposed test statistic T is invariant under scalar transformations and location shifts, $\mathbf{x}_{ij} \to \mathbf{D}\mathbf{x}_{ij} + \mathbf{c}$, i =1,2, j = 1,2,…,n_i, where $\mathbf{D}$ is nonsingular $p \times p$ diagonal matrice and $\mathbf{c}$ is a constant vector.

It is obvious that the v_k is approximate degrees of freedom its corresponding to kth block with size $q_k \times q_k$ in the block diagonal matrix $\tilde{\mathbf{S}}_{block}$. Furthermore, by nature of v_k, we found that $v_k \to \min(n_1 - 1, n_2 - 1)$ when the difference between the sample covariance matrices $\mathbf{S}_{1kk}$ and $\mathbf{S}_{2kk}$ is large, on the other hand when the difference between the sample covariance matrices is slightly different, $v_k \to n_1 + n_2 - 2$. However, it is clear that the approximate degrees of freedom v_k in (30) lies in $[\min(n_1 - 1, n_2 - 1), n_1 + n_2 - 2]$ the same as the degrees of freedom v in (6). But if this condition $q_k \le v_k - 6$, $\forall k$, $k = 1,2,\ldots,m$, is true, the proposed test statistics T will be convergence in distribution to standard normal distribution, because the third central moment of Y_k is finite, For convenience and easy to use in practice, this condition may be changed to $q_k \le \min(n_1 - 1, n_2 - 1) - 6$. Since $q_k \ge 1$, so the proposed test statistics T can be usable when the both sample size n_1 and n_2 must be greater than or equal to 8.

One point of interest here is how large the block sizes in the block diagonal matrix $\tilde{\mathbf{S}}_{block}$. Since theoretically the proposed test statistics T based on the solution to approximation distribution of T^2 by Krishnamoorthy and Yu (2004), so it only requires block sizes as $q_k \le v_k - 6$, $\forall k$, $k = 1,2,\ldots,m$, whereas they gives recommendations about their solution that this solution has the attained significance level are very close to the nominal level provided $p \le \min(n_1 - 1, n_2 - 1)/5$ in unequal sample size cases and this condition is somewhat relaxed to $p \le n/4$, in equal sample size cases ($n_1 = n_2 = n$). So, based on their suggestions and the idea of keeping more information from the sample covariance matrix $\tilde{\mathbf{S}}$ as much as possible, we can give some guidance when there is no prior information to arrange variables from the sample covariance matrix $\tilde{\mathbf{S}}$ as the block diagonal matrix $\tilde{\mathbf{S}}_{block}$, that the appropriate block sizes ones should keep maximum block size of $q_k = \lfloor \min(n_1 - 1, n_2 - 1)/5 \rfloor$, $\forall k$, $k = 1,2,\ldots,m$, when unequal sample size cases and $q_k = \lfloor \min(n_1, n_2)/4 \rfloor$, $\forall k$, $k = 1,2,\ldots,m$, when equal sample size cases, where "$\lfloor a \rfloor$" denotes the floor function of constant a.

Simulation Study

In this section, the performance of the proposed test statistic T was evaluated through a simulation study and also was compared with those of the three tests mentioned in section 1 as: T_{BS}, T_{CQ} and T_{SKK}. We performed a Monte Carlo simulation by R program version 3.5.1. The two–sample dataset are generated from the p–dimensional multivariate normal distribution with the mean vector $\boldsymbol{\mu}_i$ and the positive definite covariance matrix $\boldsymbol{\Sigma}_i$ with size n_i for $i = 1,2$ by "MASS" package version 7.3–51.1.

We set initial value of random–number seed as $2^{31} - 1$ and then repeatedly computed testing statistics of the proposed test along with three comparative tests and counted the number of rejection under the null hypothesis and the number of rejection under the alternative hypothesis 10,000 times in each of five–pair forms of population covariance matrices structures. In each of five–pair forms of population covariance matrices structures, the attained significance level $(\hat{\alpha})$ and the attained power $\left(\widehat{1-\beta}\right)$ respectively, are computed by:

$$\hat{\alpha} = \frac{\text{the number of rejection under } H_0}{10,000} \tag{32}$$

$$\widehat{1-\beta} = \frac{\text{the number of rejection under } H_1}{10,000} \tag{33}$$

Parameter Set up

For the null hypothesis, we set the mean vectors as $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ and we choose $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = \begin{bmatrix} u_1 & u_2 & \cdots & u_p \end{bmatrix}'$ where $u_{2k-1} = 0$, $u_{2k} \overset{iid}{\sim} U(-0.5, 0.5)$, $k = 1,2,\ldots,p/2$, for the alternative, when $U(a,b)$ denotes uniform distribution with the support (a,b). The five–pair forms of the population covariance matrices $\left(\boldsymbol{\Sigma}_{1j}, \boldsymbol{\Sigma}_{2j}\right)$, $j = 1,2,\ldots,5$, were considered in three characteristics as:

1. The diagonal matrix: $\boldsymbol{\Sigma}_{11} = \mathbf{K}, \boldsymbol{\Sigma}_{21} = \boldsymbol{\Psi}$
2. The population covariance matrix with a common block size $q \times q$ as:
   - 2.1 $\boldsymbol{\Sigma}_{12} = \mathbf{K}^{1/2}\mathfrak{R}_{0.2}\mathbf{K}^{1/2}$ and $\boldsymbol{\Sigma}_{22} = \boldsymbol{\Psi}^{1/2}\mathfrak{R}_{0.4}\boldsymbol{\Psi}^{1/2}$ when correlation around ± 0.2 and ± 0.4
   - 2.2 $\boldsymbol{\Sigma}_{13} = \mathbf{K}^{1/2}\mathfrak{R}_{0.6}\mathbf{K}^{1/2}$ and $\boldsymbol{\Sigma}_{23} = \boldsymbol{\Psi}^{1/2}\mathfrak{R}_{0.8}\boldsymbol{\Psi}^{1/2}$ when correlation around ± 0.6 and ± 0.8
   - 2.3 $\boldsymbol{\Sigma}_{14} = \mathbf{K}^{1/2}\mathfrak{R}_{0.9}\mathbf{K}^{1/2}$ and $\boldsymbol{\Sigma}_{24} = \boldsymbol{\Psi}^{1/2}\mathfrak{R}_{0.95}\boldsymbol{\Psi}^{1/2}$ when correlation around ± 0.9 and ± 0.95
3. The population covariance matrix with mixed block sizes, that is, $\boldsymbol{\Sigma}_{15} = \mathbf{K}^{1/2}\mathfrak{I}_1\mathbf{K}^{1/2}$ and $\boldsymbol{\Sigma}_{25} = \boldsymbol{\Psi}^{1/2}\mathfrak{I}_2\boldsymbol{\Psi}^{1/2}$

where, $\mathbf{K} = \text{diag}(\kappa_1, \kappa_2, \ldots, \kappa_p)$ is a $p \times p$ diagonal matrix with $\kappa_i = 2 + (p - i + 1)/p$; $\boldsymbol{\Psi} = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$ is also a $p \times p$ diagonal matrix with $\psi_i = 4 + (p - i + 1)/p$, $i = 1,2,\ldots,p$ and $\mathfrak{R}_t = \text{diag}(\mathfrak{R}_{t1}, \mathfrak{R}_{t2}, \ldots, \mathfrak{R}_{tm})$ is a $p \times p$ block diagonal matrix where $t = 0.2, 0.4, 0.6, 0.8, 0.9, 0.95$ and $\mathfrak{R}_{tk} = (r_{ij})$, $r_{ii} = 1$, $r_{ij} = (-1)^{i+j}(t)^{|i-j|^{0.1}}$, $i, j = 1,2,\ldots,q_k$, $i \ne j$ when $k = 1,2,\ldots,m-1$ are of dimension q and the last blocks is q_m, where $p = q(m-1) + q_m$. Lastly, $\mathfrak{I}_1$ and $\mathfrak{I}_2$ are $p \times p$ block diagonal matrix where $\mathfrak{I}_1$ construct from

mixed of $\mathfrak{R}_{0.2k}$, $\mathfrak{R}_{0.6k}$, $\mathfrak{R}_{0.9k}$ and $\mathfrak{I}_2$ construct from mixed of $\mathfrak{R}_{0.4k}$, $\mathfrak{R}_{0.8k}$ and $\mathfrak{R}_{0.95k}$. Both of the block diagonal matrix $\mathfrak{I}_1$ and $\mathfrak{I}_2$ have different block sizes, various number of block size and these blocks are randomly located on the diagonal. The third characteristics of population covariance matrix was set up to be consistent with the natural of the data or observations.

The simulations study was conducted on both equal and unequal sample size, totally 20 situations in each table. The proposed test statistic T along with comparative tests were computed for the common block size $q = 1$ when the forms of population covariance matrix are diagonal matrix (Table 1) and for the common block size $q = \lfloor \min(n_1,n_2)/4 \rfloor$ and $q = \lfloor \min(n_1 - 1,n_2 - 1)/5 \rfloor$ when the forms of population covariance matrix are block diagonal matrix with equal and unequal sample size respectively (Table 2 to 4). Finally, the set of block sizes $(q_1,q_2,\ldots,q_m)$ was set up for the block diagonal matrix $\mathfrak{I}_1$ and $\mathfrak{I}_2$ when the forms of population covariance matrix are block diagonal matrix with mixed block sizes with randomly located (Table 5). All of these works, we set up the nominal significance level as 0.05.

Simulation Results

From Table 1 to 5, we showed the attained significance level and the attained power of these four tests T_{BS}, T_{CQ}, T_{SKK} and T in totally 20 different situations set up as above. The attained significance level values which is closest to the nominal significance level 0.05 in each row in each table are shown in bold and also the last row of each table provides the Average Absolute Discrepancy (*AAD*) between the nominal significance level and the estimated attained significance over that 10 conditions computed by $AAD = \sum |\hat{\alpha} - 0.05| / 10$ (Yanagihara and Yuan, 2005), a smaller *AAD* value indicates better overall performance of the other competing tests in 10 situations of maintaining the nominal significance level.

For overall situations considered both equal and unequal sample size, it was shown that the proposed test T gave the attained significance level values close to the nominal level setting $\alpha = 0.05$ consistently more than any other three tests considered with smallest average absolute discrepancy in all situations studied. It also gave the best the attained powers when the dimension is larger than or equal to 200 ($p \geq 200$) in all cases considered.

For three comparative tests T_{BS}, T_{CQ} and T_{SKK}, they did not give consistently in the attained significance level values and most of those values are not close to the nominal level setting under conditions and situation considered. Thus, we will not consider these three comparative tests further.

Table 1: Attained significance levels and attained powers when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_{21}$ at nominal significance level $\alpha = 0.05$

		Attained significance levels				Attained powers			
p	n_1,n_2	T_{BS}	T_{CQ}	T_{SKK}	T	T_{BS}	T_{CQ}	T_{SKK}	T
$q = 1$									
60	20,20	0.0584	0.0584	0.0878	**0.0544**	0.1829	0.1830	0.2298	0.1700
100	20,20	0.0591	0.0591	0.0955	**0.0562**	0.2365	0.2365	0.3210	0.2274
	40,40	0.0604	0.0603	0.0717	**0.0591**	0.5174	0.5173	0.5491	0.5109
200	20,20	0.0538	0.0541	0.1087	**0.0499**	0.3243	0.3236	0.4525	0.3016
	40,40	0.0548	0.0547	0.0748	**0.0542**	0.7173	0.7166	0.7549	0.7055
	60,60	0.0549	0.0548	0.0665	**0.0544**	0.9348	0.9348	0.9430	0.9330
400	20,20	0.0532	0.0533	0.1379	**0.0500**	0.4683	0.4686	0.6523	0.4392
	40,40	**0.0506**	0.0507	0.0781	0.0490	0.9057	0.9055	0.9325	0.9002
	60,60	**0.0520**	0.0522	0.0681	0.0523	0.9948	0.9948	0.9962	0.9941
	80,80	0.0576	**0.0575**	0.0688	0.0586	0.9998	0.9998	1.0000	1.0000
AAD		0.0055	0.0055	0.0358	0.0040	–	–	–	–
$q = 1$									
60	26,31	0.0563	0.0563	0.0686	**0.0539**	0.2625	0.2620	0.2857	0.2510
100	26,31	0.0563	0.0558	0.0776	**0.0555**	0.3622	0.3616	0.4109	0.3515
	36,46	0.0562	0.0564	0.0650	**0.0550**	0.5533	0.5537	0.5771	0.5448
200	26,31	0.0579	0.0576	0.0875	**0.0550**	0.5077	0.5067	0.5776	0.4901
	46,51	0.0548	0.0547	0.0682	**0.0541**	0.8383	0.8385	0.8557	0.8333
	66,76	**0.0548**	0.0551	0.0619	0.0561	0.9778	0.9777	0.9801	0.9779
400	26,31	0.0552	0.0556	0.1028	**0.0549**	0.7155	0.7153	0.8023	0.7000
	46,51	0.0532	0.0530	0.0735	**0.0515**	0.9691	0.9691	0.9778	0.9677
	66,76	0.0569	0.0570	0.0714	**0.0568**	0.9994	0.9994	0.9995	0.9994
	86,106	0.0558	**0.0556**	0.0634	0.0575	1.0000	1.0000	1.0000	1.0000
AAD		0.0057	0.0057	0.0240	0.0050	–	–	–	–

Table 2: Attained significance levels and attained powers when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_{22}$ at nominal significance level $\alpha = 0.05$

		Attained significance levels				Attained powers			
p	n_1,n_2	T_{BS}	T_{CQ}	T_{SKK}	T	T_{BS}	T_{CQ}	T_{SKK}	T
$q = \lfloor \min(n_1,n_2)/4 \rfloor$									
60	20,20	0.0652	0.0646	0.0840	**0.0586**	0.1569	0.1565	0.1861	0.2052
100	20,20	0.0621	0.0620	0.0875	**0.0528**	0.1895	0.1891	0.2499	0.2731
	40,40	0.0640	0.0639	0.0653	**0.0557**	0.3626	0.3626	0.3648	0.6379
200	20,20	0.0592	0.0590	0.0965	**0.0485**	0.2695	0.2698	0.3599	0.3545
	40,40	0.0635	0.0635	0.0664	**0.0541**	0.5227	0.5226	0.5435	0.8216
	60,60	0.0648	0.0649	0.0640	**0.0551**	0.7485	0.7483	0.7400	0.9778
400	20,20	0.0559	0.0565	0.1158	**0.0478**	0.3778	0.3771	0.5236	0.5065
	40,40	0.0580	0.0580	0.0703	**0.0545**	0.7450	0.7441	0.7771	0.9626
	60,60	0.0608	0.0608	0.0640	**0.0528**	0.9331	0.9330	0.9363	0.9996
	80,80	0.0589	0.0589	0.0586	**0.0567**	0.9898	0.9898	0.9889	1.0000
AAD		0.0112	0.0112	0.0272	0.0044	–	–	–	–
$q = \lfloor \min(n_1 - 1,n_2 - 1)/5 \rfloor$									
60	26,31	0.0629	0.0629	0.0665	**0.0557**	0.2155	0.2163	0.2198	0.3160
100	26,31	0.0640	0.0639	0.0735	**0.0540**	0.2920	0.2920	0.3195	0.4446
	36,46	0.0581	0.0584	0.0610	**0.0539**	0.4243	0.4246	0.4305	0.6470
200	26,31	0.0589	0.0589	0.0769	**0.0553**	0.4123	0.4116	0.4660	0.5770
	46,51	0.0617	0.0617	0.0662	**0.0557**	0.6824	0.6826	0.6882	0.9294
	66,76	0.0609	0.0610	0.0582	**0.0559**	0.8891	0.8893	0.8816	0.9956
400	26,31	0.0588	0.0590	0.0906	**0.0523**	0.6123	0.6113	0.6909	0.8026
	46,51	0.0624	0.0623	0.0706	**0.0557**	0.8885	0.8883	0.9013	0.9931
	66,76	0.0598	0.0598	0.0624	**0.0542**	0.9877	0.9877	0.9881	1.0000
	86,106	0.0609	0.0608	0.0587	**0.0566**	0.9996	0.9996	0.9996	1.0000
AAD		0.0108	0.0109	0.0185	0.0049	–	–	–	–

Table 3: Attained significance levels and attained powers when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_{13}$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_{23}$ at nominal significance level $\alpha = 0.05$

		Attained significance levels				Attained powers			
p	n_1,n_2	T_{BS}	T_{CQ}	T_{SKK}	T	T_{BS}	T_{CQ}	T_{SKK}	T
$q = \lfloor \min(n_1,n_2)/4 \rfloor$									
60	20,20	0.0710	0.0708	0.0687	**0.0550**	0.1205	0.1209	0.1129	0.5116
100	20,20	0.0648	0.0646	0.0693	**0.0497**	0.1367	0.1365	0.1384	0.7118
	40,40	0.0686	0.0689	**0.0441**	0.0600	0.1920	0.1927	0.1357	0.9914
200	20,20	0.0635	0.0625	0.0736	**0.0492**	0.1748	0.1742	0.1936	0.8488
	40,40	0.0674	0.0674	0.0516	**0.0505**	0.2529	0.2533	0.2040	0.9996
	60,60	0.0695	0.0693	0.0424	**0.0554**	0.3282	0.3285	0.2358	1.0000
400	20,20	0.0585	0.0582	0.0784	**0.0469**	0.2323	0.2319	0.2772	0.9739
	40,40	0.0616	0.0615	**0.0502**	0.0516	0.3656	0.3652	0.3215	1.0000
	60,60	0.0627	0.0628	**0.0445**	0.0562	0.4972	0.4967	0.4064	1.0000
	80,80	0.0618	0.0617	0.0367	**0.0519**	0.6132	0.6132	0.4803	1.0000
AAD		0.0149	0.0148	0.0124	0.0035	–	–	–	–
$q = \lfloor \min(n_1 - 1,n_2 - 1)/5 \rfloor$									
60	26,31	0.0672	0.0672	0.0564	**0.0535**	0.1436	0.1437	0.1204	0.7447
100	26,31	0.0654	0.0654	0.0615	**0.0565**	0.1821	0.1816	0.1663	0.9202
	36,46	0.0668	0.0670	**0.0505**	0.0580	0.2340	0.2332	0.1881	0.9903
200	26,31	0.0625	0.0623	0.0613	**0.0547**	0.2499	0.2495	0.2471	0.9795
	46,51	0.0672	0.0672	**0.0516**	0.0555	0.3460	0.3466	0.2933	1.0000
	66,76	0.0642	0.0641	0.0433	**0.0546**	0.4625	0.4625	0.3574	1.0000
400	26,31	0.0607	0.0607	0.0692	**0.0506**	0.3655	0.3654	0.3853	0.9997
	46,51	0.0622	0.0621	**0.0505**	0.0529	0.5251	0.5244	0.4789	1.0000
	66,76	0.0629	0.0629	**0.0455**	0.0551	0.6888	0.6887	0.6106	1.0000
	86,106	0.0646	0.0645	0.0403	**0.0560**	0.8477	0.8479	0.7618	1.0000
AAD		0.0144	0.0143	0.0072	0.0047	–	–	–	–

Table 4: Attained significance levels and attained powers when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_{14}$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_{24}$ at nominal significance level $\alpha = 0.05$

		Attained significance levels				Attained powers			
p	n_1,n_2	T_{BS}	T_{CQ}	T_{SKK}	T	T_{BS}	T_{CQ}	T_{SKK}	T
$q = \lfloor \min(n_1,n_2)/4 \rfloor$									
60	20,20	0.0716	0.0714	0.0568	**0.0563**	0.1082	0.1076	0.0858	0.9998
100	20,20	0.0646	0.0644	0.0593	**0.0538**	0.1197	0.1202	0.1046	1.0000
	40,40	0.0682	0.0683	0.0325	**0.0531**	0.1558	0.1556	0.0812	1.0000
200	20,20	0.0630	0.0630	0.0604	**0.0497**	0.1493	0.1494	0.1423	1.0000
	40,40	0.0665	0.0663	0.0393	**0.0520**	0.1960	0.1960	0.1298	1.0000
	60,60	0.0697	0.0697	0.0310	**0.0547**	0.2424	0.2425	0.1273	1.0000
400	20,20	0.0589	0.0593	0.0671	**0.0458**	0.1882	0.1877	0.1961	1.0000
	40,40	0.0619	0.0618	0.0416	**0.0470**	0.2709	0.2710	0.2003	1.0000
	60,60	0.0626	0.0629	0.0326	**0.0536**	0.3525	0.3523	0.2287	1.0000
	80,80	0.0603	0.0603	0.0248	**0.0548**	0.4245	0.4247	0.2431	1.0000
AAD		0.0147	0.0147	0.0142	0.0036	–	–	–	–
$q = \lfloor \min(n_1 - 1,n_2 - 1)/5 \rfloor$									
60	26,31	0.0672	0.0672	**0.0477**	0.0539	0.1243	0.1241	0.0879	1.0000
100	26,31	0.0671	0.0667	**0.0511**	0.0556	0.1488	0.1488	0.1180	1.0000
	36,46	0.0672	0.0669	0.0398	**0.0580**	0.1849	0.1850	0.1258	1.0000
200	26,31	0.0600	0.0599	**0.0540**	0.0546	0.2033	0.2027	0.1750	1.0000
	46,51	0.0676	0.0674	0.0410	**0.0558**	0.2604	0.2600	0.1792	1.0000
	66,76	0.0638	0.0637	0.0314	**0.0571**	0.3276	0.3276	0.1941	1.0000
400	26,31	0.0605	0.0609	0.0584	**0.0534**	0.2888	0.2892	0.2724	1.0000
	46,51	0.0615	0.0615	0.0421	**0.0578**	0.3877	0.3877	0.3012	1.0000
	66,76	0.0649	0.0648	0.0350	**0.0507**	0.5010	0.5007	0.3669	1.0000
	86,106	0.0640	0.0638	0.0276	**0.0558**	0.6306	0.6303	0.4466	1.0000
AAD		0.0144	0.0143	0.0099	0.0053	–	–	–	–

Table 5: Attained significance levels and attained powers when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_{15}$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_{25}$ (different block sizes), at nominal significance level $\alpha = 0.05$

		Attained significance levels				Attained powers			
p	n_1,n_2	T_{BS}	T_{CQ}	T_{SKK}	T	T_{BS}	T_{CQ}	T_{SKK}	T
60	20,20	0.0697	0.0690	0.0653	**0.0472**	0.1246	0.1245	0.1181	0.8475
100	20,20	0.0651	0.0654	0.0711	**0.0505**	0.1433	0.1433	0.1495	0.9967
	40,40	0.0735	0.0738	0.0548	**0.0509**	0.2035	0.2027	0.1535	1.0000
200	20,20	0.0617	0.0611	0.0743	**0.0456**	0.1801	0.1798	0.1992	0.9944
	40,40	0.0684	0.0682	0.0515	**0.0491**	0.2696	0.2696	0.2202	1.0000
	60,60	0.0714	0.0714	0.0453	**0.0478**	0.3568	0.3573	0.2641	1.0000
400	20,20	0.0604	0.0609	0.0789	**0.0469**	0.2379	0.2372	0.2845	1.0000
	40,40	0.0623	0.0623	**0.0530**	0.0467	0.3808	0.3811	0.3345	1.0000
	60,60	0.0637	0.0636	0.0456	**0.0507**	0.5002	0.5003	0.4219	1.0000
	80,80	0.0635	0.0633	0.0396	**0.0500**	0.6448	0.6448	0.5203	1.0000
AAD		0.0160	0.0159	0.0118	0.0019	–	–	–	–
60	26,31	0.0667	0.0669	0.0580	**0.0506**	0.1546	0.1544	0.1334	0.9396
100	26,31	0.0650	0.0645	0.0619	**0.0533**	0.1968	0.1962	0.1855	0.9957
	36,46	0.0666	0.0663	0.0545	**0.0539**	0.2392	0.2392	0.2032	1.0000
200	26,31	0.0597	0.0596	0.0628	**0.0498**	0.2612	0.2617	0.2677	1.0000
	46,51	0.0655	0.0656	0.0519	**0.0510**	0.3803	0.3807	0.3344	1.0000
	66,76	0.0647	0.0648	0.0445	**0.0545**	0.4832	0.4832	0.3828	1.0000
400	26,31	0.0605	0.0601	0.0673	**0.0486**	0.3894	0.3888	0.4041	1.0000
	46,51	0.0643	0.0643	0.0541	**0.0525**	0.5288	0.5289	0.4873	1.0000
	66,76	0.0646	0.0645	0.0453	**0.0498**	0.7004	0.7006	0.6091	1.0000
	86,106	0.0640	0.0639	0.0433	**0.0552**	0.8706	0.8706	0.7937	1.0000
AAD		0.0142	0.0141	0.0077	0.0023	–	–	–	–

We recommended to use the proposed test T when the population covariance matrix are diagonal matrix with $p \geq 200$ and both $n_1, n_2 \geq 40$ (shown in Table 1). When the population covariance matrices have block diagonal matrix structure, we recommended to use the proposed test T for $p \geq 200$ and both $n_1, n_2 \geq 40$ as well (shown in Table 2–5).

In addition, it is obvious that the attained powers the proposed test T when the dimension (p) increased for a given sample size or vice versa. It is still true when correlations among the variables in each sample are higher.

A Real Data Example

In this section, we applied the proposed test statistic using the prostate cancer data that collects data from DNA microarray technology. The data were retrieved on November 5, 2018 from "`spls`" package version 2.2–2 in R program. This data contain 6,033 genes for 102 subjects, 50 of which are non–tumor prostate and 52 of which are prostate tumors (Dettling and Bühlmann, 2002). A selection of 1,000 genes (p) was used to test the mean vectors of two independent sample, non–tumor prostate and prostate tumors, so $n_1 = 50$ and $n_2 = 52$.

Before computing 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), we obtain $T^* = 4.338$ with corresponding p–value < 0.01 which leads to the rejection of the null hypothesis of the equality covariance matrices.

To compute the proposed test statistic T, we determine common block size of the sample covariance matrix $\tilde{\mathbf{S}}$ is $q = \lfloor \min(50-1, 52-1)/5 \rfloor = 9$. Therefore, the first 111 blocks have dimension 9 and the last block has dimension 1. The test results are shown in Table 6 which test statistic has p–values less than 0.001 leading to the rejection of the null hypothesis of no difference between the two mean vectors, i.e., the gene expression levels of non–tumor prostate are significantly different from those of prostate tumors at the 0.05 level of significance. The computing results appeared below.

Table 6: Testing the equality of the gene expression level between non–tumor prostate and prostate tumors

Test statistic	T
Test Statistic value	60.161
p–value	<0.001
Computational Time	0.17 seconds

Conclusion

In this study, we developed and proposed a new approximate test statistic for testing the equality of mean vectors from two multivariate normal distributions when the covariance matrices are unknown and unequal in high–dimensional data. The main motivation of our proposed test is to avoid $\tilde{\mathbf{S}}^{-1}$, which is not exist, from T^2 test, we replaced the sample covariance matrix $\tilde{\mathbf{S}}$ with the block diagonal sample covariance matrix $\tilde{\mathbf{S}}_{block}$. Under the null hypothesis, the asymptotic distribution of a proposed test statistic converges to a standard normal distribution when the dimension of data approach infinity, or $p \to \infty$ and the sample covariance matrices $\tilde{\mathbf{S}}$ can be arranged to block diagonal matrix structure. Our proposed test are available when both sample sizes n_1 and n_2 are greater than or equal to 8, or $\min(n_1, n_2) \geq 8$. One interesting result of our proposed test that is invariant under scalar transformations and location shifts. Simulation results indicate that our proposed test performs the best and becomes more powerful when the dimension increases for a given sample size or vice versa or correlation among the variables in each sample are trend to higher.

Acknowledgment

The authors wishes to thank Mr.Suwatthana Chongtub, Mr.Tananop Limsuwanroj, Miss Rungruang Sithongkhum, Miss Pamita Vongprasert, Miss Raweewan Boonpa and Mr.Thatchakorn Tiwaphanuchai, for providing excellent running a simulation study assistance throughout this research.

Author's Contributions

Paranut Sukcharoen: Literary review, proof theorem, computer programming, analyzing the real data example, reporting and drafted the manuscript.

Samruam Chongcharoen: Participating in all steps in conducting research, contribution to the writing of the manuscript and approved the final manuscript.

Ethics

The authors declare that there is no conflict interest regarding the publication of this manuscript.

References

Bai, Z. and H. Saranadasa, 1996. Effect of high dimension: by an example of a two sample problem. Stat. Sinica, 6: 311–329. https://pdfs.semanticscholar.org/b100/71daa6844ea85d4c49d84bea8e2103c2807b.pdf

Chaipitak, S. and S. Chongcharoen, 2013. A test for testing the equality of two covariance matrices for high–dimensional data. J. Applied Sci., 13: 270–277. DOI: 10.3923/jas.2013.270.277

Chen, S.X. and Y.L. Qin, 2010. A two–sample test for high–dimensional data with applications to gene–set testing. Ann. Statist., 38: 808–835. DOI: 10.1214/09-AOS716

Chongcharoen, S., 2011. Inversion of covariance matrix for high dimension data. J. Math. Stat., 7: 227–229. DOI: 10.3844/jmssp.2011.227.229

Dettling, M. and P. Bühlmann, 2002. Supervised clustering of genes. Genome Biol., DOI: 10.1186/gb-2002-3-12-research0069

Gregory, K.B., R.J. Carroll, V. Baladandayuthapani and S.N. Lahiri, 2015. A two–sample test for equality of means in high dimension. J. Am. Stat. Assoc., 110: 837–849. DOI: 10.1080/01621459.2014.934826

Hu, J., Z. Bai, C. Wang and W. Wang, 2017. On testing the equality of high dimensional mean vectors with unequal covariance matrices. Annals Inst. Stat. Math., 69: 365–387. DOI: 10.1007/s10463-015-0543-8

James, G.S., 1954. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. Biometrika, 41: 19–43. DOI: 10.2307/2333003

Jiamwattanapong, K. and S. Chongcharoen, 2015. A new test for the mean vector in high–dimensional data. Songklanakarin J. Sci. Technol., 37: 477–484. http://rdo.psu.ac.th/sjstweb/journal/37–4/37–4–13.pdf

Jiamwattanapong, K. and S. Chongcharoen, 2017. A two–sample test for mean vectors in high–dimensional data. Applied Sci. Innovative Res., 1: 118–130. DOI: 10.22158/asir.v1n2p118

Johansen, S., 1980. The welch–James approximation to the distribution of the residual sum of squares in a weighted linear regression. Biometrika, 67: 85–92. DOI: 10.2307/2335320

Katayama, S. and Y. Kano, 2014. A new test on high–dimensional mean vector without any assumption on population covariance matrix. Commun. Stat. Theory Meth., 43: 5290–5304. DOI: 10.1080/03610926.2012.717663

Kawasaki, T. and T. Seo, 2015. A two sample test for mean vectors with unequal covariance matrices. Commun. Stat. Simulat. Comput., 44: 1850–1866. DOI: 10.1080/03610918.2013.824587

Krishnamoorthy, K. and Y. Xia, 2006. On selecting tests for equality of two normal mean vectors. Multivariate Behav. Res., 41: 533–548. DOI: 10.1207/s15327906mbr4104_5

Krishnamoorthy, K. and J. Yu, 2004. Modified Nel and Van der Merwe test for the multivariate Behrens–Fisher problem. Stat. Probability Lett., 66: 161-169. DOI: 10.1016/j.spl.2003.10.012

Nel, D.G. and C.A. van der Merwe, 1986. A solution to the multivariate Behrens–fisher problem. Commun. Stat. Theory Meth., 15: 3719–3735. DOI: 10.1080/03610928608829342

Nel, D.G., C.A. van der Merwe and B.K. Moser, 1990. The exact distributions of the univariate and multivariate Behrens–fisher statistics with a comparison of several solutions in the univariate case. Commun. Stat. Theory Meth., 19: 279–298. DOI: 10.1080/03610929008830200

Nishiyama, T., M. Hyodo, T. Seo and T. Pavlenko, 2013. Testing linear hypotheses of mean vectors for high–dimension data with unequal covariance matrices. J. Stat. Plann. Inference, 143: 1898–1911. DOI: 10.1016/j.jspi.2013.07.008

Richard, A.J. and W.W. Dean, 2014. Applied Multivariate Statistical Analysis. 6 Edn., Prentice–Hall, London, ISBN-10: 1292024941, pp: 294.

Srivastava, M.S., 2002. Methods of Multivariate Statistics. 1st Edn., Wiley-Interscience, New York, ISBN-10: 0471223816, pp: 119.

Srivastava, M.S., 2009. A test for the mean vector with fewer observations than the dimension under non–normality. J. Multivariate Anal., 100: 518–532. DOI: 10.1016/j.jmva.2008.06.006

Srivastava, M.S., S. Katayama and Y. Kano, 2013. A two sample test in high dimensional data. J. Multivariate Anal., 114: 349–358. DOI: 10.1016/j.jmva.2012.08.014

Yamada, T. and T. Himeno, 2015. Testing homogeneity of mean vectors under heteroscedasticity in high–dimension. J. Multivariate Anal., 139: 7–27. DOI: 10.1016/j.jmva.2015.02.005

Yanagihara, H. and K.H. Yuan, 2005. Three approximate solutions to the multivariate Behrens–fisher problem. Commun. Stat. Simulat. Comput., 34: 975–988. DOI: 10.1080/03610910500308396

Yao, J., S. Zheng and Z. Bai, 2015. Large Sample Covariance Matrices and High–Dimensional Data Analysis. 1st Edn., Cambridge University Press, New York, ISBN-13: 9781107065178.

Yao, Y., 1965. An approximate degrees of freedom solution to the multivariate Behrens fisher problem. Biometrika, 52: 139–147. DOI: 10.2307/2333819

Zhang, J. and J. Xu, 2009. On the k–sample Behrens–Fisher problem for high–dimensional data. Sci. China Series A: Math., 52: 1285–1304. DOI: 10.1007/s11425-009-0091-x

Zhou, B., 2016. Linear hypothesis testing for high–dimensional data under heteroscedasticity. PhD Thesis, National University of Singapore, Singapore.

Zhou, B., J. Guo and J.T. Zhang, 2017. High–dimensional general linear hypothesis testing under heteroscedasticity. J. Stat. Plann. Inference, 188: 36–54. DOI: 10.1016/j.jspi.2017.03.005