Songklanakarin J. Sci. Technol.
44 (1), 142-148, Jan. - Feb. 2022

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

*Original Article*

# One-sided multivariate tests for high-dimensional data from two populations with unknown and unequal covariance matrices

Samruam Chongcharoen, Pawat Paksaranuwat*, and Manad Khamkong

*Department of Statistics, Faculty of Science, Chiang Mai University,*
*Mueang, Chiang Mai, 50200 Thailand*

Received: 30 June 2020; Revised: 29 March 2021; Accepted: 5 April 2021

## Abstract

In this paper, we propose a new statistic for testing a one-sided hypothesis of mean vectors from two multivariate normal populations when the covariance matrices are unknown and unequal for high-dimensional data. As we know that the sample covariance matrix is singular for high-dimensional data, the proposed test is based on the idea of keeping as much information as possible from the sample covariance matrices. The performance of the proposed test is assessed in a simulation study with varied situations. The simulation results show that the proposed test was satisfactory in attaining nominal significance values close to set levels and the attained test power was excellent in every situation considered. Finally, the efficacy of the proposed test is 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

## 1. Introduction

High-dimensional data are increasingly encountered in statistical applications in many areas, mostly in biology and finance, where the dimension is a lot larger than the sample size. When this happens, classical multivariate statistical procedures cannot be applied because they involve the inverse of sample covariance matrix, which does not exist in such high-dimensional case. In a one-sample case, Chongcharoen (2012) studied one-sided multivariate tests for high-dimensional data with unknown covariance matrix by combining Dempster's high-dimensional tests (1958, 1960) and the one-sample versions of Bai and Saranadasa (1996) and Srivastata and Du (2008) based on Follmann's test (1996). These tests do not need the inverse of the sample covariance matrix, but 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. $p/n \to c \in (0,\infty)$. However, in practice, there are many datasets that have a dimension much larger than the sample size (*p* > *n*) (Park & Ayyala, 2013). In this study, we have extended Chongcharoen's tests (2012) to cases with two independent samples and propose one-sided multivariate tests of two independent samples with unknown and unequal covariance matrices for high-dimensional data based on the idea of keeping as much information as possible from the two sample covariance matrices, by using a submatrix of the covariance matrix instead of using the diagonal of the covariance matrix. The real-life situation in which the one-sided alternative makes sense occurs when one uses a matched-pair design to compare the multivariate responses of two treatments, and one tests for difference of these two mean responses to compared treatments. The one-sided alternative may be of interest if one believes that for each coordinate, the mean response to treatment one is at least as large as that to treatment two.

## 2. Materials and Methods

Throughout this paper, we suppose that $X_{i1}, X_{i2}, \ldots, X_{in_i}$ ; $i = 1, 2$ are independent random samples from *p*-dimensional multivariate normal distributions with

*Corresponding author
Email address: pawat.pak@cmu.ac.th

unknown mean vectors $\boldsymbol{\mu}_i$ and unknown unequal positive definite covariance matrices $\boldsymbol{\Sigma}_i, X_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. We consider the comparison tests for two independent multivariate means of two high-dimensional data sets on sample populations with an alternative one-sided test. That is to say, we want to test the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against $H_1 : \boldsymbol{\mu}_1 > \boldsymbol{\mu}_2$ when both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are unknown and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. The data at hand have larger dimensionality than the number of observations. That is, $p > n_1, n_2$.

Originally, the unrestricted test for testing $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against $H_1 : \boldsymbol{\mu}_1 > \boldsymbol{\mu}_2$ when two independent random samples $X_{i1}, X_{i2}, ..., X_{in_i}$; $i = 1, 2$ are drawn from two independent *p*-dimensional multivariate normal distributions with unknown mean vectors $\boldsymbol{\mu}_i$ and unknown unequal positive definite covariance matrices (called the Behrens–Fisher problem) with $p \leq n_1 + n_2 - 2$ is

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

where the sample mean vectors $\left(\bar{\boldsymbol{X}}_i\right)$ and the sample covariance matrices $\left(\boldsymbol{S}_i\right)$ are respectively defined as

$$\bar{\boldsymbol{X}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \text{ and } \boldsymbol{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{\boldsymbol{X}}_i\right)\left(X_{ij} - \bar{\boldsymbol{X}}_i\right)'; \; i = 1, 2. \qquad (2)$$

The exact distribution of $T^2$ under the null hypothesis is obtained from Nel, van der Merwe and Moser (1990). However, the exact distribution is very complicated and computationally intractable and thus is of limited value for practical applications. However, an approximation of this test statistic is used when both sample sizes $n_1$ and $n_2$ approach infinity, and the distribution of $T^2$ converges to Chi-squared with *p* degrees of freedom. This approximation is very simple and easy to compute but suffers if either sample size $n_1$ or $n_2$ is small, while being more accurate when $\min(n_1, n_2) \to \infty$, (Srivastava, 2002; Yanagihara & Yuan, 2005). Unfortunately, in practice the sample size is often not very large, so this approximation is not recommended for practical applications.

There is a vast amount of literature devoted to the solution of this problem, and many researchers have tried to approximate the distribution of $T^2$ by a constant times *F*–distribution with numerator degrees of freedom *p* and an approximated denominator degrees of freedom depending on sample size, mean, and covariance matrix: the approximate solutions based on $T^2$ in James (1954), Yao (1965), Johansen (1980) and Yanagihara and Yuan (2005) are invariant whereas the solution by Nel and Van Der Merwe (1986) is not. Later, Krishnamoorthy and Yu (2004) modified the solution in Nel and Van Der Merwe (1986) by providing an invariant test statistic, and Kawasaki and Seo (2016) improved the solution in Yanagihara and Yuan (2005) by asymptotic expansions. From the literary review, we found that the solution in Yanagihara and Yuan (2005) was satisfactorily close to attaining the nominal significance level. Krishnamoorthy and Xia (2006), among others, showed via intensive simulation studies that this test performed the best among the approximate solutions to the multivariate Behrens–Fisher problem.

An approximation to the distribution of $T^2$ by an *F*–distribution is given by Krishnamoorthy and Yu (2004) as follows:

$$T^2 \sim \frac{vp}{v - p + 1} F_{p, v-p+1}, \qquad (3)$$

where $F_{p, v-p+1}$ denotes an *F*–distributed random variable with *p* and *v* – *p* + 1 degrees of freedom; the *v* degrees of freedom are estimated from the sample covariance matrices using the relationship

$$v = \frac{p + p^2}{\displaystyle\sum_{i=1}^{2}\frac{1}{n_i - 1}\left\{\operatorname{tr}\left[\left(\frac{\mathbf{S}_i}{n_i}\left(\frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2}\right)^{-1}\right)^2\right] + \left(\operatorname{tr}\left[\frac{\mathbf{S}_i}{n_i}\left(\frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2}\right)^{-1}\right]\right)^2\right\}} \qquad (4)$$

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

In high-dimensional data where a single population has the number of variables exceeding the sample size $p > n_i$, the sample covariance matrix $\mathbf{S}$ loses its full rank and is singular, which ensures that $\mathbf{S}_i$ does not have an inverse (Chongcharoen, 2011). Furthermore, for two populations where the number of variables is larger than the sum of the two sample sizes minus 2, $p > n_1 + n_2 - 2,$ the sample covariance matrix $\tilde{\mathbf{S}}$ is defined by

$$\tilde{\mathbf{S}} = \frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2} \qquad (5)$$

and does not have an inverse. So, the test statistic $T^2$ in Equation (1) cannot be applied to high-dimensional data.

To overcome the problem, namely needing the inverse of the sample covariance matrix that is singular for high-dimensional data, many recent efforts have been devoted to construct new tests for the multivariate Behrens–Fisher problem in high-dimensional data. The majority have tried to avoid the use of $\tilde{\mathbf{S}}^{-1}$ by, for instance, replacing $\mathbf{S}_i$ with a diagonal matrix (Bai & Saranadasa, 1996), sidestepping the covariance matrix estimation (Chen and Qin, 2010), or using the diagonal matrix of the sample covariance matrix $\tilde{\mathbf{S}}$ and the trace of the sample correlation matrix (Srivastava, Katayama, and Kano, 2013) as examples among many others in the literature.

Based on the idea of keeping as much information of $\mathbf{S}_i$ as possible, Sukcharoen and Chongcharoen (2019) proposed a testing statistic to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ under the aforementioned conditions for high-dimensional data as

$$T = \frac{T_n - \sum_{k=1}^{m} \frac{v_k q_k}{v_k - q_k - 1}}{\sqrt{\sum_{k=1}^{m} \frac{2 q_k v_k^2 (v_k - 1)}{(v_k - q_k - 1)^2 (v_k - q_k - 3)}}} \rightarrow N(0,1)$$

where $T_n = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \, \tilde{\boldsymbol{S}}_{block}^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$; $\bar{X}_i \; i = 1, 2$, as defined in Equation (1), and

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

in which $\tilde{\mathbf{S}}_{kk} \;\; k = 1, 2, \ldots, m, \;\; m \leq p,$ are $q_k \times q_k$ submatrices on the diagonal of $\tilde{\mathbf{S}}$ with $q_k < n_1 + n_2 - 2$; $\sum_{k=1}^{m} q_k = p$; and $v_k$ is the approximate number of degrees of freedom in the $k$-th block, which can be obtained by

$$v_k = \frac{q_k + q_k^2}{\sum_{i=1}^{2} \frac{1}{n_i - 1} \left\{ \mathrm{tr}\left[ \left( \frac{\mathbf{S}_{ikk}}{n_i} \left( \frac{\mathbf{S}_{1kk}}{n_1} + \frac{\mathbf{S}_{2kk}}{n_2} \right)^{-1} \right)^2 \right] + \left( \mathrm{tr}\left[ \frac{\mathbf{S}_{ikk}}{n_i} \left( \frac{\mathbf{S}_{1kk}}{n_1} + \frac{\mathbf{S}_{2kk}}{n_2} \right)^{-1} \right] \right)^2 \right\}}$$

$$= \frac{q_k + q_k^2}{\frac{\mathrm{tr}\left[ \left( \mathbf{S}_{1kk} \tilde{\mathbf{S}}_{kk}^{-1} \right)^2 \right] + \left[ \mathrm{tr}\left( \mathbf{S}_{1kk} \tilde{\mathbf{S}}_{kk}^{-1} \right) \right]^2}{n_1^2 (n_1 - 1)} + \frac{\mathrm{tr}\left[ \left( \mathbf{S}_{2kk} \tilde{\mathbf{S}}_{kk}^{-1} \right)^2 \right] + \left[ \mathrm{tr}\left( \mathbf{S}_{2kk} \tilde{\mathbf{S}}_{kk}^{-1} \right) \right]^2}{n_2^2 (n_2 - 1)}}$$

This statistic will reject $H_0$ at significance level $\alpha$ if observed $T \geq z_{1-\alpha}$, where $z_{1-\alpha}$ denotes the upper $1-\alpha$ quantile. Based on the idea of keeping as much information of $\mathbf{S}_i$ as possible, Sukcharoen and Chongcharoen (2019) proposed a testing statistic to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ under the aforementioned conditions for high-dimensional data as a standard normal distribution. They also showed that the proposed test statistic *T* is invariant under location shift and scaling transformation, $x_{ij} \rightarrow D x_{ij} + c, \; i = 1, 2; \; j = 1, 2, \ldots, n_i$, where *C* is a constant vector and *D* is a nonsingular *p x p* diagonal matrix. They also showed that their test works very well when both sample sizes $n_1$ and $n_2$ are larger than 8. Based on a simulation study and the idea of keeping as much information as possible from the sample covariance matrix $\tilde{\mathbf{S}}$, they also suggest that their proposed test can be used although there is no prior information to arrange variables from the sample covariance matrix $\tilde{\mathbf{S}}$ to the block diagonal matrix $\tilde{\mathbf{S}}_{block}$. For appropriate block sizes, they suggest to keep maximum block size $q_k = \lfloor \min(n_1, n_2)/4 \rfloor, \; \forall k, \; k = 1, 2, \ldots, m,$ for equal sample size cases, and $q_k = \lfloor \min(n_1 - 1, n_2 - 1)/5 \rfloor,$ $\forall k, \; k = 1, 2, \ldots, m,$ for unequal sample size cases, where $\lfloor a \rfloor$ denotes the floor function (i.e. rounding by deleting the decimal part) applied to $a$.

Motivated by this kind of data, prior literature, and Follmann's test, we combined the unrestricted alternative test for the high-dimensional multivariate tests mentioned previously with Follmann's idea, to propose a test statistic for testing the one-sided multivariate hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ for high-dimensional data as

$$T = \frac{T_n - \sum_{k=1}^{m} \frac{v_k q_k}{v_k - q_k - 1}}{\sqrt{\sum_{k=1}^{m} \frac{2 q_k v_k^2 (v_k - 1)}{(v_k - q_k - 1)^2 (v_k - q_k - 3)}}}$$ and $\sum_{i=1}^{p} (\bar{X}_{1i} - \bar{X}_{2i})$, with which $H_0$ will be rejected at significance level $\alpha$ if

$$T \geq z_{1-2\alpha} \text{ and } \sum_{i=1}^{p}\left(\bar{X}_{1i} - \bar{X}_{2i}\right) > 0 \cdot \qquad (7)$$

We obtain $1'\theta = 0$ and the significance level is approximated by

$$\begin{aligned} P\left(T \geq z_{1-2\alpha} \cap 1'\left(\bar{X}_1 - \bar{X}_2\right) > 0\right) &= P\left(T \geq z_{1-2\alpha}\right) P\left(1'\left(\bar{X}_1 - \bar{X}_2\right) > 0\right) \\ &= (2\alpha)\left(\frac{1}{2}\right) \\ &= \alpha, \end{aligned}$$

please see Theorem 2.1 in Follmann (1996).

One point of interest here is how large the block sizes in the block diagonal matrix $\tilde{\mathbf{S}}_{block}$ are: If both population covariance matrices $\Sigma_1$ and $\Sigma_2$ are block diagonal matrices with *m* blocks and of small sizes $q_1, q_2, \ldots, q_m$ that are known, then we should set $\tilde{\mathbf{S}}_{block}$ to be similar to population covariance matrices $\Sigma_1$ and $\Sigma_2$. However, the population covariance matrices in a real-life situation may not be block diagonal and we have no prior information. Since, theoretically, the proposed test statistic *T* is based on the solution to approximate the distribution of $T^2$ proposed by Krishnamoorthy and Yu (2004), we only require block sizes $q_k \leq v_k - 6,\ k = 1, 2, \ldots, m,$ as in the recommendations of Sukcharoen and Chongcharoen (2019), to attain a significance level very close to the nominal, provided that $p \leq \min\left(n_1 - 1, n_2 - 1\right)/5$ in unequal sample size cases; this condition is somewhat relaxed to $p \leq n/4$ in the equal sample size cases $n_1 = n_2 = n.$ Hence, based on these suggestions and the idea of keeping as much information as possible from the sample covariance matrix $\tilde{\mathbf{S}}$ when there is no prior information to arrange the variables to a block diagonal matrix $\tilde{\mathbf{S}}_{block}$, appropriate block sizes can be kept at the maximum $q_k$ by

$$q^* = \begin{cases} \left\lfloor \min\left(n_1 - 1, n_2 - 1\right)/5 \right\rfloor, & when\ n_1 \neq n_2, \\ \left\lfloor n/4 \right\rfloor, & when\ n_1 = n_2 = n. \end{cases}$$

Of course, ideally in each diagonal block of the matrix $\tilde{\mathbf{S}}_{block}$, we want to have a set of highly correlated variables.

## 3. Simulation Studies

The performance of the proposed test was evaluated using a simulation study with a variety of parameters settings for the mean vectors and the population covariance matrices. The mean vectors $\mu_1$ and $\mu_2$ were set as $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$, and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 + \boldsymbol{\upsilon}$ with $\boldsymbol{\mu} = \begin{bmatrix} u_1 & u_2 & \cdots & u_p \end{bmatrix}'$, $u_i \sim \mathrm{U}(-1.5, 1.5)$ and $\boldsymbol{\upsilon} = \begin{bmatrix} v_1 & v_2 & \cdots & v_p \end{bmatrix}'$, $v_i \sim \mathrm{U}(0.1, 0.5)$. The estimates of type I error rate and test power for the proposed test statistics for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ were evaluated by computing $\dfrac{\text{number of rejections}}{m}$, where *m* is the number of iterations of the datasets simulated under the null hypothesis or the alternative hypothesis. The population covariance matrices $\Sigma_i$, for $i = 1, 2$, were set up as one of the following**:**

Type A: Both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were in diagonal matrix form, $diag\left(\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{22}, \cdots, \mathbf{\Sigma}_{mm}\right)$, where $\mathbf{\Sigma}_{kk} = c_k I + \left(1 - c_k\right) J$, $c_k \sim U\left(0, 1\right)$, $J$ is a $q_k \times q_k$ matrix of 1's, and block size $q_k$ is the maximum $q^*$. So $\tilde{\mathbf{S}}_{block}$ is set in diagonal matrix form as $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ with the same subblock sizes of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.

Type B: Both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were also in diagonal matrix form as Type A but the number of blocks *m* and block sizes $q_1, q_2, \ldots, q_m$ are random. So $\tilde{\mathbf{S}}_{block}$ is also set in diagonal matrix with different numbers of blocks and block sizes as in $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.

Type C: Both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were not in diagonal matrix form as $\mathbf{\Sigma}_i = \boldsymbol{D}_\sigma^{1/2} \mathfrak{R} \boldsymbol{D}_\sigma^{1/2}$, where $\boldsymbol{D}_\sigma^{1/2} = diag\left(\sigma_1, \sigma_2, \ldots, \sigma_p\right)$, $\sigma_j = 2 + \left(-1\right)^{j-1}\left(p - j + 1\right)/p$ and $\mathfrak{R} = \left[r_{ij}\right]$, $r_{ij} = \left(-1\right)^{i+j}\left(c^{|i-j|}\right)$, for $c \sim U\left(0, 1\right)$. Thus $\tilde{\mathbf{S}}_{block}$ is set as block diagonal form in different pattern of block sizes.

We set the nominal significance level as $\alpha = 0.05$ for all of the simulation studies.The simulated data for each combination of designed means and designed covariance matrices were generated with dimension $p \in \{60, 100, 200, 400\}$ and for both equal and unequal sample sizes $(n_1, n_2)$ with each $p > n_1 + n_2 - 2$, then the proposed statistic for testing the equality of two population means was computed and the number of rejections counted; each procedure was repeated $m = 10,000$ times. The attained significance level and power of the proposed test are reported in Tables 1–3.

In Tables 1 and 2, both unknown and unequal covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were block diagonal with the same known pattern of correlations between the variables. This means we could set the block sizes in the block diagonal matrix $\tilde{\mathbf{S}}_{block}$ in the same pattern to match both covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Subsequently, all attained significance levels were close to the nominal significance level and all attained test powers were close to 1 in every situation considered. It can be seen that the proposed test gave a reasonable result.

In Table 3, both unknown and unequal covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were not block diagonal (as is likely to occur in a real-life situation). The proposed test achieved a significance level close to the nominal significance level and a test power close to 1 in every situation considered, so we can say that the proposed test still works very well although neither of the two unknown population covariance matrices is block diagonal.

Table 1. The attained significance level and power of the test with Type A covariance matrices

| $p$ | $(n_1, n_2)$ | $q_k$ | Type I Error | | Power of the Test |
|---|---|---|---|---|---|
| | | | $\mu_1 = \mu_2 = \underline{0}$ | $\mu_1 = \mu_2 = u$ | $\mu_2 = u$ and $\mu_1 = \mu_2 + v$ |
| 60 | (20,20) | 5 | 0.0486 | 0.0504 | 0.8684 |
| 100 | (20,20) | 5 | 0.0522 | 0.0460 | 0.9617 |
| | (26,36) | 5 | 0.0519 | 0.0579 | 0.9986 |
| | (40,40) | 10 | 0.0551 | 0.0493 | 0.9992 |
| 200 | (20,20) | 5 | 0.0521 | 0.0484 | 0.9991 |
| | (26,36) | 5 | 0.0496 | 0.0532 | 1.0000 |
| | (40,40) | 10 | 0.0531 | 0.0532 | 1.0000 |
| | (51,71) | 10 | 0.0510 | 0.0517 | 1.0000 |
| | (80,80) | 20 | 0.0503 | 0.0521 | 1.0000 |
| 400 | (20,20) | 5 | 0.0474 | 0.0455 | 1.0000 |
| | (26,36) | 5 | 0.0501 | 0.0527 | 1.0000 |
| | (40,40) | 10 | 0.0537 | 0.0525 | 1.0000 |
| | (51,71) | 10 | 0.0539 | 0.0507 | 1.0000 |
| | (80,80) | 20 | 0.0539 | 0.0559 | 1.0000 |
| | (101,141) | 20 | 0.0513 | 0.0532 | 1.0000 |
| | (160,160) | 40 | 0.0583 | 0.0540 | 1.0000 |

Table 2. The attained significance level and power of the test with Type B covariance matrices

| $p$ | $(n_1, n_2)$ | $q_k$ | Type I Error | | Power of the Test |
|---|---|---|---|---|---|
| | | | $\mu_1 = \mu_2 = \underline{0}$ | $\mu_1 = \mu_2 = u$ | $\mu_2 = u$ and $\mu_1 = \mu_2 + v$ |
| 60 | (20,20) | 5 | 0.0527 | 0.0512 | 0.9240 |
| 100 | (20,20) | 5 | 0.0509 | 0.0517 | 0.9890 |
| | (26,36) | 5 | 0.0538 | 0.0508 | 0.9998 |
| | (40,40) | 10 | 0.0541 | 0.0508 | 0.9998 |
| 200 | (20,20) | 5 | 0.0509 | 0.0513 | 0.9999 |
| | (26,36) | 5 | 0.0495 | 0.0462 | 1.0000 |
| | (40,40) | 10 | 0.0521 | 0.0492 | 1.0000 |
| | (51,71) | 10 | 0.0552 | 0.0511 | 1.0000 |
| | (80,80) | 20 | 0.0525 | 0.0568 | 1.0000 |
| 400 | (20,20) | 5 | 0.0457 | 0.0483 | 1.0000 |
| | (26,36) | 5 | 0.0511 | 0.0526 | 1.0000 |
| | (40,40) | 10 | 0.0535 | 0.0535 | 1.0000 |
| | (51,71) | 10 | 0.0530 | 0.0466 | 1.0000 |
| | (80,80) | 20 | 0.0520 | 0.0497 | 1.0000 |
| | (101,141) | 20 | 0.0497 | 0.0567 | 1.0000 |
| | (160,160) | 40 | 0.0498 | 0.0514 | 1.0000 |

Table 3. The attained significance level and power of the test with Type C covariance matrices

| $p$ | $(n_1, n_2)$ | $q_k$ | Type I Error | | Power of the Test |
|---|---|---|---|---|---|
| | | | $\mu_1 = \mu_2 = \underline{0}$ | $\mu_1 = \mu_2 = u$ | $\mu_2 = u$ and $\mu_1 = \mu_2 + v$ |
| 60 | (20,20) | 5 | 0.0551 | 0.0566 | 0.9007 |
| 100 | (20,20) | 5 | 0.0526 | 0.0551 | 0.9585 |
| | (26,36) | 5 | 0.0598 | 0.0548 | 0.9959 |
| | (40,40) | 10 | 0.0551 | 0.0536 | 0.9999 |
| 200 | (20,20) | 5 | 0.0542 | 0.0576 | 0.9936 |
| | (26,36) | 5 | 0.0582 | 0.0571 | 0.9999 |
| | (40,40) | 10 | 0.0516 | 0.0566 | 1.0000 |
| | (51,71) | 10 | 0.0540 | 0.0569 | 1.0000 |
| | (80,80) | 20 | 0.0543 | 0.0553 | 1.0000 |
| 400 | (20,20) | 5 | 0.0557 | 0.0553 | 1.0000 |
| | (26,36) | 5 | 0.0589 | 0.0534 | 1.0000 |
| | (40,40) | 10 | 0.0543 | 0.0498 | 1.0000 |
| | (51,71) | 10 | 0.0564 | 0.0566 | 1.0000 |
| | (80,80) | 20 | 0.0487 | 0.0540 | 1.0000 |
| | (101,141) | 20 | 0.0581 | 0.0565 | 1.0000 |
| | (160,160) | 40 | 0.0562 | 0.0535 | 1.0000 |

From these results, we recommend using the proposed test as a one-sided alternative multivariate test for high-dimensional data when two independent high-dimensional random samples are obtained from a *p*-dimensional multivariate normal distribution with an unknown mean vector and an unknown unequal positive definite covariance matrix. Table 3 indicates that although the arrangement of the group of variables is random, the proposed test is still sensitive in detecting the difference of the two means. In practice and ideally, we should arrange the variables so that high correlates are in the same block as much as possible, and Jiamwattanapong and Chongcharoen (2017) have recommend to arrange the variables in blocks in such a way that the correlation coefficient of any two adjacent variables in the same block is greater than or equal to 0.5.

## 4. An Example Using Real-Life Data

We used DNA microarray data from an oncology study to demonstrate the efficacy of the proposed test (Notterman *et al*., 2001). A selection of 100 genes (p) was used to test the mean vectors of two independent groups: tumor tissue and normal tissue. Each group has a sample size of 18, i.e. $n_1 = n_2 = 18$. For our example, both covariance matrices are assumed to be unequal, although the equality of covariance matrices can be tested by using the method presented in Chaipitak and Chongcharoen (2013). Suppose we want to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ where $\mu_1, \mu_2$ are the mean vectors of tumor tissue and normal tissue respectively. The two test statistics computed using Equation (7) are $T$ = 21.6046 and $\sum_{i=1}^{p}\left(\bar{X}_{1i} - \bar{X}_{2i}\right)$= 589.3333 which led to the rejection of the null hypothesis ($T > Z_{0.90}$ and $\sum_{i=1}^{p}\left(\bar{X}_{1i} - \bar{X}_{2i}\right) > 0$); i.e. for the same 100 gene expression levels, the tumor tissue mean is significantly greater than the normal tissue mean at the 0.05 significance level.

## 5. Conclusions

In this study, we developed and proposed a test statistic for hypothesis testing when two sampled high-dimensional multivariate normal distributions have covariance matrices that are unknown and unequal. The main motivation of our proposed test is to avoid need to invert the singular covariance matrix. Based on the test statistic by Sukcharoen and Chongcharoen (2019) for unrestricted alternative test of the high-dimensional data, we combined it with Follmann's idea to propose a test statistic for testing a one-sided multivariate hypothesis for high-dimensional data.

The results of a simulation study indicated that our proposed test had a good performance with a higher power when both of the unknown covariance matrices are block diagonal with the same known block pattern. Moreover, in general situations when the unknown and unequal covariance matrices are not block diagonal, our proposed test still gave a good performance under the conditions considered. Therefore, we recommend this test for one-sided multivariate testing to compare means in high dimensional data, as long as better alternative tests are not available.

## References

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

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.

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.

Chongcharoen, S. (2012). One-sided multivariate tests for high dimensional data. *Journal of Mathematics and Statistics*, *8*(2), 274–282.

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

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics*, 16(1), 41–50.

Follmann, D. (1996). A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association*, *91*(434), 854–861.

James, A. T. (1954). Normal multivariate analysis and the orthogonal group. *Annals of Mathematics and Statistics*, *25*(1), 40–75.

Jiamwattanapong, K. & Chongcharoen, S. (2017). A two-sample test for mean vectors in high-dimensional data. *Applied Science and Innovative Research*, *1*(2), 118-130.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, *67*(1), 85–92.

Kawasaki, T. & Seo, T. (2015). Bias correction for $T^2$ type statistic with two-step monotone missing data. *Statistics*, *50*(1), 76–88.

Krishnamoorthy, K., & Xia, Y. (2006). On selecting tests for equality of two normal mean vectors. *Multivariate Behavioral Research*, *41*(4), 533–548.

Krishnamoorthy, K., & Yu, J. (2004). Modified Nel and Van der Merwe test for the multivariate Behrens–Fisher problem. *Statistics and Probability Letters*, *66*(2), 161–169.

Nel, D. G., & Van der Merwe, C. A. (1986). A solution to the multivariate Behrens-Fisher problem. *Communications in Statistics-Theory and Methods*, *15*(12), 3719–3735.

Nel, D. G., Van der Merwe, C. A., & Moser, B. K. (1990). The exact distributions of the univariate and multivariate Behrens-Fisher statistics with a comparison of several solutions in the univariate case. *Communication in Statistics-Theory and Methods*, *19*(1), 279–298.

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*(7), 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.

Richard, A. J., & Dean, W. W. (2014). *Applied multivariate statistical analysis (6th ed.)*. London, England: Prentice–Hall.

Srivastava, M. S. (2002). *Methods of multivariate statistics*. New York, NY: Wiley–Interscience.

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.

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

Sukcharoen, P. & Chongcharoen, S. (2019). A test on the multivariate Behrens–Fisher problem in high–dimensional data by block covariance estimation. *Journal of Mathematics and Statistics*, *15*, 44-54.

Yanagihara, H., & Yuan, K. H. (2005). Three approximate solutions to the multivariate Behrens–Fisher problem. *Communications in Statistics-Simulation and Computation*, *34*(4), 975–988.

Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens Fisher problem. *Biometrika*, *52*(1/2), 139–147.