Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 243-257
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Dimension Reduction for Detecting a Difference in Two High-Dimensional Mean Vectors

Whitney V. Worley[1], Dean M. Young[2], Phil D. Young[3]

[1]Department of Mathematics, University of Central Arkansas, Conway, AR, USA
[2]Department of Statistical Science, Baylor University, Waco, TX, USA
[3]Department of Information Systems and Business Analytics, Baylor University, Waco, TX, USA
Email: wworley1@uca.edu

How to cite this paper: Worley, W.V., Young, D.M. and Young, P.D. (2021) Dimension Reduction for Detecting a Difference in Two High-Dimensional Mean Vectors. *Open Journal of Statistics*, **11**, 243-257.
https://doi.org/10.4236/ojs.2021.111013

Received: January 14, 2021
Accepted: February 23, 2021
Published: February 26, 2021

Abstract

We consider the efficacy of a proposed linear-dimension-reduction method to potentially increase the powers of five hypothesis tests for the difference of two high-dimensional multivariate-normal population-mean vectors with the assumption of homoscedastic covariance matrices. We use Monte Carlo simulations to contrast the empirical powers of the five high-dimensional tests by using both the original data and dimension-reduced data. From the Monte Carlo simulations, we conclude that a test by Thulin [1], when performed with post-dimension-reduced data, yielded the best omnibus power for detecting a difference between two high-dimensional population-mean vectors. We also illustrate the utility of our dimension-reduction method real data consisting of genetic sequences of two groups of patients with Crohn's disease and ulcerative colitis.

Keywords

Homoscedastic Covariance Matrices, Test Power, Monte Carlo Simulation, Moore-Penrose Inverse, Singular Value Decomposition

1. Introduction

When considering two multivariate-normal populations, researchers often attempt to determine if a difference exists between population-mean vectors. Traditionally, Hotelling's T^2 test has been utilized to determine if a difference exists. However, this test can be applied only when the data has a combined sample size that is greater than the original feature dimension because Hotelling's T^2 test depends on the non-singularity of the sample covariance matrix. When the sample-data dimension is greater than the sum of the sample sizes, we say the data is

"high-dimensional". For a fixed sample size, increasing the data dimension increases the covariance-matrix estimator variability, thus yielding statistical hypothesis tests for the difference in two population-mean vectors that are less powerful. Also, if the data dimension is greater than the sum of the group sample sizes, then the corresponding pooled-sample covariance matrix is singular and, therefore, one cannot conduct Hotelling's T^2 test for a mean difference. Hence, alternative tests for detecting the difference of two high-dimensional mean vectors, data dimension reduction (DR), or alternative tests combined with DR must be utilized.

In this paper, we investigate the efficacy of linear DR via the singular value decomposition (SVD) applied to a concatonated data matrix. Specifically, we examine the change in test powers for five tests for the difference of two high-dimensional mean vectors after implementing our DR to the original data. When applicable, we also apply the traditional Hotelling's T^2 test to the dimension-reduced data. Thus, using Monte Carlo power simulation studies, we contrast the powers of five high-dimensional tests with and without our proposed SVD-DR method. We concluded that SVD-DR, when applied to the data prior to conducting a test proposed by Thulin [1], yielded the largest omnibus power of the five considered tests whose empirical powers are contrasted here.

Throughout the paper, we use the notation $\mathbb{R}_{m\times n}$ and $\mathbb{R}_n$ to represent the matrix space of all $m\times n$ and $n\times n$ matrices, respectively, over the real field $\mathbb{R}$. Also, we let $\mathbb{R}_k^{\geq}$ represent the cone of all $k\times k$ nonnegative-definite real matrices, and we let $\mathbb{R}_k^{>}$ represent the cone interior of positive-definite matrices. Additionally, $\mathrm{tr}(\boldsymbol{A})$ represents the trace of the matrix $\boldsymbol{A}$, and $\|\boldsymbol{a}-\boldsymbol{b}\|$ denotes the Euclidean distance between the vectors $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}_{p\times 1}$. We assume that the data of the form $\boldsymbol{x}_{ij}\in\mathbb{R}_{p\times 1}$ with $i=1,2$ and $j=1,\cdots,N_i$, are randomly sampled from two distinct p-dimensional normal distributions denoted by $N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$, $i=1,2$, where the corresponding population-mean vectors and common population covariance matrix are denoted by $\boldsymbol{\mu}_i\in\mathbb{R}_{p\times 1}$, where $i=1,2$, and $\boldsymbol{\Sigma}\in\mathbb{R}_p^{>}$, respectively.

For $i=1,2$, the i^{th} sample mean vectors and sample covariance matrices are given by

$$\bar{\boldsymbol{x}}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{x}_{ij}$$

and

$$\boldsymbol{S}_i=\frac{1}{N_i-1}\sum_{j=1}^{N_i}\left(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i\right)\left(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i\right)^{\mathrm{T}},$$

respectively. An unbiased estimator of $\boldsymbol{\Sigma}$ is

$$\boldsymbol{S}_p=\frac{(N_1-1)\boldsymbol{S}_1+(N_2-1)\boldsymbol{S}_2}{N_1+N_2-2}. \tag{1}$$

The hypothesis test of interest is

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

Provided $\left(N_1 + N_2\right) > p$, Hotelling's T^2 statistic is

$$T^2 = \frac{N_1 N_2}{N_1 + N_2}\left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)^{\mathrm{T}} \boldsymbol{S}_p^{-1}\left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right), \tag{2}$$

where

$$\frac{N_1 + N_2 - p - 1}{\left(N_1 + N_2 - 2\right)p} T^2 \sim F\left(p, N_1 + N_2 - p + 1\right).$$

If $p > \left(N_1 + N_2\right)$, then (2) is incalculable.

Because of the increasing availability of high-dimensional data, especially in biological applications, researchers have proposed tests for the high-dimensional two-mean-vector problem. Dempster [2] first proposed a test for normally-distributed observation vectors where $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. The high-dimensional problem for contrasting two population mean vectors has been explored in the literature in such articles as Bai and Saranadasa [3], who proposed a test with the same asymptotic power as the test proposed by Dempster [2] but without relying on the assumption of normality.

Also, Srivastava [4] proposed a test similar to Hotelling's T^2 test where the inverse of (1) was replaced by its corresponding Moore-Penrose inverse. In addition, Srivastava and Du [5] have proposed replacing (1) with $\operatorname{diag}\left(\boldsymbol{S}_p\right)$ in (2), which results in a test statistic that is invariant under the group of non-singular $p \times p$ diagonal matrices. Park and Ayyala [6] and Chen and Quin [7] have also proposed test statistics for two high-dimensional population-mean vectors that do not rely on the assumption of equal covariance matrices, but these tests lose all information in the correlations between variables. Bickel and Levina [8] and Cai and Liu [9] have proposed tests using sparse estimators of the common covariance structure. In addition, Feng *et al.* [10] and Chen *et al.* [11] have proposed regularized versions of Hotelling's T^2 test. Moreover, Thulin [1] has proposed a modification to the test from Lopes *et al.* [12] by using random subspaces to improve test-statistic invariance properties. Zhang and Pan [13] followed the work of Thulin [1] by proposing a test using hierarchical clustering that more efficiently employs the covariance structure information in high dimensions. Srivastava *et al.* [14], He *et al.* [15], and others have also proposed tests for a difference in two high-dimensional population-mean vectors.

We have organized the remaining sections of the paper as follows. In Section 2, we present each of the five high-dimensional tests used to determine the utility of our proposed SVD-DR method. In Section 3, we present our SVD-DR approach, and in Section 4, we contrast the estimated power curves of each of the five considered tests using Monte Carlo simulations. Each test with and without SVD-DR is then applied to a bowel disease data set in Section 5. We then discuss the computational benefits of our proposed SVD-DR method in Section 6 and conclude with a brief discussion in Section 7.

2. Five Two-Sample Tests for a Difference between Two High-Dimensional Mean Vectors from Populations with Equal Covariance Matrices

We next describe five hypothesis tests for identifying differences in two high-dimensional mean vectors. Namely, we consider the tests derived in Bai and Saranadasa [3], Srivastava [4], Srivastava and Du [5], Thulin [1], and Zhang and Pan [13].

2.1. The Bai-Saranadasa Test

Dempster's test for $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ when $(N_1 + N_2) < p$ under the assumption of equal population covariance matrices is

$$T_D = \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} \frac{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^{\mathrm{T}} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{\mathrm{tr}\boldsymbol{S}_p}. \tag{3}$$

Let $n = (N_1 + N_2 - 2)$. Under specified conditions, Bai and Saranadasa [3] proposed an asymptotic-based version of (3) given by

$$T_{BS} = \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} \frac{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^{\mathrm{T}} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \mathrm{tr}\boldsymbol{S}_p}{\sqrt{2\left[\mathrm{tr}\boldsymbol{S}_p^2 - \frac{1}{n}(\mathrm{tr}\boldsymbol{S}_p)^2\right]}}.$$

2.2. Srivastava's T^{+2} Test

Srivastava [4] presented a test similar to (2) in which $\boldsymbol{S}_p^{-1}$ is replaced by the Moore-Penrose inverse $\boldsymbol{S}_p^{+}$ because $\mathrm{rank}(\boldsymbol{S}_p) < p$. His test statistic is

$$T^{+2} = \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^{\mathrm{T}} \boldsymbol{S}_p^{+} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2).$$

2.3. The Srivastava-Du Test

The tests given by Bai and Saranadasa [3] and Srivastava [4] are invariant under transformations of the type $\boldsymbol{x}_{ij} \to a\boldsymbol{\Gamma}\boldsymbol{x}_{ij}$, provided $a \neq 0$ and $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{I}_p$. However, these tests are not transformation-invariant for $p \times p$ non-singular, diagonal matrices. This fact implies that a change in units of measurement will affect the powers of both tests. To rectify this impediment, Srivastava and Du [5] proposed an invariant test statistic under $p \times p$ non-singular, diagonal matrices, which is

$$T_{SD} = \frac{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^{\mathrm{T}} \boldsymbol{D}_{\boldsymbol{S}_p}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - p}{\left[2\left\{\mathrm{tr}\hat{\boldsymbol{R}}^2 - \left(\frac{p^2}{n}\right)\right\} C_{p,n}\right]^{\frac{1}{2}}},$$

where $\boldsymbol{S}_p = (s_{ij})$, $i, j = 1, 2, \cdots, p$, $\boldsymbol{D}_{\boldsymbol{S}_p} = \mathrm{diag}(s_{11}, \cdots, s_{pp})$, $\hat{\boldsymbol{R}} = \boldsymbol{D}_{\boldsymbol{S}_p}^{-1/2} \boldsymbol{S}_p \boldsymbol{D}_{\boldsymbol{S}_p}^{-1/2}$,

and $C_{p,n} = 1 + \frac{\mathrm{tr}\hat{\boldsymbol{R}}^2}{p^{3/2}} \xrightarrow{p} 1$, as $(n, p) \to \infty$.

2.4. Thulin's Random-Subspaces Test

Lopes *et al.* [12] also presented a test for identifying a difference between two high-dimensional mean vectors that use random projections of the data onto k-dimensional subspaces, where k is a sufficiently small positive integer. However, this test is not invariant under a linear transformation of the marginal distributions.

Moreover, Thulin [1] proposed a modification to the test by Lopes *et al.* [12] that uses random subspaces in lieu of random pseudo-projections and is invariant under linear transformations of the marginal distributions. We can use **Algorithm 1** to compute the random-subspaces test presented by Thulin [1]. In the algorithm pseudo-code, we use the fact that $\boldsymbol{X}_1 \in \mathbb{R}_{N_1 \times p}$ and $\boldsymbol{X}_2 \in \mathbb{R}_{N_2 \times p}$ are the original-data matrices whose rows are composed of randomly sampled vectors of observations from their respective multivariate normal distributions.

The choice of the number of subspaces, k, directly affects the power of the random-subspaces test. If k is too small, much of the information contained in the multivariate structure is lost. However, if k is too large, test power is lost. For the random-subspaces test, Thulin [1] numerically verified that a good choice for k is $k = \lfloor (N_1 + N_2)/2 \rfloor$, where $\lfloor . \rfloor$ is the "floor" function. Also, let B represent the number of randomly-selected subspaces. Thulin [1] showed that the random-subspaces test is essentially stable for $B \geq 100$. **Algorithm 1** describes the steps needed to perform the random-subspaces test.

Algorithm 1: Computing the random-subspaces test.

1: Randomly select $k \leq n$ positive integers $\nu_1, \ldots, \nu_k$ from $\{1, 2, \ldots, p\}$, without replacement.
2: Create a random submatrix of each actual data set by creating the concatenated-vector-variable subset data sets $\mathbf{X}^*_{1(\nu)} = [\mathbf{x}_{1\nu_1} : \mathbf{x}_{1\nu_2} : \cdots : \mathbf{x}_{1\nu_k}]^T$ and $\mathbf{X}^*_{2(\nu)} = [\mathbf{x}_{2\nu_1} : \mathbf{x}_{2\nu_2} : \cdots : \mathbf{x}_{2\nu_k}]^T$, where $\mathbf{x}_{i\nu_j}$ is the j^{th} column selected from the original data matrix $\mathbf{X}_i, i = 1, 2$.
3: Calculate the empirical Hotelling's test score T^2_ν based on the variable-selected data sets $\mathbf{X}^*_{1(\nu)}$ and $\mathbf{X}^*_{2(\nu)}$.
4: Repeat steps 1-3 some B times, obtaining empirical test-statistic values for $T^2_1, \ldots, T^2_B$.
5: Obtain the final empirical test value $T_{RS} = B^{-1} \sum_{i=1}^{B} T^2_i$.
6: Contrast T_{RS} to an appropriate simulated critical value.

2.5. Zhang and Pan's Clustered Subspaces Test

Zhang and Pan [13] followed the work of Thulin [1] and made more efficient use of the information in (1) by proposing a test using hierarchical clustering in lieu of random subspaces. In particular, highly-correlated variables are clustered together to create subspaces. Hotelling's T^2 statistic is then applied to each clustered subspace.

Clusters are initially calculated based on a cutoff distance d_c. For their power-study simulations, Zhang and Pan [13] used the tuning parameter value $d_c = 1 - r_c$, where $r_c := \left(e^{2z_c^*} - 1\right) \Big/ \left(e^{2z_c^*} + 1\right)$, $z_c^* := t_c / \sqrt{n-1}$, $t_c := \Phi^{-1}\left(\left[1 - 2/\left(p(p-1)\right)\right]\right)$, and $\Phi(\cdot)$ is the standard normal cumulative

distribution function. After one performs clustering on the variables, some clusters may contain an excessive number of variables. If a cluster contains more than $k_c := \lfloor 2n/3 \rfloor$ variables, we partition the cluster into two sub-clusters. This partitioning process continues until each cluster or sub-cluster contains no more than k_c variables. One can use **Algorithm 2** to compute the clustered subspaces test presented in Zhang and Pan [13].

Algorithm 2: Computing the cluster-subspaces test.

1: Determine the tuning parameters d_c and k_c.
2: Perform hierarchical clustering based on the selected cutoff distance using average linkage.
3: For each cluster or sub-cluster with more than k_c variables, subdivide into two sub-clusters.
4: Repeat Step 3 until each cluster or sub-cluster has no more than k_c variables.
5: For each cluster or sub-cluster γ, calculate Hotelling's statistic T_γ^2.
6: Obtain the value of the empirical test statistic by calculating $T_{CS} = \sum_\gamma T_\gamma^2$.
7: Contrast T_{CS} to an appropriate simulated critical value.

3. Linear Dimension Reduction via the SVD of a Concatonated Matrix of Two High-Dimensional Data Sets

Below, we describe how we apply the SVD to a concatenated data matrix composed of the two sample-data sets to reduce the original data dimension. Through the deletion of a subset of right and left singular vectors associated with the SVD of the total data matrix, we eliminate information concerning the mean difference, which is of minor importance, while maintaining the bulk of the information for detecting that $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \neq \mathbf{0}$. Thus, to reduce the dimensionality of the two considered datasets, we propose the following DR method. First, we horizontally concatenate the two data matrices, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, to form the $N \times p$ data matrix $\boldsymbol{X}$ and calculate $\text{rank}(\boldsymbol{X})$. Next, given r, where $r < p$, we determine $\text{SVD}(\boldsymbol{X})$ and subsequently eliminate the $(p-r)$ columns of the left and right eigenvector matrices associated with the smallest $(p-r)$ eigenvalues of $\boldsymbol{X}$. This process yields an $(N \times r)$ approximation of $\boldsymbol{X}$. Below in **Algorithm 3**, we present steps for calculating the SVD-DR method.

Algorithm 3: SVD-DR of two-sample high-dimensional data.

1: Concatenate the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ so that $\mathbf{X} = \left[\mathbf{X}_1^T \vdots \mathbf{X}_2^T\right]^T$.
2: Calculate $SVD(\mathbf{X}) = \mathbf{U}\mathbf{D}\mathbf{V}^T$.
3: Partition $\mathbf{U}$ such that $\mathbf{U} = \left[\mathbf{U}_1 \vdots \mathbf{U}_2,\right]$, where $\mathbf{U}_1 \in \mathbb{R}_{N\times r}$ and $r = rank(\mathbf{X})$.
4: Partition $\mathbf{D}$ such that $\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_2 \end{bmatrix}$ with $\mathbf{D}_1 \in \mathbb{R}_{r\times r}$.
5: Partition $\mathbf{V}$ such that $\mathbf{V} = \left[\mathbf{V}_1 \vdots \mathbf{V}_2\right]$ with $\mathbf{V}_1 \in \mathbb{R}_{N\times r}$.
6: Reduce $\mathbf{X}$ to r dimensions by calculating $\mathbf{X}_r = \mathbf{U}_1\mathbf{D}_1\mathbf{V}_1^T$.
7: Partition $\mathbf{X}_r$ such that $\mathbf{X}_r = \left[\mathbf{X}_{1_r}^T \vdots \mathbf{X}_{2_r}^T\right]^T$ with $\mathbf{X}_{1_r} \in \mathbb{R}_{N_1\times r}$ and $\mathbf{X}_{2_r} \in \mathbb{R}_{N_2\times r}$.

4. A Monte Carlo Power Contrast of Five Tests for High-Dimensional Means with and without DR

To examine the powers of the five considered tests for a difference between two high-dimensional population-mean vectors before and after applying DR on the

concatenated data matrix $\boldsymbol{X}$, we simulated the powers of the five two-sample-mean tests over varying values of $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|$. Below, we describe the Monte Carlo simulation design that we used to examine the efficacy of the SVD-DR method.

4.1. Empirical Power Calculations

For our power simulations, we simulated critical values to compute the empirical power for each of the five considered tests for a difference between two high-dimensional population means. More specifically, we simulated θ replications of the data under H_0 and applied SVD-DR to each simulated data matrix. The $(\theta\alpha)^{th}$ largest value of the empirical test values was selected as the critical value, $\hat{c}_\alpha$. We then generated another θ replication of the data under H_1 and applied our SVD-DR approach to each of the simulated data sets. The empirical power was then calculated by

$$\hat{\beta} = \frac{\#\text{ of } t_A \geq \hat{c}_\alpha}{\theta},$$

where t_A represents the empirical test values generated under H_1. We used $\theta = 1000$, which corresponded to the number of simulation iterations incorporated for power simulations by both Thulin [1] and Zhang and Pan [13].

4.2. The Monte Carlo Power Simulation Design

For the power-contrast study to evaluate the efficacy of the SVD-DR method, we utilized a Monte Carlo configuration similar to that used by Thulin [1] and Zhang and Pan [13] to generate simulated data sets. We generated $N_i = 50$ observations of the form $\boldsymbol{x}_{ij} = \left(x_{i1}, \cdots, x_{iN_i}\right)^{\mathrm{T}}$ from $N_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$, $i = 1,2$. We also let $p = 200$, $\boldsymbol{\mu}_1 = \mathbf{0}_{p\times 1}$, and $\boldsymbol{\mu}_2 = \left(\mu_{2j}\right)$, where $j = 1, \cdots, p$, with

$$\mu_{2j} = \begin{cases} d, & \text{for } \lceil j/25 \rceil \leq m, \operatorname{mod}(j-1, 25) < 20, \\ 0, & \text{otherwise} \end{cases}$$

where $m = 1, 5, 8$ and $\lceil \cdot \rceil$ is the "ceiling" function. That is, we shifted 20 of 25 variable means in m of the $p/25$ subvectors. Also, we let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{r,s} = \left(\sigma_{rs}\right)$, where $r, s = 1, \cdots, p$, and

$$\sigma_{rs} = \begin{cases} 1, & \text{for } r = s, \\ r, & \text{for } r \neq s, \lceil r/25 \rceil = \lceil s/25 \rceil, \\ s, & \text{otherwise.} \end{cases}$$

Therefore, $\boldsymbol{\Sigma}_{r,s}$ denotes a covariance matrix with unit variances and $p/25$ equal-sized non-diagonal submatrices. The off-diagonal covariance elements in each submatrix are equal to r if r and s belong to the same submatrix block and equal to s, otherwise. Here, we consider the covariance structures $\boldsymbol{\Sigma}_{0,0}, \boldsymbol{\Sigma}_{0.5,0.2}$, and $\boldsymbol{\Sigma}_{0.9,0.1}$. For the Monte Carlo power simulations in Section 4.2, we used $k = \lfloor (N_1 + N_2)/2 \rfloor$ and $B = 100$.

4.3. The Monte Carlo Power-Simulation Results

In **Figure 1** and **Figure 3**, we display power plots for each of the considered tests

corresponding to each parameter configuration for which we have applied SVD-DR prior to conducting the tests. From the results shown in Figure 1, where all covariance matrices are diagonal, we observed that four of the five tests produced similar power curves. In particular, the test from Bai and Saranadasa [3] yielded the predominant power curve, and the test proposed by Srivastava [4]

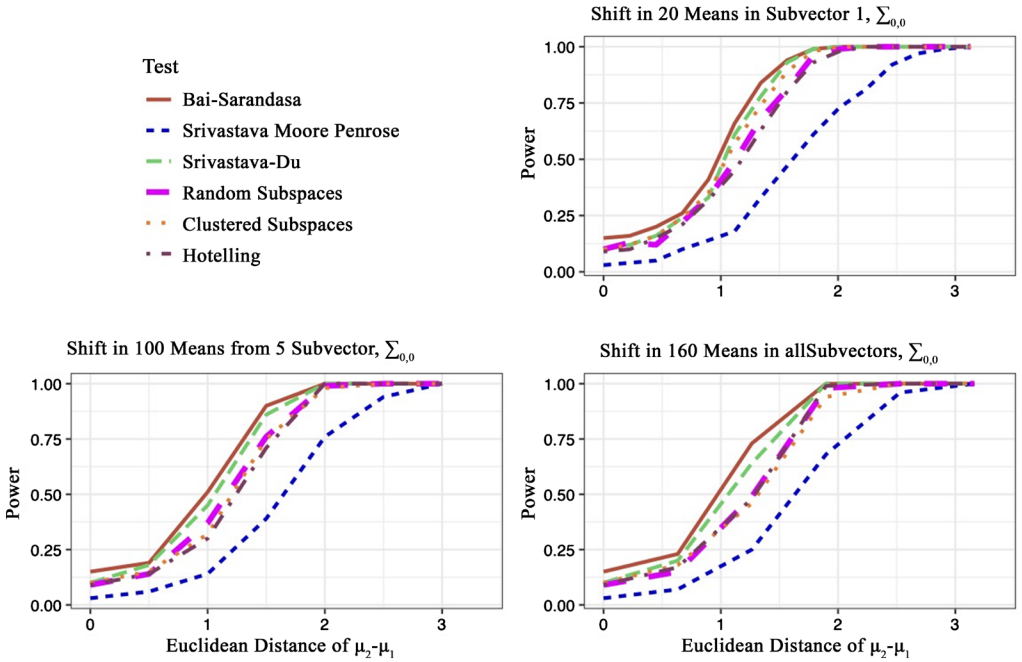

Figure 1. Power curves of the five tests for mean differences with post-SVD-DR data when $p = 200$, $n_1 = n_2 = 50$ and $\alpha = 0.05$ with diagonal covariance matrices.

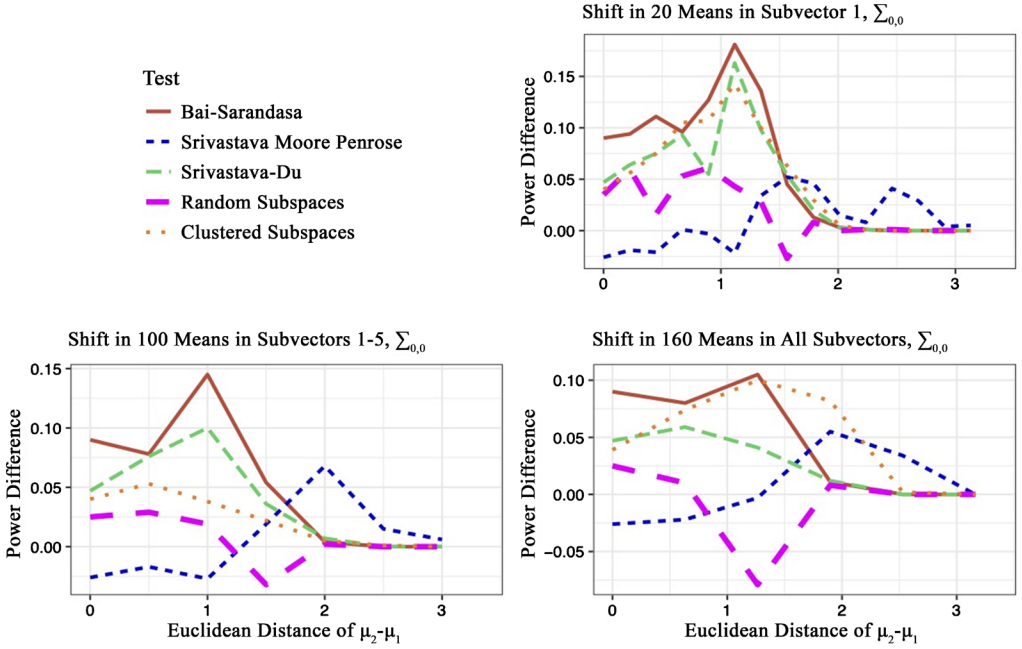

Figure 2. Power-difference plots of the five tests for mean differences conducted with post-SVD-DR data for $p = 200$, $n_1 = n_2 = 50$ and $\alpha = 0.05$ with diagonal covariance matrices.

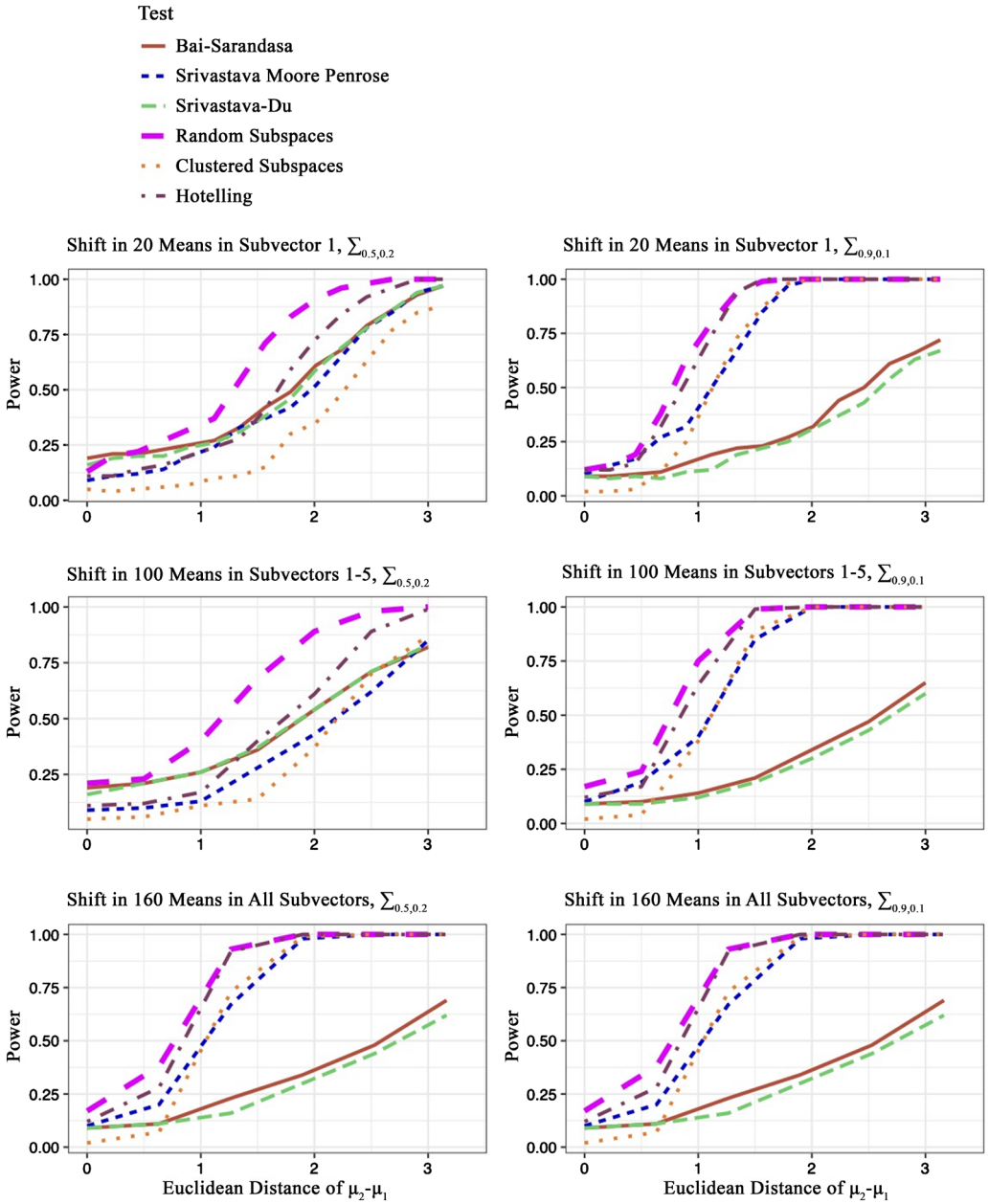

Figure 3. Power curves of the five tests for mean differences conducted with post-SVD-DR data when $p = 200$, $n_1 = n_2 = 50$ and $\alpha = 0.05$ with non-diagonal covariance matrices.

invariably produced the least-prominent power curve. All five high-dimensional tests for a difference between two mean vectors contrasted here yielded increased power as m increased. Also, when the common covariance matrix was diagonal, we found that performing Hotelling's T^2 test on the SVD-DR data yielded powers comparable to those of the competing tests for two high-dimensional population means using post-SVD-DR data.

In Figure 2, we displayed power-difference plots for six parameter configurations with diagonal common covariance structures. For each test, the plots illustrate differences among the power curves using SVD-DR data minus the power

curves using the original unreduced data. As shown in **Figure 2**, the tests by Bai and Saranadasa [3], Srivastava and Du [5], and the random subspaces test by Thulin [1] yielded a slight to moderate power increase when $0.01 \le \|\mu_2 - \mu_1\| \le 2.00$. In contrast, the test proposed by Zhang and Pan [13] produced slightly increased power and the test by Srivastava [4] actually yielded decreased power for some values of $\|\mu_2 - \mu_1\|$.

In **Figure 3**, we display power curves for the five tests conducted on SVD-DR data for six parameter configurations with non-diagonal covariance structures. We see that Thulin's random-subspaces test, when applied to the SVD-DR data, yielded the best omnibus power curve for all six parameter configurations considered here. For most parameter configurations shown in **Figure 3**, Hotelling's T^2 test, conducted with SVD-DR data, yielded power curves similar to those of the random subspaces test by Thulin [1] across all mean and covariance matrix configurations with non-diagonal covariance structures. As within-block correlation and m increased, the test by Bai and Saranadasa [3] and the cluster-subspaces test of Srivastava and Du [5] generally provided the two smallest power curves.

More importantly, we contrasted the powers of each test for the difference between two high-dimensional mean vectors with and without the application of the SVD-DR to the original data in **Figure 4**. The plots display the power of each test with the SVD-DR data minus the corresponding test power without the SVD-DR method applied to the full-dimensional data. That is, the graphed plots represent the average power-difference for the powers before and after applying SVD-DR on the $\theta = 1000$ data sets for each of the five tests for a difference between two high-dimensional population-mean vectors.

In each of the power-difference plots in **Figure 4**, we observed moderate to large increased test power for the random subspaces test proposed by Thulin [1] for all six parameter configurations. Specifically, Thulin's [1] random subspaces test showed substantial maximal gains in power that ranged between 0.20 and 0.70, depending on $\|\mu_2 - \mu_1\|$ and on the type of population covariance structure. The tests given in Bai and Saranadasa [3] and Srivastava [4] yielded moderate power gains in three of the six parameter configurations. However, the clustered-random-subspaces test by Zhang and Pan [13] produced little increase in power and some decreased power in five of the six parameter configurations. The most significant result from **Figure 4** was that SVD-DR consistantly and substantially improved the power of the random subspaces test by Thulin [1], which was already the most powerful considered test on the unreduced data. In addition, we see that when $\|\mu_2 - \mu_1\|$ is relatively large, the degree of power improvement is considerably lessened because the tests considered already have large power.

5. A Contrast of Test Performance with and without SVD-DR for Bowel Disease Data

Burczynski *et al.* [16] studied patients with two common inflammatory bowel

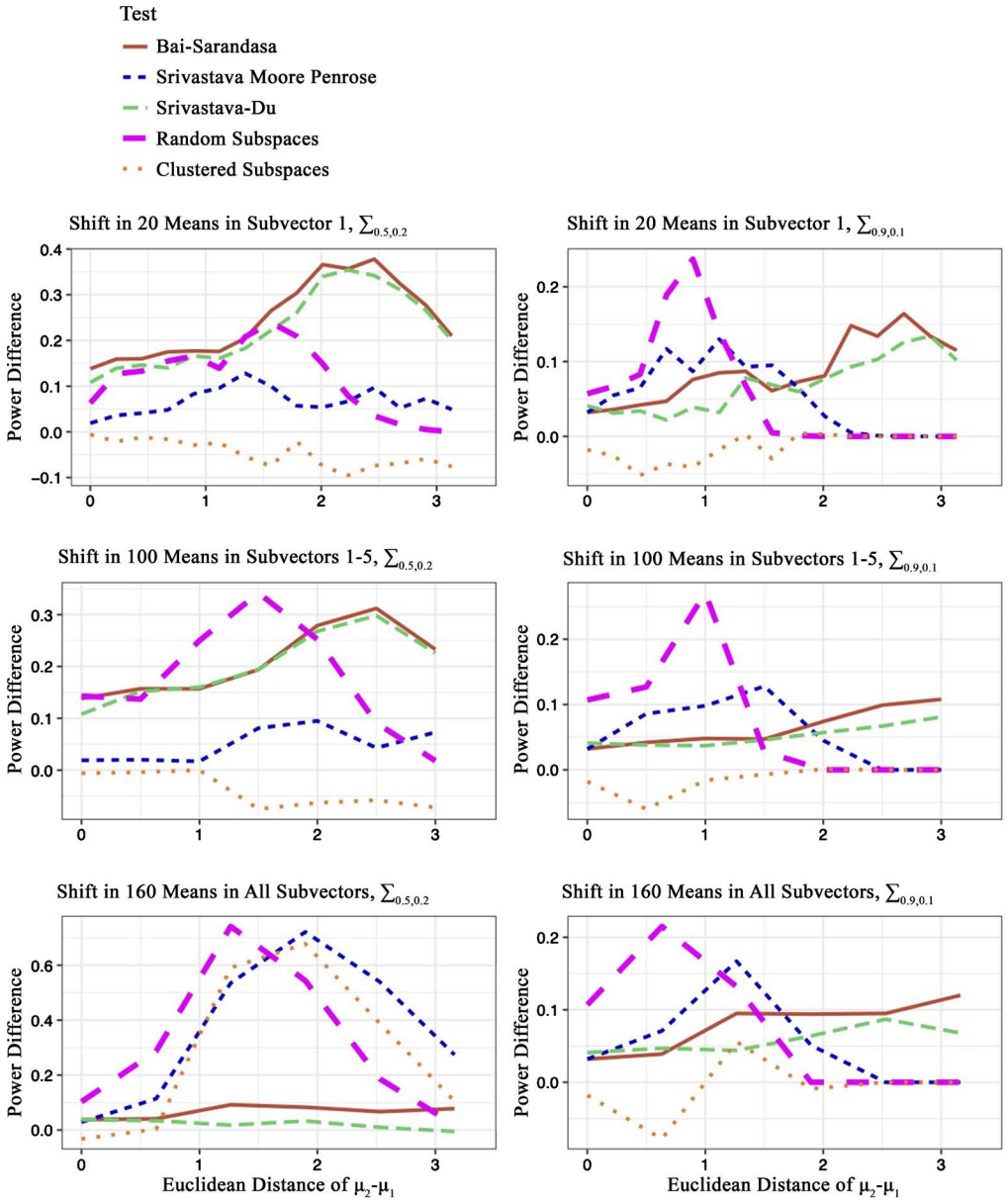

Figure 4. Power-difference plots of the five tests for mean differences conducted with post-SVD-DR for $p = 200$, $n_1 = n_2 = 50$ and $\alpha = 0.05$ with non-diagonal covariance matrices.

diseases that produce intestinal inflammation and cause tissue damage: Crohn's disease and ulcerative colitis. For patients with inflammatory bowel disease, approximately 10% were diagnosed with diseases that were medically classified as indeterminate following a colonoscopy even though these two diseases are distinct. Burczynski *et al.* [16] analyzed transcriptional profiles in peripheral blood mononuclear cells for patients with either Crohn's disease or ulcerative colitis by hybridization to microarrays of more than 22,000 genetic sequences.

To illustrate that the proposed SVD-DR method can increase the test powers for identifying a difference between two high-dimensional population-mean vectors, we applied each of the five tests to both SVD-reduced and full-dimensional

data. We used $N_1 = 10$ randomly-selected patients with Crohn's disease and $N_2 = 10$ randomly-selected patients with ulcerative colitis. We randomly chose $p = 2000$ features to demonstrate the SVD-DR efficacy. In addition, we applied Hotelling's T^2 test to the post-SVD-DR data.

We present the results of the five tests for two high-dimensional mean vectors with and without SVD-DR and Hotelling's T^2 with SVD-DR in **Table 1**. No difference between the two population mean-vector gene-expression levels of the patient groups was detected by any of the five high-dimensional tests when the tests were applied to the full-dimensional data. However, when SVD-DR was applied to the data prior to performing the five tests, a difference between the two high-dimensional mean vectors of patients with Crohn's disease and ulcerative colitis was observed for three of the five tests: Thulin [1], Zhang and Pan [13], and Srivastava [4]. On the other hand, the tests from Bai and Saranadasa [3] and Srivastava and Du [5] yielded a relatively small reduction in the test p-value and, therefore, failed to produce a statistically significant result.

The increased power for the random subspaces test proposed by Thulin [1] was not surprising, given the power-curve and power-difference plots in **Figure 3** and **Figure 4**, respectively. However, the increased power for the test proposed by Srivastava [4] was surprising and seems to have occurred because the common sample covariance structure of the reduced data contained many relatively large off-diagonal elements. In addition, Hotelling's T^2 test, conducted with the post-SVD-DR data, detected a difference in the two high-dimensional mean vectors for patients with Crohn's disease and patients with ulcerative colitis.

6. The Computational Benefit of SVD-DR

In conjunction with improvements in the power, an additional benefit of the SVD-DR application is a reduced computational intensity needed to conduct tests for a difference between two high-dimensional population-mean vectors. For the tests proposed in Bai and Saranadasa [3], Srivastava [4], and Srivastava and Du [5], computation of (1) is time-consuming. The random subspaces test by Thulin [1] and the clustered subspaces Zhang and Pan [13] are also computationally intense because of the increased number of data projections required for high-dimensional data. The application of the SVD-DR method before testing

Table 1. Test results for bowel disease data with and without SVD-DR.

Test	Full dim. T-score	Full dim. p-value	SVD-DR T-score	SVD-DR p-value
Bai-Saranadasa	0.03	0.488	0.13	0.448
Srivastava T^{*2}	35.00	0.873	123.77	0.013
Srivastava-Du	0.04	0.485	0.53	0.298
Random Subspaces	24.66	0.310	97.30	0.000
Cluster Subspaces	5136.26	0.076	22.76	0.044
Hotelling's T^2	NA	NA	5.57	0.008

for a difference between two high-dimensional population means using hierarchical cluster subspaces (**Algorithm 2**) drastically reduced the computational demand for this test.

To demonstrate the computational efficacy of the SVD-DR method before performing a hypothesis test for the difference in two high-dimensional means, we summarized the computation times for the five tests on the real data introduced in Section 5 with and without the application of SVD-DR to the data in Table 2. The computational times for the tests using SVD-DR include both the time to reduce the data dimension and the time to conduct the test itself.

In Table 2, we observed decreased computation times for all five tests for detecting a difference between two high-dimensional population mean-vectors. In particular, for the Zhang and Pan [13] test, we saw a drastic computational-time reduction from 13.8 hours to 3.04 seconds.

7. Discussion

We have contrasted the changes in powers for five previously-proposed tests for the difference of two high-dimensional population-mean vectors when using the SVD to reduce the dimensionality of the two sample-data sets and with the original high-dimensional data. From Figure 2, we observed that under the configurations with diagonal covariance structures, the tests proposed by Bai and Saranadasa [3], Thulin [1], Srivastava and Du [5], and Zhang and Pan [13] all displayed moderately-increased power for relatively small values of $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|$, when applied to post-SVD-DR data. The Moore-Penrose inverse test proposed by Srivastava [4] demonstrated a mixture of very slightly-increased power and decreased power when the test was performed on SVD-DR data.

For non-diagonal covariance matrices, when SVD-DR was applied to the original data prior to conducting each of the five tests, the random-subspaces test displayed the largest power increase as shown in Figure 4. Also, the test proposed by Zhang and Pan [13] actually lost power for certain values of $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|$ for five of the six parameter configurations considered here.

We also applied the five tests for a difference between two high-dimensional population means to actual data from Burczynski *et al.* [16], which consisted of genetic data from patients with two types of inflammatory bowel disease. After

Table 2. Computational times in seconds for executing the five tests for two high-dimensional mean vectors on the bowel disease data with SVD-DR data and without first applying the SVD-DR method to reduce the data dimension.

Test	No SVD-DR	With SVD-DR
Bai-Saranadasa	5.93	3.27
Srivastava T^{+2}	6.20	0.17
Srivastava-Du	28.70	0.18
Random Subspaces	5.93	3.27
Cluster Subspaces	49790.00	3.04

applying SVD-DR to a random subset of variables sampled from the original data, we found that Thulin's random-subspaces test, Srivastava's T^{+2} test, and Zhang and Pan's cluster-subspaces test identified a difference in the two high-dimensional population-mean vectors of DNA for both patients with Crohn's disease and patients with ulcerative colitis at the 5% significance level. However, the tests proposed by Bai and Saranadasa [3] and Srivastava and Du [5] did not find evidence of a difference between the two high-dimensional mean vectors when conducted with SVD-DR data. Finally, we demonstrated the computational benefit of applying SVD-DR on a subset of the data from Burczynski *et al.* [16] prior to conducting the five tests.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Thulin, M. (2014) A High-Dimensional Two-Sample Test for the Mean Using Random Subspaces. *Computational Statistics and Data Analysis*, **74**, 26-38. https://doi.org/10.1016/j.csda.2013.12.003

[2] Dempster, A.P. (1958) A High Dimensional Two Sample Significance Test. *Annals of Mathematical Statistics*, **29**, 995-1010. https://doi.org/10.1214/aoms/1177706437

[3] Bai, Z. and Saranadasa, H. (1996) Effect of High Dimension: By an Example of a Two Sample Problem, *Statistica Sinica*, **6**, 311-329.

[4] Srivastava, M.S. (2007) Multivariate Theory for Analyzing High Dimensional Data. *The Journal of the Japan Statistical Society*, **37**, 53-86. https://doi.org/10.14490/jjss.37.53

[5] Srivastava, M.S. and Du, M. (2008) A Test for the Mean Vector with Fewer Observations than the Dimension. *Journal of Multivariate Analysis*, **99**, 386-402. https://doi.org/10.1016/j.jmva.2006.11.002

[6] 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**, 929-943. https://doi.org/10.1016/j.jspi.2012.11.001

[7] Chen, S.X. and Qin, Y.-L.N. (2010) A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing. *Annals of Statistics*, **38**, 808-835. https://doi.org/10.1214/09-AOS716

[8] Bickel, P. and Levina, E.N. (2008) Regularized Estimation of Large Covariance Matrices. *Annals of Statistics*, **36**, 199-227. https://doi.org/10.1214/009053607000000758

[9] Cai, T. and Liu, W.N. (2011) Adaptive Thresholding for Sparse Covariance Matrix Estimation. *Journal of the American Statistical Association*, **106**, 672-684. https://doi.org/10.1198/jasa.2011.tm10560

[10] Feng, L., Zou, C.N. and Wang, Z. (2016) Multivariate-Sign-Based High-Dimensional Tests for the Two-Sample Location Problem. *Journal of the American Statistical Association*, **111**, 721-735. https://doi.org/10.1080/01621459.2015.1035380

[11] Chen, L.S., Paul, D., Prentice, R.L. and Wang, P. (2011) A Regularized Hotelling's T^2 Test for Pathway Analysis in Proteomic Studies. *Journal of the American Statis-*

tical Association, **106**, 1345-1360. https://doi.org/10.1198/jasa.2011.ap10599

[12] Lopes, M., Jacob, L. and Wainright, M.J. (2011) A More Powerful Two-Sample Test in High Dimensions Using Random Projection. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F. and Weinberger, K.Q., Eds., *Advances in Neural Information Processing Systems*, Vol. 24, Curran Associates Inc., Red Hook, 1206-1214.

[13] Zhang, J. and Pan, M. (2016) A High-Dimension Two-Sample Test for the Mean Using Cluster Subspaces. *Computational Statistics and Data Analysis*, **97**, 87-97. https://doi.org/10.1016/j.csda.2015.12.004

[14] Srivastava, R., Li, P. and Ruppert, D. (2016) RAPPT: An Exact Two-Sample Test in High Dimensions Using Random Projections. *Journal of Computational and Graphical Statistics*, **25**, 954-970. https://doi.org/10.1080/10618600.2015.1062771

[15] He, Y., Zhang, M., Zhang, X. and Zhou, W. (2020) High-Dimensional Two-Sample Mean Vectors Test and Support Recovery with Factor Adjustment. *Computational Statistics and Data Analysis*, **151**, Article ID: 107004. https://doi.org/10.1016/j.csda.2020.107004

[16] Burczynski, M.E., Peterson, R.L., Twine, N.C., Zuberek, K.A., Brodeur, B.J., Casciotti, L., Maganti, V., Reddy, P.S., Strahs, A., Immermann, F., Spinelli, W., Schwertschlag, U., Slager, A.M., Cotreau, M.M. and Dorner, A.J. (2006) Molecular Classification of Crohn's Disease and Ulcerative Colitis Patients Using Transcriptional Profiles in Peripheral Blood Mononuclear Cells. *The Journal of Molecular Diagnostics*, **8**, 51-61. https://doi.org/10.2353/jmoldx.2006.050079