J. Japan Statist. Soc.
Vol. 37 No. 1 2007 53–86

MULTIVARIATE THEORY FOR ANALYZING HIGH DIMENSIONAL DATA

M. S. Srivastava*

In this article, we develop a multivariate theory for analyzing multivariate datasets that have fewer observations than dimensions. More specifically, we consider the problem of testing the hypothesis that the mean vector $\boldsymbol{\mu}$ of a p-dimensional random vector $\boldsymbol{x}$ is a zero vector where N, the number of independent observations on $\boldsymbol{x}$, is less than the dimension p. It is assumed that $\boldsymbol{x}$ is normally distributed with mean vector $\boldsymbol{\mu}$ and unknown nonsingular covariance matrix Σ. We propose the test statistic $F^+ = n^{-2}(p-n+1)N\overline{\boldsymbol{x}}'S^+\overline{\boldsymbol{x}}$, where $n = N-1 < p$, $\overline{\boldsymbol{x}}$ and S are the sample mean vector and the sample covariance matrix respectively, and S^+ is the Moore-Penrose inverse of S. It is shown that a suitably normalized version of the F^+ statistic is asymptotically normally distributed under the hypothesis. The asymptotic non-null distribution in one sample case is given. The case when the covariance matrix Σ is singular of rank r but the sample size N is larger than r is also considered. The corresponding results for the case of two-samples and k samples, known as MANOVA, are given.

Key words and phrases: Distribution of test statistics, DNA microarray data, fewer observations than dimension, multivariate analysis of variance, singular Wishart.

1. Introduction

In DNA microarray data, gene expressions are available on thousands of genes of an individual but there are only few individuals in the dataset. For example, in the data analyzed by Ibrahim *et al.* (2002), gene expressions were available on only 14 individuals, of which 4 were normal tissues and 10 were endometrial cancer tissues. But even after excluding many genes, the dataset consisted of observations on 3214 genes. Thus, the observation matrix is a 3214×14 matrix. Although these genes are correlated, it is ignored in the analysis. The empirical Bayes analysis of Efron *et al.* (2001) also ignores the correlations among the genes in the analysis of their data. Similarly, in the comparison of discrimination methods for the classification of tumors using gene expression data, Dudoit *et al.* (2002) considered in their Leukemia example 6817 human genes on 72 subjects from two groups, 38 from the learning set and 34 from the test set. Although it gives rise to a 6817×6817 correlation matrix, they examined a 72×72 correlation matrix. Most of the methods used in the above analyses ignore the correlations among the genes.

The above cited papers use some criteria for reducing the dimension of the data. For example, Dudoit *et al.* (2002) reduce the dimension of their Leukemia data from 3571 to 40, before they applied their methods of classification on

Received July 15, 2005. Revised October 21, 2005. Accepted January 25, 2006.

*Department of Statistics, University of Toronto, 100 St. George Street Toronto, Ontario, Canada M5S 3G3. Email: srivastava@utstat.toronto.edu

the data. Methods for reducing the dimension in microarray data have also been given by Alter *et al.* (2000) in their singular value decomposition methods. Similarly, Eisen *et al.* (1998) have given a method of clustering the genes using a measure similar to the Pearson correlation coefficient. Dimension reduction has also been the objective of Efron *et al.* (2001) and Ibrahim *et al.* (2002). Such reduction of the dimension is important in microarray data analysis because even though there are observations on thousands of genes, there are relatively few genes that reflect the differences between the two groups of data or several groups of data. Thus, in analyzing any microarray data, the first step should be to reduce the dimension. The theory developed in this paper provides a method for reducing the dimension.

The sample covariance matrix for the microarray datasets on N individuals with p genes in which p is larger than N is a singular matrix of rank $n = N - 1$. For example if x_{ik} denotes the gene expression of the i-th gene on the k-th individual, $i = 1, \dots, p$, $k = 1, \dots, N$, then the sample covariance matrix is of the order $p \times p$ given by

$$S = (s_{ij}),$$

where

$$ns_{ij} = \sum_{k=1}^{N} (x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j),$$

and $\overline{x}_i$ and $\overline{x}_j$ are the sample means of the i-th and j-th genes' expressions, $\overline{x}_i = N^{-1} \sum_{k=1}^{N} x_{ik}$, $\overline{x}_j = N^{-1} \sum_{k=1}^{N} x_{jk}$. The corresponding $p \times p$ population covariance matrix Σ will be assumed to be positive-definite. To understand the difficulty in analyzing such a dataset, let us divide the p genes arbitrarily into two groups, one consisting of $n = N-1$ genes and the other consisting of $q = p-N+1$ genes. We write the sample covariance matrix for the two partitioned groups as

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}' & S_{22} \end{pmatrix}, \qquad S_{22} = (s_{22ij}),$$

where $S_{12} = (\boldsymbol{s}_{121}, \dots, \boldsymbol{s}_{12q}) : n \times q$, $q = p - n$, $n = N - 1$. Then the sample multiple correlation between the i-th gene in the second group with the n genes in the first group is given by

$$s_{22ii}^{-1} \boldsymbol{s}_{12i}' S_{11}^{-1} \boldsymbol{s}_{12i}.$$

However, since the matrix S is of rank n, it follows from Srivastava and Khatri (1979, p. 11) that

$$S_{22} = S_{12}' S_{11}^{-1} S_{12},$$

giving $s_{22ii} = \boldsymbol{s}_{12i}' S_{11}^{-1} \boldsymbol{s}_{12i}$. Thus, the sample multiple correlation between any gene i in the second set with the first set is always equal to one for all $i = 1, \dots, p - n$. Similarly, all the n sample canonical correlations between the two sets are one for $n < (p - n)$. It thus appears that any inference procedure that takes into account the correlations among the genes may be based on n genes or

their linear combinations such as n principal components. On the other hand, since the test procedure discussed in this article is based on the Moore-Penrose inverse of the singular sample covariance matrix $S = n^{-1}YY'$, which involves the non-zero eigenvalues of YY' or equivalently of $Y'Y$, $Y : p \times n$, it would be desirable to have a larger p than dictated by the above consideration.

Another situation that often arises in practice is the case when $n \simeq p$. For in this case, even if $n > p$ and even though theoretically the covariance matrix S is positive definite, the smaller eigenvalues are really small as demonstrated by Johnston (2001) for $p = n = 10$. Since this could also happen if the covariance matrix is singular of rank $r \leq p$, we shall assume that the covariance matrix is singular of rank $r \leq n$. In the analysis of microarrays data, this situation may arise since the selected number of characteristics could be close to n.

The objective of this article is to develop a multivariate theory for analyzing high-dimensional data. Specifically, we consider the problem of testing the hypothesis concerning the mean vector in one-sample, equality of the two mean vectors in the two-sample case, and the equality of several mean vectors—the so called MANOVA problem in Sections 2, 3 and 4 respectively. The null distributions of these statistics are also given in these sections. In the one-sample case, the non-null distribution of the test statistic is also given in Section 2. The confidence intervals for linear combinations of the mean vectors or for linear combinations of contrasts in one, two, and more than two samples are given in Section 5.

Tests for verifying the assumption of the equality of covariances are given in Section 6. In Section 7, we give a general procedure for reducing the dimension of the data. An example from Alon *et al.* (1999) on colon, where the gene expressions are taken from normal and cancerous genes is discussed in Section 8. The paper concludes in Section 9.

2. One-sample problem

Let p-dimensional random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be independent and identically distributed (hereafter referred to as iid) as normal with mean vectors $\boldsymbol{\mu}$ and unknown nonsingular covariance matrix Σ. Such a distribution will be denoted by $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$. We shall assume that $N \leq p$ and consider the problem of testing the hypothesis $H : \boldsymbol{\mu} = \mathbf{0}$ against the alternative $A : \boldsymbol{\mu} \neq \mathbf{0}$. The sample mean vector and the sample covariance matrix are respectively defined by

$$\overline{\boldsymbol{x}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i, \qquad \text{and} \qquad S = n^{-1}V, \tag{2.1}$$

where

$$V = \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})', \qquad n = N - 1. \tag{2.2}$$

Let B^+ denote the Moore-Penrose inverse of the $m \times r$ matrix B satisfying the four conditions (i) $BB^+B = B$, (ii) $B^+BB^+ = B^+$, (iii) $(B^+B)' = (B^+B)$,

(iv) $(BB^+)' = BB^+$. The Moore-Penrose inverse is unique and is equal to the inverse of the nonsingular $m \times m$ square matrix B. We propose the following test statistics for testing the hypothesis that $\boldsymbol{\mu} = \mathbf{0}$ vs $\boldsymbol{\mu} \neq \mathbf{0}$. Define for $n < p$,

$$\begin{aligned} T^{+2} &= N\overline{\boldsymbol{x}}' S^+ \overline{\boldsymbol{x}} \\ &= nN\overline{\boldsymbol{x}}' V^+ \overline{\boldsymbol{x}}. \end{aligned} \tag{2.3}$$

Let

$$F^+ = \frac{p-n+1}{n}\frac{T^{+2}}{n}. \tag{2.4}$$

Thus, for $n < p$, we propose the statistic F^+ or equivalently the T^{+2} statistic for testing the hypothesis that $\boldsymbol{\mu} = \mathbf{0}$, against the alternative that $\boldsymbol{\mu} \neq \mathbf{0}$. It is shown in Subsection 2.1 that F^+ is invariant under the linear transformation $\boldsymbol{x}_i \to c\Gamma\boldsymbol{x}_i$, $c \neq 0$, and $\Gamma\Gamma' = I_p$. Let

$$\hat{b} = \frac{(n-1)(n+2)}{n^2}\frac{(\operatorname{tr} S/p)^2}{p^{-1}\left[\operatorname{tr} S^2 - \frac{1}{n}(\operatorname{tr} S)^2\right]}. \tag{2.5}$$

An asymptotic distribution of the F^+ statistic, given later in Theorem 2.3, is given by

$$\lim_{n,p\to\infty} P\left[\left(\frac{n}{2}\right)^{1/2}(p(p-n+1)^{-1}\hat{b}F^+ - 1) \le z_{1-\alpha}\right] = \Phi(z_{1-\alpha}), \tag{2.6}$$

where Φ denotes the cumulative distribution function (cdf) of a standard normal random variable with mean 0 and variance 1. In most practical cases $n = O(p^\delta)$, $0 < \delta < 1$, and in this case (2.6) simplifies to

$$\lim_{n,p\to\infty} P\left[c_{p,n}\left(\frac{n}{2}\right)^{1/2}(\hat{b}F^+ - 1) < z_{1-\alpha}\right] = \Phi(z_{1-\alpha}), \tag{2.7}$$

where we choose

$$c_{p,n} = [(p-n+1)/(p+1)]^{1/2}$$

for fast convergence to normality, see Corollary 2.1.

When

$$\boldsymbol{\mu} = \left(\frac{1}{nN}\right)^{1/2}\boldsymbol{\delta},$$

where $\boldsymbol{\delta}$ is a vector of constants, then the asymptotic power of the F^+ test as given in Subsection 2.3 is given by

$$\begin{aligned} \beta(F^+) &= \lim_{n,p\to\infty} P\left[\frac{n\hat{b}p(p-n+1)^{-1}F^+ - n}{(2n)^{1/2}} > z_{1-\alpha} \mid \boldsymbol{\mu}\right] \\ &\simeq \Phi\left(-z_{1-\alpha} + \left(\frac{n}{p}\right)\left(\frac{n}{2}\right)^{1/2}\frac{\boldsymbol{\mu}'\wedge\boldsymbol{\mu}}{a_2}\right), \end{aligned}$$

where $\wedge = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and λ_i are the eigenvalues of the covariance matrix.

A competitive test proposed by Dempster (1958, 1960) is given by

$$T_D = \left(\frac{N \overline{\boldsymbol{x}}' \overline{\boldsymbol{x}}}{\text{tr}\, S} \right).$$

The asymptotic power of the T_D test as given by Bai and Saranadasa (1996) is given by

$$\beta(T_D) \simeq \Phi \left(-z_{1-\alpha} + \frac{n \boldsymbol{\mu}' \boldsymbol{\mu}}{\sqrt{2pa_2}} \right).$$

Thus, when $\Sigma = \gamma^2 I$,

$$\beta(T_D) = \left[-z_{1-\alpha} + \left(\frac{n}{\sqrt{2p}} \right) \frac{\boldsymbol{\mu}' \boldsymbol{\mu}}{\gamma^2} \right],$$

and

$$\beta(F^+) = \left[-z_{1-\alpha} + \left(\frac{n}{p} \right)^{1/2} \left(\frac{n}{\sqrt{2p}} \right) \frac{\boldsymbol{\mu}' \boldsymbol{\mu}}{\gamma^2} \right].$$

Thus, in this case, Dempster's test is superior to the F^+ test, unless $(n/p) \to 1$. It is as expected since Dempster's test is uniformly most powerful among all tests whose power depends on $(\boldsymbol{\mu}' \boldsymbol{\mu}/\gamma^2)$, see Simaika (1941). This test is also invariant under the linear transformations $\boldsymbol{x}_i \to c\Gamma \boldsymbol{x}_i$, $c \neq 0$, $\Gamma\Gamma' = I_p$. In other cases, F^+ may be preferred if

$$\left(\frac{n}{pa_2} \right)^{1/2} \boldsymbol{\mu}' \wedge \boldsymbol{\mu} > \boldsymbol{\mu}' \boldsymbol{\mu}; \tag{2.8}$$

For example, if $\boldsymbol{\mu} \simeq N_p(\mathbf{0}, \wedge)$, then on the average (2.8) implies that

$$\left(\frac{n}{pa_2} \right)^{1/2} (\text{tr} \wedge^2) > \text{tr} \wedge,$$

that is

$$n > \frac{a_1^2}{a_2} p = bp.$$

Since $a_1^2 < a_2$, such an n exists. Similarly, if $\boldsymbol{\mu} = \lambda_i^{1/2}$, the same inequality is obtained and F^+ will have better power than the Dempster test.

Next, we compare the power of T_D and F^+ tests by simulation where the F^+ statistic in (2.7) is used. From the asymptotic expressions of the power given above and in (2.29), it is clear that asymptotically, the tests T_D and T_{BS} (defined in equation (2.26)) have the same power. Thus, in our comparison of power, we include only the Dempster's test T_D as it is also known that if $\Sigma = \sigma^2 I_p$, then the test T_D is the best invariant test under the transformation $\boldsymbol{x} \to c\Gamma \boldsymbol{x}$, $c \neq 0$, $\Gamma\Gamma' = I_p$, among all tests whose power depends on $\boldsymbol{\mu}' \boldsymbol{\mu}/\sigma^2$ irrespective of the size of n and p. Therefore it is better than the Hotelling's T^2-test (when $n > p$), T_{BS} and the F^+ tests. However, when $\Sigma \neq \sigma^2 I$, then no

ordering between the two tests T_D and F^+ exist. Thus, we shall compare the power of the T_D test with the F^+ test by simulation when the covariance matrix $\Sigma = \text{diag}(d_1, \ldots, d_p) \neq \sigma^2 I_p$. The diagonal elements $d_1, \ldots, d_p$ are obtained as an iid sample from several distributions. However, once the values of $d_1, \ldots, d_p$ are obtained by a single simulation, they are kept fixed in our simulation study. Similarly, the values of the elements of the mean vector $\boldsymbol{\mu}$ for the alternative hypothesis are obtained as iid observations from some distributions. Once these values are obtained, they are also held fixed throughout the simulation. In order that both tests have the same significance level, we obtain by simulation the cut-off points for the distribution of the statistic under the hypothesis. For example, for the statistic F^+, we obtain F^+_α such that

$$\frac{(\#\text{ of } F^+ \geq F^+_\alpha)}{1000} = \alpha,$$

where F^+ is calculated from the $(n+1)$ samples from $N_p(\mathbf{0}, \Sigma)$ for each 1000 replications. The power is then calculated from the $(n+1)$ samples from $N_p(\boldsymbol{\mu}, \Sigma)$ replicated again 1000 times. We have chosen α to be 0.05. The power of the two tests are shown in Tables 1–3. The mean vectors for the alternative are obtained as

$$\begin{aligned}
\boldsymbol{\mu}_1 &= \{x_1, \ldots, x_p\}, \qquad i = 1, \ldots, p, \;\; x_i \sim U(-0.5, 0.5). \\
\boldsymbol{\mu}_2 &= \{x_1, \ldots, x_p\}, \qquad i = 1, \ldots, p, \;\; \text{for } i = 2k, \; x_i \sim U(-0.5, 0.5); \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{for } i = 2k+1, \; x_i = 0, \; k = 0, 1, \ldots, p-1.
\end{aligned}$$

In our power comparison, we have used the statistic given in Corollary 2.1.

Table 1. Power, $\Sigma = I_p$.

		$\boldsymbol{\mu}_1$		$\boldsymbol{\mu}_2$	
p	n	T_D	F^+	T_D	F^+
60	30	1.0000	0.9780	0.9987	0.8460
100	40	1.0000	1.0000	1.0000	0.9530
	60	1.0000	1.0000	1.0000	1.0000
	80	1.0000	1.0000	1.0000	0.9970
150	40	1.0000	1.0000	0.9317	0.9830
	60	1.0000	1.0000	1.0000	1.0000
	80	1.0000	1.0000	1.0000	1.0000
200	40	1.0000	1.0000	1.0000	0.9870
	60	1.0000	1.0000	1.0000	1.0000
	80	1.0000	1.0000	1.0000	1.0000
400	40	1.0000	1.0000	1.0000	0.9960
	60	1.0000	1.0000	1.0000	1.0000
	80	1.0000	1.0000	1.0000	1.0000

Table 2. Power, $\Sigma = D$, $D = \text{diag}\{d_1, \ldots, d_p\}$, and, $d_i \sim U(2,3)$, $i = 1, \ldots, p$.

		$\boldsymbol{\mu}_1$		$\boldsymbol{\mu}_2$	
p	n	T_D	F^+	T_D	F^+
60	30	0.6424	1.0000	0.2909	1.0000
100	40	0.9417	1.0000	0.5934	1.0000
	60	0.9979	1.0000	0.8608	1.0000
	80	1.0000	1.0000	0.9622	1.0000
150	40	0.9606	1.0000	0.6605	1.0000
	60	0.9991	1.0000	0.9127	1.0000
	80	1.0000	1.0000	0.9829	1.0000
200	40	0.9845	1.0000	0.7892	1.0000
	60	0.9998	1.0000	0.9668	1.0000
	80	1.0000	1.0000	0.9981	1.0000
400	40	1.0000	1.0000	0.9032	1.0000
	60	1.0000	1.0000	0.9951	1.0000
	80	1.0000	1.0000	0.9981	1.0000

Table 3. Power, $\Sigma = D$, $D = \text{diag}\{d_1, \ldots, d_p\}$, and, $d_i \sim \chi^2_3$, $i = 1, \ldots, p$.

		$\boldsymbol{\mu}_1$		$\boldsymbol{\mu}_2$	
p	n	T_D	F^+	T_D	F^+
60	30	0.9541	1.0000	0.5420	1.0000
100	40	0.9998	1.0000	0.9394	1.0000
	60	1.0000	1.0000	0.9983	1.0000
	80	1.0000	1.0000	1.0000	1.0000
150	40	0.9999	1.0000	0.9622	1.0000
	60	1.0000	1.0000	0.9999	1.0000
	80	1.0000	1.0000	1.0000	1.0000
200	40	1.0000	1.0000	0.9925	1.0000
	60	1.0000	1.0000	1.0000	1.0000
	80	1.0000	1.0000	1.0000	1.0000
400	40	1.0000	1.0000	0.9996	1.0000
	60	1.0000	1.0000	1.0000	1.0000
	80	1.0000	1.0000	1.0000	1.0000

In other words, the hypothesis $H : \boldsymbol{\mu} = \mathbf{0}$ is rejected if

$$\begin{aligned} F^+ &> \left\{ 1 + \left[\left(\frac{2}{n} \right)^{1/2} z_{1-\alpha}/c_{p,n} \right] \right\} \\ &= F^+_\alpha. \end{aligned}$$

Thus, ideally under the hypothesis that $\boldsymbol{\mu} = 0$, we should have

$$P\{F^+ \geq F^+_\alpha\} = \alpha$$

if the normal approximation given in Corollary 2.1 is a good approximation. Thus, in order to ascertain how good this approximation is, we do a simulation in which we calculate the statistic F^+ by taking $n+1$ samples from $N(\mathbf{0}, D)$ and replicating it 1,000 times. We then calculate

$$\frac{\#\text{ of } F^+ \geq F_\alpha^+}{1000} = \hat{\alpha}$$

and compare it with α. We call $\hat{\alpha}$ the attained significance level (ASL). We choose $\alpha = 0.05$. Tables 4–6 show the closeness of $\hat{\alpha}$ with α.

2.1. Invariance and other properties of the F^+ test

For testing the hypothesis that the mean vector $\boldsymbol{\mu}$ is equal to a zero vector against the alternative that $\boldsymbol{\mu} \neq \mathbf{0}$, Hotelling's T^2-test based on the statistic

$$F = \frac{n-p+1}{np} N \overline{\boldsymbol{x}}' S^{-1} \overline{\boldsymbol{x}} \tag{2.9}$$

Table 4. Attained significant level of F^+ test under H, sample from $N(0, I)$.

	$n=30$	$n=40$	$n=50$	$n=60$	$n=70$	$n=80$	$n=90$
$p=100$	0.077	0.062	0.057	0.058	0.078	0.098	0.117
$p=150$	0.069	0.069	0.053	0.073	0.067	0.052	0.059
$p=200$	0.053	0.052	0.053	0.047	0.056	0.048	0.039
$p=300$	0.068	0.057	0.064	0.069	0.054	0.037	0.039
$p=400$	0.071	0.064	0.053	0.067	0.053	0.048	0.048

Table 5. Attained significant level of F^+ test under H, (1) sample from $N(0, D)$. (2) $D = \text{diag}(d_1, \ldots, d_p)$, where $D_i \sim U(2, 3)$.

	$n=30$	$n=40$	$n=50$	$n=60$	$n=70$	$n=80$	$n=90$
$p=100$	0.072	0.060	0.055	0.062	0.071	0.096	0.108
$p=150$	0.053	0.047	0.057	0.048	0.052	0.046	0.060
$p=200$	0.062	0.060	0.053	0.058	0.050	0.039	0.047
$p=300$	0.068	0.067	0.064	0.052	0.053	0.068	0.058
$p=400$	0.071	0.061	0.067	0.052	0.051	0.058	0.061

Table 6. Attained significant level of F^+ test under H, (1) sample from $N(0, D)$. (2) $D = \text{diag}(d_1, \ldots, d_p)$, where $D_i \sim \chi^2$ of $2df$.

	$n=30$	$n=40$	$n=50$	$n=60$	$n=70$	$n=80$	$n=90$
$p=100$	0.049	0.032	0.014	0.008	0.018	0.015	0.030
$p=150$	0.028	0.017	0.017	0.015	0.025	0.006	0.002
$p=200$	0.019	0.024	0.030	0.018	0.001	0.002	0.010
$p=300$	0.043	0.030	0.013	0.013	0.008	0.028	0.012
$p=400$	0.055	0.042	0.033	0.021	0.018	0.018	0.002

is used when $n \geq p$, since S is positive definite with probability one and F has an F-distribution with p and $n-p+1$ degrees of freedom. The F-test in (2.9) can be interpreted in many ways. For example, $\overline{\boldsymbol{x}}' S^{-1} \overline{\boldsymbol{x}}$ is the sample (squared) distance of the sample mean vector from the zero vector. It can also be interpreted as the test based on the p sample principal components of the mean vector, since

$$\overline{\boldsymbol{x}}' S^{-1} \overline{\boldsymbol{x}} = \overline{\boldsymbol{x}}' H_0' L_0^{-1} H_0 \overline{\boldsymbol{x}},$$

where

$$S = H_0' L_0 H_0, \qquad H_0 H_0' = I_p.$$

$L_0 = \text{diag}(l_1, \ldots, l_p)$ and $H_0 \overline{\boldsymbol{x}}$ is the vector of the p sample principal components of the sample mean vector. It can also be shown to be equivalent to a test based on $(\boldsymbol{a}' \overline{\boldsymbol{x}})^2$ where $\boldsymbol{a}$ is chosen such that $\boldsymbol{a}' S \boldsymbol{a} = 1$ and $(\boldsymbol{a}' \overline{\boldsymbol{x}})^2$ is maximized.

When $n < p$, S has a singular Wishart distribution, see Srivastava (2003) for its distribution. For the singular case, a test corresponding to the F-test in (2.9) can be proposed as,

$$F^- = cN \overline{\boldsymbol{x}}' S^- \overline{\boldsymbol{x}},$$

for some constant c, where S^- is a generalized inverse of S and $SS^-S = S$. No such test has been proposed in the literature so far as it raises two obvious questions, namely which g-inverse to use and what is its distribution. For example, the $p \times p$ sample covariance matrix S can be written as

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}' & S_{22} \end{pmatrix},$$

where S_{11} can be taken as an $n \times n$ positive definite matrix. When $n < p$, it follows from Srivastava and Khatri (1979, p. 11) that $S_{22} = S_{12}' S_{11}^{-1} S_{12}$. Thus, a g-inverse of S can be taken as

$$S^- = \begin{pmatrix} S_{11}^{-1} & 0 \\ 0' & 0 \end{pmatrix},$$

see Rao (1973, p. 27), Rao and Mitra (1971, p. 208), Schott (1997), or Siotani *et al.* (1985, p. 595). In this case,

$$F^- = cN \overline{\boldsymbol{x}}_1{}' S_{11}^{-1} \overline{\boldsymbol{x}}_1,$$

where $\overline{\boldsymbol{x}}' = (\overline{\boldsymbol{x}}_1', \overline{\boldsymbol{x}}_2')$ with $\overline{\boldsymbol{x}}_1 : n \times 1$ and $\overline{\boldsymbol{x}}_2 : (p-n) \times 1$. With $c = 1/n^2$, it follows that when $\boldsymbol{\mu} = \boldsymbol{0}$,

$$F^- = \frac{1}{n^2} N \overline{\boldsymbol{x}}_1{}' S_{11}^{-1} \overline{\boldsymbol{x}}_1$$

has an F-distribution with n and 1 degrees of freedom. Thus, the distribution of $\overline{\boldsymbol{x}}_1{}' S_{11}^{-1} \overline{\boldsymbol{x}}_1$ does not depend on the covariance matrix Σ. In fact, since $A'(ASA')^- A$ is a generalized inverse of S for any $p \times p$ nonsingular matrix A,

it is possible that the distribution of $\overline{\boldsymbol{x}}'S^{-}\overline{\boldsymbol{x}}$ may not depend on the covariance matrix Σ. When more restrictions are placed on the generalized inverse, such as in the case of the Moore-Penrose inverse, then the above property may not hold since in the case of Moore-Penrose inverse $A'(ASA')^{+}A$ may not be the Moore-Penrose inverse of S. However, the statistic $\overline{\boldsymbol{x}}_1'S_{11}^{-1}\overline{\boldsymbol{x}}_1$ is not invariant under the nonsingular linear transformations $\overline{\boldsymbol{x}} \to A\overline{\boldsymbol{x}}$ and $S \to ASA'$ for any $p \times p$ nonsingular matrix A.

In fact, when $N \leq p$, no invariant test exists under the linear transformation by an element of the group Glp of non-singular $p \times p$ matrices. The sample space χ consists of $p \times N$ matrices of rank $N \leq p$, since Σ is nonsingular, and any matrix in χ can be transformed to any other matrix of χ by an element of Glp. Thus the group Glp acts transitively on the sample space. Hence the only α-level test that is affine invariant is the test $\Phi \equiv \alpha$, see Lehmann (1959, p. 318).

The generalized inverse used to obtain F^{-}, however, does not use all the information available in the data. The sufficient statistic for the above problem is $(\overline{\boldsymbol{x}}, S)$ or equivalently $(\overline{\boldsymbol{x}}; H'LH)$, where $H : n \times p$, $HH' = I_n$, and

$$nS = V = H'LH, \tag{2.10}$$

$L = \operatorname{diag}(l_1, \ldots, l_n)$, a diagonal matrix. The Moore-Penrose inverse of S is given by

$$S^{+} = nH'L^{-1}H. \tag{2.11}$$

Thus, we define the sample (squared) distance of the sample mean vector from the zero vector by

$$\begin{aligned} D^{+2} &= \overline{\boldsymbol{x}}'S^{+}\overline{\boldsymbol{x}} \\ &= n\overline{\boldsymbol{x}}'H'L^{-1}H\overline{\boldsymbol{x}}. \end{aligned} \tag{2.12}$$

It may be noted that $H\overline{\boldsymbol{x}}$ is the vector of the n principal components of the sample mean vector, and $n^{-1}L$ is the sample covariance of these n components. Thus the distance function, as in the case of nonsingular S, can also be defined in terms of the n principal components. For testing the hypothesis that $\boldsymbol{\mu} = \mathbf{0}$ against the alternative that $\boldsymbol{\mu} \neq \mathbf{0}$, we propose a test statistic based on $(\boldsymbol{a}'\overline{\boldsymbol{x}})^2$, where $\boldsymbol{a}$ belonging to the column space of H', $\boldsymbol{a} \in \rho(H')$, is chosen such that $\boldsymbol{a}'S\boldsymbol{a} = 1$ and $(\boldsymbol{a}'\overline{\boldsymbol{x}})^2$ is maximized. It will be shown next that this maximum is equal to D^{+2}.

Since $\boldsymbol{a} \in \rho(H')$, $\boldsymbol{a} = H'\boldsymbol{b}$ for some n-vector $\boldsymbol{b}$. Hence, since $HH' = I_n$

$$\begin{aligned} (\boldsymbol{a}'\overline{\boldsymbol{x}})^2 &= [\boldsymbol{b}'(HH'L^{1/2}HH'L^{-1/2})H\overline{\boldsymbol{x}}]^2 \\ &= [\boldsymbol{b}'H(S)^{1/2}(S^{+})^{1/2}\overline{\boldsymbol{x}}]^2 \\ &\leq (\boldsymbol{a}'S\boldsymbol{a})(\overline{\boldsymbol{x}}'S^{+}\overline{\boldsymbol{x}}) = D^{+2} \end{aligned}$$

from the Cauchy-Schwarz inequality. The equality holds at

$$\boldsymbol{a} = (S^{+})\overline{\boldsymbol{x}}/(\overline{\boldsymbol{x}}'S^{+}\overline{\boldsymbol{x}})^{1/2}.$$

Hence, for testing the hypothesis that $\boldsymbol{\mu} = \mathbf{0}$ against the alternative that $\boldsymbol{\mu} \neq \mathbf{0}$, we propose the test statistic

$$
\begin{aligned}
F^{+} &= \frac{\max(p,n) - \min(p,n) + 1}{n \min(p,n)} N \overline{\boldsymbol{x}}' S^{+} \overline{\boldsymbol{x}} \\
&= \frac{p - n + 1}{n^2} N \overline{\boldsymbol{x}}' S^{+} \overline{\boldsymbol{x}}, \qquad \text{if} \quad n < p, \\
&= \frac{n - p + 1}{np} N \overline{\boldsymbol{x}}' S^{-1} \overline{\boldsymbol{x}}, \qquad \text{if} \quad n \geq p,
\end{aligned} \tag{2.13}
$$

which is the same as F defined in (2.8) when the sample covariance matrix is nonsingular. Thus, when $n < p$, we propose the statistics T^{+2} or F^{+}, as defined in (2.3) and (2.4) respectively, for testing the hypothesis that $\boldsymbol{\mu} = \mathbf{0}$ vs $\boldsymbol{\mu} \neq \mathbf{0}$. We note that the statistic T^{+2} is invariant under the transformation $\boldsymbol{x}_i \to c\Gamma \boldsymbol{x}_i$, where $c \in R_{(0)}$, $c \neq 0$ and $\Gamma \in O_p$; $R_{(0)}$ denotes the real line without zero and O_p denotes the group of $p \times p$ orthogonal matrices. Clearly $c\Gamma \in Glp$. To show the invariance, we note that

$$
\begin{aligned}
\frac{T^{+2}}{n} &= N \overline{\boldsymbol{x}}' V^{+} \overline{\boldsymbol{x}} \\
&= N \overline{\boldsymbol{x}}' H' L^{-1} H \overline{\boldsymbol{x}} \\
&= N \overline{\boldsymbol{x}}' \Gamma' \Gamma H' L^{-1} H \Gamma' \Gamma \overline{\boldsymbol{x}} \\
&= N \overline{\boldsymbol{x}}' \Gamma' (\Gamma V \Gamma')^{+} \Gamma \overline{\boldsymbol{x}},
\end{aligned}
$$

since the eigenvalues of $\Gamma V \Gamma'$ are the same as that of V. The invariance under scalar transformation obviously holds.

2.2. Distribution of F^{+} when $\boldsymbol{\mu} = \mathbf{0}$

We first consider the case when the covariance matrix Σ is of rank $r \leq n$. In this case, we get the following theorem.

THEOREM 2.1. *Suppose that the covariance matrix Σ is singular of rank $r \leq n$. Then under the hypothesis that the mean vector $\boldsymbol{\mu} = \mathbf{0}$,*

$$
\frac{n - r + 1}{nr} (N \overline{\boldsymbol{x}}' S^{+} \overline{\boldsymbol{x}}) \sim F_{r, n-r+1}.
$$

The proof is given in the Appendix.

In the above theorem, it is assumed that the covariance matrix Σ is not only singular but is of rank $r \leq n$. At the moment, no statistical test is available to check this assumption. However, in practical applications, we look at the eigenvalues of the sample covariance matrix S, and delete the eigenvalues that are zero or very small, as is done in the selection of principal components.

Tests are available in Srivastava (2005) to check if $\Sigma = \gamma^2 I$, γ^2 unknown, or if $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_p)$ under the assumption that $n = O(p^{\delta})$, $o < \delta \leq 1$. A test for the first hypothesis that is of spherecity is also available in Ludoit

and Wolf (2002) under the assumption that $n = O(p)$. More efficient tests can be constructed based on the statistics $(N\overline{\boldsymbol{x}}'\overline{\boldsymbol{x}}/\operatorname{tr} S)$, and $N\overline{\boldsymbol{x}}' D_S^{-1}\overline{\boldsymbol{x}}$, depending upon which of the above two hypotheses is true; here $D_s = \operatorname{diag}(s_{11}, \ldots, s_{pp})$, $S = (s_{ij})$. Thus the F^+ test may not be used when the covariance matrix is either a constant times an identity matrix or a diagonal matrix. Nevertheless, we derive the distribution of the F^+ test when $\Sigma = \gamma^2 I$, and γ^2 is unknown because in this case we obtain an exact distribution which may serve as a basis for comparison when only asymptotic or approximate distributions are available.

THEOREM 2.2. *Let the F^+ statistic be as defined in* (2.4). *Then when the covariance matrix $\Sigma = \gamma^2 I$, and γ^2 is unknown, the F^+ statistic is distributed under the hypothesis H, as an F-distribution with n and $p - n + 1$ degrees of freedom, $n \le p$.*

To derive the distribution of the F^+ statistic when $\Sigma \ne \gamma^2 I$ and $n < p$, we first note that for any $p \times p$ orthogonal matrix $G, GG' = I$, the statistic defined by ${T^+}^2$ is invariant under linear transformations, $\boldsymbol{x}_i \to G\boldsymbol{x}_i$. Hence, we may assume without any loss of generality that

$$\Sigma = \wedge = \operatorname{diag}(\lambda_1, \ldots, \lambda_p).$$

Let

$$\boldsymbol{z} = N^{1/2} A H \overline{\boldsymbol{x}}, \qquad \text{where} \quad A = (H \wedge H')^{-1/2}. \tag{2.14}$$

Then under the hypothesis H,

$$\boldsymbol{z} \sim N_n(\boldsymbol{0}, I),$$

and

$$\begin{aligned} \left(\frac{{T^+}^2}{n}\right) &= N\overline{\boldsymbol{x}}' V^+ \overline{\boldsymbol{x}} \\ &= \boldsymbol{z}'(ALA)^{-1}\boldsymbol{z}. \end{aligned} \tag{2.15}$$

The n eigenvalues $l_1, \ldots, l_n$ of the diagonal matrix L are the n non-zero eigenvalues of $V = YY'$, where the n columns of the $p \times n$ matrix Y are iid $N_p(\boldsymbol{0}, \wedge)$.

To derive the asymptotic results in which p may also go to infinity, we assume that

$$0 < a_{i,0} = \lim_{p\to\infty} a_i < \infty, \qquad i = 1, \ldots, 4 \tag{2.16}$$

where

$$a_i = (\operatorname{tr} \Sigma^i / p). \tag{2.17}$$

Let

$$b = (a_1^2 / a_2). \tag{2.18}$$

Under the assumption (2.16), consistent and unbiased estimators of a_2 and a_1, as $(n,p) \to \infty$, are given by

$$\hat{a}_2 = \frac{n^2}{(n-1)(n+2)} \left[\left(\frac{\operatorname{tr} S^2}{p} \right) - \frac{p}{n} \left(\frac{\operatorname{tr} S}{p} \right)^2 \right] \tag{2.19}$$

and

$$\hat{a}_1 = (\operatorname{tr} S/p) \tag{2.20}$$

respectively, see Srivastava (2005). Thus, clearly $(\operatorname{tr} S^2/p)$ is not a consistent estimator of a_2 unless p is fixed and $n \to \infty$. Thus, under the assumption (2.16), a consistent estimator of b as $(n,p) \to \infty$ is given by

$$\hat{b} = \left(\frac{\hat{a}_1^2}{\hat{a}_2} \right). \tag{2.21}$$

In the next theorem, we give an asymptotic distribution of the F^+ statistic, the proof of which is given in the Appendix.

THEOREM 2.3. *Under the condition* (2.16) *and when* $\boldsymbol{\mu} = \mathbf{0}$,

$$\lim_{n,p \to \infty} P \left[\left(\frac{n}{2} \right)^{1/2} (p(p-n+1)^{-1} \hat{b} F^+ - 1) \le z_{1-\alpha} \right] = \Phi(z_{1-\alpha}).$$

COROLLARY 2.1. *Let* $n = O(p^\delta)$, $0 < \delta < 1$, *then under the condition* (2.16) *and when* $\boldsymbol{\mu} = \mathbf{0}$,

$$\lim_{n,p \to \infty} P \left[c_{p,n} \left(\frac{n}{2} \right)^{1/2} (\hat{b} F^+ - 1) < z_{1-\alpha} \right] = \Phi(z_{1-\alpha})$$

where

$$c_{p,n} = [(p-n+1)/(p+1)]^{1/2}.$$

In the following theorem, it is shown that $F^+ > F_{n,p-n+1}$ with probability one, see the Appendix for its proof.

THEOREM 2.4. *Let* $F_{n,p-n+1}(\alpha)$ *be the upper* 100% *point of the* F*-distribution with* n *and* $(p-n+1)$ *degrees of freedom. Then*

$$P\{F^+ > F_{n,p-n+1}(\alpha)\} \ge \alpha.$$

2.3. Asymptotic distribution of F^+ when $\boldsymbol{\mu} \neq \mathbf{0}$

In this section, we obtain the asymptotic distribution under the local alter-

native,

$$\boldsymbol{\mu} = E(\overline{\boldsymbol{x}}) = \left(\frac{1}{nN}\right)^{1/2} \boldsymbol{\delta}. \tag{2.22}$$

From (2.15)

$$\frac{T^{+2}}{n} = \boldsymbol{z}'(ALA)^{-1}\boldsymbol{z},$$

where given H,

$$\boldsymbol{z} \sim N_n(n^{-1/2}AH\boldsymbol{\delta}, I), \qquad A = (H \wedge H')^{-1/2}.$$

Let Γ be an $n \times n$ orthogonal matrix, $\Gamma\Gamma' = I_n$, with the first row of Γ given by

$$\boldsymbol{\delta}H'A/(\boldsymbol{\delta}'H'A^2H\boldsymbol{\delta})^{1/2}.$$

Then, given H,

$$\boldsymbol{w} = \Gamma\boldsymbol{z} \sim N_n\left(\begin{pmatrix}\theta\\ \mathbf{0}\end{pmatrix}, I\right),$$

where

$$\theta = (n^{-1}\boldsymbol{\delta}'H'A^2H\boldsymbol{\delta})^{1/2}. \tag{2.23}$$

As before,

$$\lim_{p\to\infty}(pbT^{+2}/n) = \lim_{p\to\infty}\boldsymbol{z}'\boldsymbol{z} = \lim_{p\to\infty}\boldsymbol{w}'\boldsymbol{w}.$$

We note that from Lemma A.2, given in the Appendix,

$$\begin{aligned}\lim_{n,p\to\infty}\theta^2 &= \lim_{n,p\to\infty}(n^{-1}\boldsymbol{\delta}'H'A^2H\boldsymbol{\delta}), \qquad A = (H\wedge H')^{-1/2}\\ &= \lim_{n,p\to\infty}\left(\frac{\boldsymbol{\delta}'\wedge\boldsymbol{\delta}}{pa_2}\right) = \theta_0^2.\end{aligned}$$

We shall assume that $\theta_0^2 < \infty$. Thus,

$$\begin{aligned}&\lim_{n,p\to\infty} P\left\{\frac{(pbT^{+2}/n) - n}{\sqrt{2n}} > z_{1-\alpha}\right\}\\ &\quad= \lim_{n,p\to\infty} P\left\{\frac{(pbT^{+2}/n) - n - \theta_0^2}{\sqrt{2n+4\theta_0^2}}\left(\frac{2n+4\theta_0^2}{2n}\right)^{1/2} > z_{1-\alpha} - \frac{\theta_0^2}{\sqrt{2n}}\right\}\\ &\quad= \lim_{n,p\to\infty}\Phi[-z_{1-\alpha} + \theta_0^2(2n)^{-1/2}].\end{aligned}$$

Hence, the power of the F^+ test is given by

$$\beta(F^+) \simeq \Phi\left[-z_{1-\alpha} + (n/p)(n/2)^{1/2}\frac{\boldsymbol{\mu}'\wedge\boldsymbol{\mu}}{a_2}\right]. \tag{2.24}$$

2.4. Dempster's test

We assume as before that $\boldsymbol{x}_i$ are iid $N_p(\boldsymbol{\mu}, \Sigma)$. Suppose $\Sigma = \sigma^2 I$, σ^2 unknown. Then, the sufficient statistics are $\overline{\boldsymbol{x}}$ and $(\operatorname{tr} S/p)$. The problem remains invariant under the transformation $\boldsymbol{x} \to c\Gamma\boldsymbol{x}_i$, where $c \neq 0$, and $\Gamma\Gamma' = I_p$. The maximal invariant statistic is given by

$$T_D = \frac{N\overline{\boldsymbol{x}}'\overline{\boldsymbol{x}}}{(\operatorname{tr} S/p)}. \tag{2.25}$$

The maximal invariant in the parameter space is given by

$$\gamma = (\boldsymbol{\mu}'\boldsymbol{\mu}/\sigma^2).$$

T_D is distributed as an F-distribution with p and np degrees of freedom and noncentrality parameter γ. Thus T_D is the uniformly most powerful invariant test. This is also the likelihood ratio test. But when $\Sigma \neq \sigma^2 I$, then T_D has none of the properties mentioned above. Dempster (1958), nevertheless proposed the same test F_D when $\Sigma \neq \sigma^2 I$, since it can be computed for all values of p irrespective of whether $n \leq p$ or $n > p$. To obtain the distribution of F_D, we note that

$$F_D = \frac{Q_1}{Q_2 + \cdots + Q_N},$$

where Q_i's are independently and identically distributed but not as a chi-square χ^2 random variable. Dempster (1958) approximated the distribution of Q_i as $m\chi_r^2$, where r is the degrees of freedom associated with χ^2 and m is a constant. Since F_D does not depend on m, he gave two equations from which an iterative solution of r can be obtained. Alternatively, since

$$E(m\chi_r^2) = mr = \operatorname{tr}\Sigma,$$

and

$$\operatorname{var}(m\chi_r^2) = 2m^2 r = 2\operatorname{tr}\Sigma^2,$$

r is given by

$$r = \frac{(\operatorname{tr}\Sigma)^2}{(\operatorname{tr}\Sigma^2)} = p\frac{(\operatorname{tr}\Sigma/p)^2}{(\operatorname{tr}\Sigma^2/p)} = pb.$$

A consistent estimator of b has been given in (2.21). Thus, r can be estimated by

$$\hat{r} = p\hat{b}.$$

Hence, an approximate distribution of F_D under the hypothesis that $\boldsymbol{\mu} = \mathbf{0}$ is given by an F-distribution with $[\hat{r}]$ and $[n\hat{r}]$ degrees of freedom, where $[a]$ denotes the largest integer $\leq a$. The asymptotic distribution of the F_D test under the alternative that $\boldsymbol{\mu} \neq \mathbf{0}$, is given by

$$\lim_{n,p\to\infty}\left[P\left\{\left(\frac{r}{2}\right)^{1/2}(F_D - 1) > z_{1-\alpha} \mid \boldsymbol{\mu} = (nN)^{-1/2}\boldsymbol{\delta}\right\} - \Phi\left(-z_{1-\alpha} + \frac{n\boldsymbol{\mu}'\boldsymbol{\mu}}{\sqrt{2pa_2}}\right)\right] = 0,$$

see Bai and Saranadasa (1996). This also gives the asymptotic distribution of Dempster's test under the hypothesis.

2.5. Bai-Saranadasa test (BS test)

Bai and Saranadasa (1996) proposed a standardized version of the Dempster test which is much easier to handle in order to obtain the asymptotic null and non-null distributions, with performance almost identical to the Dempster test in all simulation results given by Bai and Saranadasa (1996). The test statistic proposed by Bai and Saranadasa (1996) is given by

$$T_{BS} = \frac{N\overline{\boldsymbol{x}}'\overline{\boldsymbol{x}} - \operatorname{tr} S}{(p\hat{a}_2)^{1/2}(2(n+1)/n)^{1/2}}, \tag{2.26}$$

where

$$p\hat{a}_2 = \frac{n^2}{(n+2)(n-1)}\left[\operatorname{tr} S^2 - \frac{1}{n}(\operatorname{tr} S)^2\right] \equiv \frac{n^2}{(n+2)(n-1)}\hat{C},$$

and is an unbiased and ratio consistent estimator of $\operatorname{tr}\Sigma^2$. This test is also invariant under the transformation $\boldsymbol{x} \to c\Gamma\boldsymbol{x}_i$, $c \neq 0$, $\Gamma\Gamma' = I$ as was the Dempster test T_D. To obtain the distribution of T_{BS}, we need the following lemma.

LEMMA 2.1. *Let a_{in} be a sequence of constants such that*

$$\lim_{n\to\infty} \max_{1\le i\le n}(a_{in}^2) = 0, \qquad \textit{and} \qquad \lim_{n\to\infty}\sum_{i=1}^{n} a_{in}^2 = 1.$$

Then for any iid random variables u_i with mean zero and variance one,

$$\lim_{n\to\infty} P\left[\sum_{i=1}^{n} a_{in}u_i < z\right] = \Phi(z),$$

see Srivastava (1970) *or Srivastava* (1972, *Lemma* 2.1).

Thus, if

$$\lim_{p\to\infty} \frac{\max_{1\le i\le p}\lambda_i}{(\operatorname{tr}\Sigma^2)^{1/2}} = 0, \tag{2.27}$$

then

$$\lim_{n,p\to\infty} P\{T_{BS} \le z\} = \Phi(z).$$

It may be noted that the condition (2.27) will be satisfied if

$$\lambda_i = O(p^\gamma), \qquad 0 \le \gamma < \frac{1}{2} \tag{2.28}$$

since a_2 is finite. To obtain the power of the test, we need to obtain the distribution of the statistic T_{BS} under the alternative hypothesis $\boldsymbol{\mu} \neq \boldsymbol{0}$. As before, we consider the local alternatives in which

$$E(\overline{\boldsymbol{x}}) = \boldsymbol{\mu} = (1/nN)^{1/2}\boldsymbol{\delta}.$$

From the above results, we get under condition (2.28),

$$\lim_{n,p\to\infty}\left[\frac{N[\overline{\boldsymbol{x}}-(nN)^{-1/2}\boldsymbol{\delta}]'[\overline{\boldsymbol{x}}-(nN)^{-1/2}\boldsymbol{\delta}]-\operatorname{tr}S}{(2\operatorname{tr}\Sigma^2)^{1/2}}<z\right]=\Phi(z).$$

Since,

$$\lim_{n,p\to\infty}\operatorname{var}\left(\frac{(N/n)^{1/2}\boldsymbol{\delta}'\overline{\boldsymbol{x}}}{\sqrt{2\operatorname{tr}\Sigma^2}}\right)=\lim_{n,p\to\infty}\left(\frac{\boldsymbol{\delta}'\wedge\boldsymbol{\delta}}{2n\operatorname{tr}\Sigma^2}\right)=\lim_{n,p\to\infty}\left(\frac{\theta_0^2}{2n}\right)=0,$$

it follows that

$$\lim_{n,p\to\infty}\Bigg[P\left\{\frac{(N\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}')-\operatorname{tr}S}{\sqrt{2p\hat{a}_2}}\left(\frac{\sqrt{p\hat{a}_2}}{\sqrt{\operatorname{tr}\Sigma^2}}\right)>z_{1-\alpha}-\frac{\boldsymbol{\delta}'\boldsymbol{\delta}}{n\sqrt{2pa_2}}\right\}-\Phi\left(z_{1-\alpha}+\frac{\boldsymbol{\delta}'\boldsymbol{\delta}}{n\sqrt{2pa_2}}\right)\Bigg]=0.$$

Thus, the asymptotic power of the BS test under condition (2.28) is given by

$$\beta(BS)\simeq\Phi\left(-z_{1-\alpha}+\frac{n\boldsymbol{\mu}'\boldsymbol{\mu}}{\sqrt{2pa_2}}\right). \tag{2.29}$$

2.6. Läuter, Glimm and Kropf test

Läuter *et al.* (1998) proposed a test based on principal components. The components are, obtained by using the eigenvectors corresponding to the non-zero eigenvalues of $V+N\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}'$, the complete sufficient statistic under the hypothesis. In place of using the sample covariance of the principal components, they use O_1SO_1', where $V+N\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}'=O'\tilde{L}O$, $OO'=I_{n+1}$, $O'=(O_1',O_2')$, and $\tilde{L}=\operatorname{diag}(\tilde{l}_1,\ldots,\tilde{l}_{n+1})$, $\tilde{l}_1>\cdots,\tilde{l}_{n+1}$. The matrix O_1' is of $p\times k$ dimension, $k\le n$. When O_1' is a $p\times k$ matrix such that O_1SO_1' is a positive definite matrix with probability one, then it follows from Läuter *et al.* (1998) that the statistic

$$T_{LGK}=\frac{n-k+1}{nk}N\overline{\boldsymbol{x}}'O_1'(O_1SO_1')^{-1}O_1\overline{\boldsymbol{x}}\sim F_{k,n-k+1},\qquad k\le n.$$

The non-null distribution of this statistic is not available. A comparison by simulation also appears difficult as no guidance is available as to how many components to include. For example, if we choose only one component from the $n+1$ components, denoted by $\boldsymbol{a}$, then one may use the t^2-test given by

$$t^2=N(\boldsymbol{a}'\overline{\boldsymbol{x}})^2/(\boldsymbol{a}'S\boldsymbol{a}),$$

where $\boldsymbol{a}$ is a function of the statistic $V+N\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}'$. This has a t^2-distribution with n degrees of freedom. Such a test however, has a poor power when the components of the mean vector do not shift in the same direction and the shifts are not of equal magnitude; see Srivastava *et al.* (2001) for some power comparison.

3. Two-sample F^+ test

In this section, we consider the problem of testing that the mean vector $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of two populations with common covariance Σ are equal. For this, we shall define the sample (squared) distance between the two populations with sample mean vectors $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ and pooled sample covariance matrix S by

$$D^{+2} = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' S^+ (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2).$$

Thus, for testing the equality of the two mean vectors, the test statistic F^+ becomes

$$F^+ = \frac{p-n+1}{n^2}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' S^+ (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2), \qquad n \le p,$$

where $\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1N_1}$ are iid $N_p(\boldsymbol{\mu}_1, \Sigma)$, $\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2N_2}$ are iid $N_p(\boldsymbol{\mu}_2, \Sigma)$, $\Sigma > 0$, $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ are the sample mean vectors,

$$nS = \sum_{i=1}^{N_1}(\boldsymbol{x}_{1i} - \overline{\boldsymbol{x}}_1)(\boldsymbol{x}_{1i} - \overline{\boldsymbol{x}}_1)' + \sum_{i=1}^{N_2}(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2)(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2)',$$

and

$$n = N_1 + N_2 - 2.$$

All the results obtained in Theorems 2.1 to 2.4 for the one-sample case are also available for the two-sample case, except that T^{+2} is now defined as

$$T^{+2} = \frac{N_1 N_2}{N_1 + N_2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' S^+ (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2).$$

4. Multivariate analysis of variance (MANOVA)

For MANOVA, that is for testing the equality of k mean vectors, we shall denote the independently distributed observation vectors by $\boldsymbol{x}_{ij}$, $j = 1, \ldots, N_i$, $i = 1, \ldots, k$ where $\boldsymbol{x}_{ij} \sim N(\boldsymbol{\mu}_i, \Sigma)$, $\Sigma > 0$. The between sum of squares will be denoted by

$$B = \sum_{i=1}^{k} N_i (\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})',$$

where $\overline{\boldsymbol{x}}_i$ are the sample mean vectors,

$$\overline{\boldsymbol{x}} = \left(\sum_{i=1}^{k} N_i \overline{\boldsymbol{x}}_i / N\right), \qquad N = N_1 + \cdots + N_k,$$

and within sum of squares or sum of squares due to error will be denoted by

$$V = nS = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)',$$

where $n = N - k$.

The test statistic we propose for testing the hypothesis that $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k$ against the alternative that $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$ for at least one pair of $(i, j, i \neq j)$ is given by

$$U^+ = \prod_{i=1}^{m}(1 + d_i)^{-1} = |I + BV^+|^{-1},$$

where d_i are the non-zero eigenvalues of the matrix BV^+, $m = k - 1 \leq n$, and V^+ is the Moore-Penrose inverse of V. To obtain the distribution of U^+ under the hypothesis that all the mean vectors are equal, we note that the between sum of squares B defined above can be written as

$$B = UU',$$

where $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$, $m = k - 1$ and is distributed as $N_{p,m}(\theta, \Sigma, I_m)$ where $\theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ is a $p \times m$ matrix and a random matrix U is said to have $N_{p,m}(\theta, \Sigma, C)$ if the pdf of U can be written as

$$(2\pi)^{-(1/2)pm}|C|^{-(1/2)p}|\Sigma|^{-(1/2)m}\, etr\left(-\frac{1}{2}\Sigma^{-1}[(U-\theta)C^{-1}(U-\theta)']\right),$$

where $etr(A)$ stands for the exponential of the trace of the matrix A, see S & K (1979, p. 54). Under the hypothesis of the equality of the k mean vectors $\theta = 0$ and $U \sim N_{p,m}(0, \Sigma, I_m)$.

Similarly, the sum of squares due to error given in (2.19) can be written as

$$V = nS = YY',$$

where $Y : p \times n$, the n columns of Y are iid $N_p(\mathbf{0}, \Sigma)$. It is also known that B and V are independently distributed. Let

$$HVH' = L = \text{diag}(l_1, \ldots, l_n), \qquad n \leq p,$$

where $HH' = I_n$, $H : n \times p$. Hence, the non-zero eigenvalues of BV^+ are the non-zero eigenvalues of

$$U'H'A(ALA)^{-1}AHU = Z'(ALA)^{-1}Z, \tag{4.1}$$

where

$$A = (H\Sigma H')^{-1/2}, \tag{4.2}$$

and the m columns of Z are iid $N_n(\mathbf{0}, I)$ under the hypothesis that $\boldsymbol{\mu} = 0$.

From Srivastava and von Rosen (2002), the results corresponding to the case when Σ is singular becomes available. Thus, we get the following theorem.

THEOREM 4.1. *Suppose the covariance matrix Σ is singular of rank $r \leq n$. Let $d_1, \ldots, d_m$ be the non-zero eigenvalues of BV^+, where $V^+ = H'L^{-1}H$, $L =$*

$\operatorname{diag}(l_1,\dots,l_r)$, $H: r\times p$, $HH' = I_r$ *and* $m = k-1 \le r$. *Then, in the notation of Srivastava* (2002, p. XV),

$$U^+ = \prod_{i=1}^{m}(1+d_i)^{-1} \sim U_{r,m,n},$$

and, under the hypothesis that $\mu = 0$, r *fixed and* $n\to\infty$,

$$\begin{aligned} &P\left\{-\left(n-\frac{1}{2}(r-m+1)\right)\log U_{r,m,n} > c\right\} \\ &\quad = P\{\chi^2_{rm} > c\} + n^{-2}\eta[P(\chi^2_{rm+4} > c) - P(\chi^2_{rm} > c)] + O(n^{-4}), \end{aligned} \tag{4.3}$$

where $\eta = rm(r^2+m-5)/48$ *and* χ^2_f *denotes a chi-square random variable with* f *degrees of freedom.*

When $\Sigma = \gamma^2 I$, $A = I$, and hence we get the following theorem.

THEOREM 4.2. *Suppose the covariance matrix* $\Sigma = \gamma^2 I$, γ *is unknown and* $m \le n < p$, *then* U^+ *has the distribution of* $U_{n,m,p}$. *Thus, the asymptotic expression of* $P\{-l\log U^+ \le z\}$ *under the hypothesis that* $\mu = 0$, *is given by*

$$\begin{aligned} P\{-l\hat{b}\log U^+ \le z\} = {} & P\{\chi^2_g \le z\} + p^{-2}\beta[P(\chi^2_{g+4}\le z) - P(\chi^2_g \le z)] \\ & + O(p^{-4}), \end{aligned} \tag{4.4}$$

where n *fixed,* $p\to\infty$ *and* $g = nm$,

$$l = p - \frac{1}{2}(n-m+1), \qquad m = k-1, \qquad \text{and} \qquad \beta = \frac{g}{48}(n^2+m-5).$$

THEOREM 4.3. *Under the null hypothesis and* (2.16)

$$\lim_{n,p\to\infty} P\left[\frac{-p\hat{b}\log U^+ - mn}{\sqrt{2mn}} < z_{1-\alpha}\right] = \Phi(z_{1-\alpha}).$$

For some other tests and power, see Srivastava and Fujikoshi (2006), *and when* $(p/n)\to c$, $c\in(0,\infty)$, *see Fujikoshi et al.* (2004).

5. Confidence intervals

When a hypothesis is rejected, it is desirable to find out which component or components may have caused the rejection. In the context of DNA microarrays, it will be desirable to find out which genes have been affected after the radiation treatment. Since p is very large, the confidence intervals obtained by the Bonferroni inequality, which corresponds to the maximum of t-tests, are of no practical value as they give very wide confidence intervals.

As an alternative, many researchers such as Efron *et al.* (2001) use the 'False Discovery Rate', FDR, proposed by Benjamini and Hockberg (1995), although

the conditions required for its validity may not hold in many cases. On the other hand, we may use approximate confidence intervals for selecting the variables or to get some idea as to the variable or variables that may have caused the rejection of the hypothesis. These kind of confidence intervals use the fact that

$$\sup_{\boldsymbol{a}\in R^p} g^2(\boldsymbol{a}) < c \Rightarrow \sup_{\boldsymbol{a}\in \rho(H')} g^2(\boldsymbol{a}) < c,$$

for any real valued function $g(\cdot)$ of the p-vector $\boldsymbol{a}$, where $H' : p \times n$, and R^p is the p-dimensional Euclidean space. Thus, it gives a confidence coefficient less than $100(1-\alpha)\%$. However, we shall assume that they are approximately equal.

5.1. Confidence intervals in one-sample

Let

$$A_\alpha^2 = [n + (2n)^{1/2} z_{1-\alpha}], \tag{5.1}$$

where $z_{1-\alpha}$ denotes the upper $100\alpha\%$ point of the standard normal distribution. It may be noted that $A_\alpha^2 \simeq \chi^2_{n,\alpha}$. In the notation and assumptions of Section 2, approximate simultaneous confidence intervals for linear combinations $\boldsymbol{a}'\boldsymbol{\mu}$ at $100(1-\alpha)\%$ confidence coefficient are given by

$$\boldsymbol{a}'\overline{\boldsymbol{x}} \pm N^{-1/2}(n/p)^{1/2}\hat{b}^{-1/2}(\boldsymbol{a}'S\boldsymbol{a})^{1/2}A_\alpha,$$

when $\boldsymbol{a}'S\boldsymbol{a} \neq 0$ and $\boldsymbol{a}\in \rho(H')$. Thus, approximate simultaneous confidence intervals for the components μ_i of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ with confidence coefficient approximately $100(1-\alpha)\%$ are given by

$$\overline{x}_i \pm N^{-1/2}(n/p)^{1/2}\hat{b}^{-1/2}s_{ii}^{1/2}A_\alpha, \qquad i = 1, \ldots, p,$$

where $\overline{\boldsymbol{x}} = (\overline{x}_1, \ldots, \overline{x}_p)'$, and

$$S = (s_{ij}).$$

5.2. Confidence intervals in two-samples

In the two-sample case and again under the assumption made in the beginning of this section, approximate simultaneous confidence intervals for the linear combinations $\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ at $100(1-\alpha)\%$ confidence coefficient are given by

$$\boldsymbol{a}'(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \pm \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{1/2} (n/p)^{1/2}\hat{b}^{-1/2}(\boldsymbol{a}'S\boldsymbol{a})^{1/2}A_\alpha,$$

when $\boldsymbol{a}'S\boldsymbol{a} \neq 0$ and $\boldsymbol{a}\in \rho(H')$. Here $n = N_1 + N_2 - 2$, and S is the pooled estimate of the covariance matrix Σ. Thus, approximate simultaneous confidence intervals for the differences in the components of the means $\mu_{1i} - \mu_{2i}$ are given by

$$(\overline{x}_{1i} - \overline{x}_{2i}) \pm \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{1/2} (n/p)^{1/2}\hat{b}^{-1/2}(s_{ii})^{1/2}A_\alpha^{1/2}, \qquad i = 1, \ldots, p,$$

where

$$\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})', \qquad \boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2p})',$$
$$\boldsymbol{x}_1 = (\overline{x}_{11}, \ldots, \overline{x}_{1p})', \qquad \boldsymbol{x}_2 = (\overline{x}_{21}, \ldots, \overline{x}_{2p})',$$

and

$$S = (s_{ij}).$$

5.3. Confidence intervals in MANOVA

From the results of Section 2, it follows that the eigenvalues of V^+B, under the hypothesis of the equality of means, are asymptotically the eigenvalues of $p^{-1}U$, where $U \sim W_n(m, I)$, $m = k - 1$. Approximately, this is equal to the eigenvalues of $W^{-1}U$, where $W \sim W_n(p, I)$ for large p. With this result in mind, we now proceed to obtain the confidence intervals in MANOVA.

In the MANOVA, we are comparing k means. Thus, we need simultaneous confidence intervals for $\sum_{j=1}^k \boldsymbol{a}'\boldsymbol{\mu}_j q_j$, where $\sum_{j=1}^k q_j = 0$, that is q_j's for m contrasts. These approximate $100(1-\alpha)\%$ simultaneous confidence intervals for the contrasts in the means $\boldsymbol{a}'(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)\boldsymbol{q}$, $\boldsymbol{q} = (q_1, \ldots, q_k)'$ are given by

$$\sum_{j=1}^k \boldsymbol{a}'\overline{\boldsymbol{x}}_j q_j \pm \left[\left(\sum_{j=1}^k \frac{b_j^2}{N_j}\right)\left(\frac{\boldsymbol{a}'V\boldsymbol{a}}{p}\right)\left(\frac{c_\alpha}{1-c_\alpha}\right)\right]^{1/2} \hat{b}^{-1/2},$$

where V is defined in the beginning of Section 4, and $\frac{c_\alpha}{1-c_\alpha}$ is the upper $100\alpha\%$ point of the distribution of the largest eigenvalue of $W^{-1}U$. The value of c_α can be obtained from Table B.7 in Srivastava (2002) with $p_0 = n$, $m_0 = k - 1$, and n corresponds to p. The vector $\boldsymbol{a}$ of constants can be chosen as in the previous two subsections.

6. Testing the equality of covariances

For testing the equality of the two covariance matrices, let S_1 and S_2 be the sample covariances based on n_1 and n_2 degrees of freedom. Then a test for the equality of the two covariance matrices can be based on the test statistic

$$T_2^{(1)} = (p\hat{b}\,\mathrm{tr}\, V_1^+ V_2 - n_1 n_2)/(2n_1 n_2)^{1/2} \tag{6.1}$$

where $V_1 = n_1 S_1 = H_1' L_1 H_1$, $H_1 H_1' = I_{n_1}$, $V_2 = n_2 S_2$ and $V_1^+ = H_1' L_1^{-1} H_1$. The statistic T_2 is distributed asymptotically as $N(0, 1)$ under the hypothesis that the two covariances are equal. The estimate $\hat{b}$ of b can be obtained from the pooled estimate. For the equality of the k covariances in MANOVA, let $S_i = n_i^{-1} V_i$ be the sample covariances of the k populations based on n_i degrees of freedom, $i = 1, \ldots, k$. Then, a test for the equality of the k covariances can be based on the test statistic.

$$T_3^{(1)} = [p\hat{b}\,\mathrm{tr}\, V_1 V_{(1)}^+ - n_1 n_{(1)}]/\sqrt{2n_1 n_{(1)}} \tag{6.2}$$

where $V_{(1)} = \sum_{\alpha=2}^{k} V_\alpha$, and $n_{(1)} = n_2 + \cdots + n_k$. It can be shown that under the hypothesis of equality of all covariances, T_3 is asymptotically distributed as $N(0,1)$. Alternatively, we may use the approximate distribution of

$$P_2 = p\hat{b}(\operatorname{tr} V_1^+ V_2), \tag{6.3}$$

and

$$P_3 = p\hat{b}\operatorname{tr}(V_1 V_{(1)}^+) \tag{6.4}$$

which are approximately distributed as $\chi^2_{n_1 n_2}$ and $\chi^2_{n_1 \times n_{(1)}}$ respectively. The tests in (6.1) and (6.2) are arbitrary. To make them less arbitrary, we may define

$$T_2^{(i)} = [p\hat{b}\operatorname{tr} V_i^+ V_{(3-i)} - n_1 n_2]/(2n_1 n_2)^{1/2}, \qquad i = 1, 2,$$

and

$$T_3^{(i)} = [p\hat{b}\operatorname{tr} V_i V_{(i)}^+ - n_i n_{(i)}]/(2n_i n_{(i)})^{1/2}, \qquad i = 1, \ldots, k.$$

Then, conservative tests based on

$$R_2 = \max(|T_2^{(1)}|, |T_2^{(2)}|) \tag{6.5}$$

for testing the equality of two covariance matrices, and

$$R_3 = \max(|T_3^{(1)}|, \ldots, |T_3^{(k)}|) \tag{6.6}$$

for testing the equality of k covariance matrices may be considered. The hypothesis of the equality of the two covariance matrices is rejected if

$$R_2 > z_{1-\alpha/4}$$

and the hypothesis of equality of k covariance matrices is rejected if

$$R_3 > z_{1-\alpha/2k}$$

where $z_{1-\alpha}$ is the upper $100\alpha\%$ point of the $N(0,1)$ distribution.

7. Selection of variables

In order to have good power for any test in multivariate analysis, it is important to have as few characteristics as possible. Especially those variables that do not have discriminating ability, should be removed. We begin with $\hat{b}_{n,n}$ based on the n characteristics chosen for the three cases, as n largest value of $N\overline{x}_r^2/S_{r,r}$, $[N_1 N_2/(N_1 + N_2)]\ (\overline{x}_{1r} - \overline{x}_{2r})^2/S_{rr}$, and $(\boldsymbol{a}_r' B \boldsymbol{a}_r)/(\boldsymbol{a}_r' S \boldsymbol{a}_r)$, where $r = 1, \ldots, p$ and $\boldsymbol{a}_r = (0 \ldots 0, 1, 0 \ldots 0)'$ is a p-vector with all zeros except the r-th place which is one. Let p^* be the number of characteristics for which

$$(p/n)\hat{b}_{n,n}(N\overline{x}_r^2/S_{rr}) > A_\alpha^2, \tag{7.1}$$

$$(p/n)\hat{b}_{n,n}[N_1 N_2/(N_1 + N_2)](\overline{x}_{1r} - \overline{x}_{2}r)^2 > A_\alpha^2, \tag{7.2}$$

and

$$(p\hat{b}_{n,n})(\boldsymbol{a}_r' B \boldsymbol{a}_r)/(\boldsymbol{a}_r' S \boldsymbol{a}_r) > c_\alpha/(1-c_\alpha), \tag{7.3}$$

for the three cases respectively where A_α^2 is defined in (5.1) and c_α in Subsection 5.3. All the testing, procedures, and confidence intervals etc, may now be carried out with N observations and p^* characteristics. Other values of p^* around it may also be tried for seperation of groups etc. If the selected p^* is less than n, then the usual multivariate methods apply. However, if $p^* \simeq n$ or $p^* > n$, then the methods proposed in this paper apply. This method of selection of variables is illustrated in Example 8.2 in Section 8.

8. An example

Alon *et al.* (1999) used Affymetrix oligonucleotide arrays to monitor absolute measurements on expressions of over 6500 human gene expressions in 40 tumour and 22 normal colon tissues. Alon *et al.* considered only 2000 genes with highest minimal intensity across the samples. We therefore consider only data on these 2000 genes on 40 patients and 22 normal subjects.

To analyze the data, we first calculate the pooled sample covariance matrix assuming that the two populations of tumour and normal tissues have the same covariance matrix Σ. Although, this assumption can be ascertained by the test statistic given in Section 6, it may not be necessary at this stage as we will be testing for it after the selection of variables. From the pooled sample covariance matrix S, we found that $b_{60,60} = 0.1485$. Using the formula (7.2) in Section 7, we select 103 characteristics. For testing the equality of means, we calculated $b_{60,103} = 0.110$, and

$$\left(\frac{N_1 N_2}{N_1+N_2}\right)\left(\frac{p}{n}\right)\hat{b}_{60,103}(\overline{\boldsymbol{x_1}} - \overline{\boldsymbol{x_2}})' S^+ (\overline{\boldsymbol{x_1}} - \overline{\boldsymbol{x_2}}) = 94.760. \tag{8.1}$$

Using the approximation that $A_\alpha^2 \simeq \chi_{n,\alpha}^2$, we find that the value of $\chi_{60,0.05}^2 = 79.08$. Hence, the p-value is 0.0028. Thus, the hypothesis of the equality of the two means is rejected. Similarly, for testing the equality of the covariances of the two populations, we calculate the sample covariances of the two populations, S_1 and S_2 on 39 and 21 degrees of freedom respectively. The value of $P_2 = p\hat{b}_{60,103}\,\mathrm{tr}\, V_1^+ V_2 = 523.01$, where $V_1 = 39S_1$ and $V_2 = 21S_2$. This is approximately chi-square with 819 degrees of freedom. The p-value is 0.999 for this case, and the hypothesis of the equality of covariances of the two groups is accepted. Thus, the final analysis may be based on 103 characteristics with both groups having the common covariance. We can also try a few other values of p if the seperation of the two groups and equality of the two covariances are maintained. We first tried $p = 200$. For this case, we find that $b_{60,200} = 0.0096$, and the value of the statistic in (8.1) is 59.376, with a p-value of 0.4984. Thus, $p = 200$ does not provide the separation of the two groups. Next we tried $p = 100$. For this case, $b_{60,100} = 0.113$, and the value of the statistic in (8.1) is 86.0128 with a p-value of 0.015. Thus, $p = 100$ is a possible alternative, since the value of the test statistic P_2 for $p^* = 100$ is 317.3079 with a p-value of one, accepting the hypothesis of equality of the two covariances of the two groups. Next, we consider

the case when $p = n = 60$. We now have the case of $p \simeq n$. We found that one eigenvalue of the sample covariance matrix is very small. Thus, we assume that the population covariance matrix Σ is of rank 59 and apply the results given in Theorem 2.1. The value of the test statistic is 21.1035 with a p-value of 0.04646. The corresponding value of the test statistic P_2 for $p^* = 60$ is 666.0173 with a p-value of one. Consequently, $p = 60$ is also a possibility on which we can base our future analysis. However, it appears from the above analysis that $p = 103$ is the best choice.

9. Concluding remarks

In this paper we define sample (squared) distance using the Moore-Penrose inverse instead of any generalized inverse of the sample covariance matrix which is singular due to fewer observations than the dimension. For normally distributed data, it is based on sufficient statistic while a sample (squared) distance using any generalized inverse is not. These distances are used to propose tests in one-sample, two-sample and multivariate analysis of variance. Simultaneous confidence intervals for the mean parameters are given. Using the proposed methods, a dataset is analyzed. This shows that the proposed test statistic performs well. In addition, it provides confidence intervals which have been used to select the relevant characteristics from any dataset.

Appendix A: Proof of Theorem 2.1

Here, we consider the case when the covariance matrix Σ is singular of rank $r \leq n$. That is, for an $r \times p$ matrix M, $MM' = I_r$, $\Sigma = M'D_\lambda M$ where $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, $\lambda_i > 0$, $i = 1, \ldots, r$. Also, $nS = V = YY'$ where the n columns of Y are iid $N_p(\mathbf{0}, \Sigma)$. Since Σ is of rank r, we can write $\Sigma = BB'$ where $B : p \times r$ of rank r. Hence, $Y = BU$, where the n columns of U are iid $N_r(\mathbf{0}, I)$. Then, the matrix S is of rank r, and $\rho(S) = \rho(\Sigma) = \rho(M') = \rho(H')$ where $\rho(S)$ denotes the column vector space of S, and $HVH' = \mathrm{diag}(l_1, \ldots, l_r)$, $H : r \times p$. Thus, there exists an $r \times r$ orthogonal matrix A such that $M = AH$. Let $\boldsymbol{u} \sim N_p(\mathbf{0}, I)$ and $W \sim W_p(I, n)$ be independently distributed.

Then, since $\Sigma^{1/2} = M'D_\lambda^{1/2}M$, $MH' = A$, and $HM' = A'$, we get

$$\begin{aligned}
\left(\frac{N}{n}\right)\overline{\boldsymbol{x}}'S^+\overline{\boldsymbol{x}} &= \boldsymbol{u}'\Sigma^{1/2}H'(HVH')^{-1}H\Sigma^{1/2}\boldsymbol{u} \\
&= \boldsymbol{u}'M'D_\lambda^{1/2}MH'(H\Sigma^{1/2}W\Sigma^{1/2}H')^{-1}H\Sigma^{1/2}\boldsymbol{u} \\
&= \boldsymbol{u}'M'D_\lambda^{1/2}A(A'D_\lambda^{1/2}MWM'D_\lambda^{1/2}A)^{-1}A'D_\lambda^{1/2}M\boldsymbol{u} \\
&= \boldsymbol{u}'M'(MWM')^{-1}M\boldsymbol{u} \\
&= \boldsymbol{z}'U^{-1}\boldsymbol{z}
\end{aligned}$$

where $\boldsymbol{z} \sim N_r(\mathbf{0}, I)$ and $U \sim W_r(I, n)$. This proves Theorem 2.1.

To prove Theorem 2.2, we need the following lemma.

LEMMA A.1. *Let* $\tilde{D} = \operatorname{diag}(\tilde{d}_1, \ldots, \tilde{d}_n)$ *whose distribution is that of the* n *non-zero eigenvalues of a* $p \times p$ *matrix distributed as* $W_p(I, n)$, $n \le p$. *That is, the joint pdf of* $\tilde{d}_1 > \cdots > \tilde{d}_n$ *is given by*

$$\text{(A.1)} \qquad \left[\pi^{(1/2)n^2}/2^{(1/2)pn}\Gamma_n\left(\frac{1}{2}n\right)\Gamma_p\left(\frac{1}{2}n\right)\right] \times \left[\prod_{i<j}^{n}(\tilde{d}_i - \tilde{d}_j)\right]\left[\prod_{i=1}^{n}\tilde{d}_i^{(1/2)(p-n-1)}e^{-(1/2)\tilde{d}_i}\right].$$

Let P *be an* $n \times n$ *random orthogonal matrix with positive diagonal elements distributed independently of* D *with density given by*

$$\text{(A.2)} \qquad \pi^{-(1/2)n^2}\Gamma_n\left(\frac{1}{2}n\right) g_n(P)dP, \qquad PP' = I,$$

where

$$\text{(A.3)} \qquad g_n(P) = J(P'(dP) \to dP),$$

the Jacobian of the transformation from $P'(dP)$ *to* dP *as defined in Srivastava and Khatri* (1979, p. 31). *Then* $U = P\tilde{D}P'$ *is distributed as* $W_n(I, p)$, $n \le p$.

PROOF OF LEMMA A.1. The Jacobian of the transformation $J(P, \tilde{D} \to U)$ can be obtained from Theorem 1.11.5 of Srivastava and Khatri (1979, p. 31). It is given by

$$\text{(A.4)} \qquad J(P, \tilde{D} \to U) = \left[\prod_{i<j}^{n}(\tilde{d}_i - \tilde{d}_j)g_n(P)\right]^{-1}.$$

The joint pdf of $\tilde{D}$ and P is given by

$$\begin{aligned} &\frac{1}{2^{(1/2)pn}\Gamma_n\left(\frac{1}{2}p\right)}g_n(P)\prod_{i<j}^{n}(\tilde{d}_i - \tilde{d}_j)\left[\prod_{i=1}^{n}\tilde{d}_i^{(1/2)(p-n-1)}e^{-(1/2)\tilde{d}_i}\right] \\ &\quad = \frac{1}{2^{(1/2)pn}\Gamma_n\left(\frac{1}{2}p\right)}g_n(P) \\ &\qquad \times \prod_{i<j}^{n}(\tilde{d}_i - \tilde{d}_j)|P\tilde{D}P'|^{(1/2)(p-n-1)}\left(etr - \frac{1}{2}P\tilde{D}P'\right). \end{aligned} \qquad \text{(A.5)}$$

Putting $U = P\tilde{D}P'$ and using the above Jacobian of the transformation given in (A.4), we get the pdf of U which is $W_n(I, p)$, $n \le p$.

COROLLARY A.1. *Let* $\boldsymbol{w} \sim N_n(\mathbf{0}, I_n)$ *and* $\tilde{D} = \operatorname{diag}(\tilde{d}_1, \ldots, \tilde{d}_n)$, $\tilde{d}_1 > \cdots > \tilde{d}_n$, *whose pdf is given by* (A.1) *and* $\boldsymbol{w}$ *and* $\tilde{D}$ *are independently distributed. Then*

$$\frac{p-n+1}{n}\boldsymbol{w}'\tilde{D}^{-1}\boldsymbol{w} \sim F_{n,p-n+1}.$$

PROOF OF COROLLARY A.1. Let P be any $n \times n$ orthogonal matrix distributed independently of $\tilde{D}$ whose pdf is given by (A.2). Then

$$\begin{aligned} \boldsymbol{w}'\tilde{D}^{-1}\boldsymbol{w} &= \boldsymbol{w}'P(P\tilde{D}P')^{-1}P\boldsymbol{w} \\ &= \boldsymbol{z}'U^{-1}\boldsymbol{z}, \end{aligned}$$

where

$$\boldsymbol{z} \sim N_n(\mathbf{0}, I_n)$$

and

$$U \sim W_n(I, p)$$

are independently distributed. Hence,

$$\begin{aligned} \frac{p-n+1}{n}\boldsymbol{w}'\tilde{D}^{-1}\boldsymbol{w} &= \frac{p-n+1}{n}\boldsymbol{z}'U^{-1}\boldsymbol{z} \\ &\sim F_{n,p-n+1}. \end{aligned}$$

PROOF OF THEOREM 2.2. Since $\Sigma = \gamma^2 I = \wedge$ and since T^{+2} is invariant under scalar transformation, we may assume without any loss of generality that $\gamma^2 = 1$. Hence,

$$A = (H \wedge H')^{-1/2} = (HH')^{-1/2} = I.$$

Thus, from (2.15) with a slight change of notation ($\boldsymbol{w}$ instead of $\boldsymbol{z}$), it follows that

$$\frac{T^{+2}}{n} = \boldsymbol{w}'L^{-1}\boldsymbol{w},$$

where $\boldsymbol{w} \sim N_n(\mathbf{0}, I)$ is independently distributed of the diagonal matrix L. Also, the diagonal elements of L are the non-zero eigenvalues of YY', where the n columns of Y are iid $N_p(\mathbf{0}, I_n)$. Thus, L is the diagonal matrix of the eigenvalues of $U = Y'Y \sim W_n(p, I_n)$. Using Corollary A.1, we thus have

$$\begin{aligned} \frac{p-n+1}{n}\frac{T^{+2}}{n} &= \frac{p-n+1}{n}\boldsymbol{w}'L^{-1}\boldsymbol{w} \\ &= \frac{p-n+1}{n}\boldsymbol{z}'U^{-1}\boldsymbol{z} \sim F_{n,p-n+1}. \end{aligned}$$

To prove Theorems 2.3, we need the results stated in the following lemma.

LEMMA A.2. *Let $V = YY' \sim W_p(\wedge, n)$, where the columns of Y are iid $N_p(\mathbf{0}, \wedge)$. Let $l_1, \ldots, l_n$ be the n non-zero eigenvalues of $V = H'LH$, $HH' = I_n$, $L = \operatorname{diag}(l_1, \ldots, l_n)$ and the eigenvalues of $W \sim W_n(I_n, p)$, are given by*

the diagonal elements of the diagonal matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$. *Then in probability*

$$\text{(a)} \quad \lim_{p\to\infty}\left(\frac{Y'Y}{p}\right) = \lim_{p\to\infty}\left(\frac{\operatorname{tr}\wedge}{p}\right) I_n = a_{10} I_n$$

$$\text{(b)} \quad \lim_{p\to\infty}\left(\frac{1}{p}L\right) = a_{10} I_n$$

$$\text{(c)} \quad \lim_{p\to\infty}\left(\frac{1}{p}D\right) = I_n$$

$$\text{(d)} \quad \lim_{p\to\infty}(H \wedge H') = (a_{20}/a_{10}) I_n$$

$$\text{(e)} \quad \lim_{n,p\to\infty}\left(\frac{1}{n}\boldsymbol{a}'H'H\boldsymbol{a}\right) = \lim_{n,p\to\infty}\left(\frac{\boldsymbol{a}'\wedge\boldsymbol{a}}{pa_1}\right)$$

for a non-null vector $\boldsymbol{a} = (a_1, \ldots, a_p)'$ *of constants.*

Proof of Lemma A.2. The n eigenvalues $l_1, \ldots, l_n$ of the diagonal matrix L are the n non-zero eigenvalues of $V = YY'$, where the n columns of the $p \times n$ matrix Y are iid $N_p(\mathbf{0}, \wedge)$. The n non-zero eigenvalues of YY' are also the n eigenvalues of $Y'Y$. Let U denote a $p \times n$ matrix where its n columns are iid $N_p(\mathbf{0}, I)$. Then, the eigenvalues of $Y'Y$ are in distribution the eigenvalues of

$$\begin{aligned} U' \wedge U &= \begin{pmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_n' \end{pmatrix} \wedge (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n), \\ &= \begin{pmatrix} \boldsymbol{u}_1'\wedge\boldsymbol{u}_1, & \boldsymbol{u}_1'\wedge\boldsymbol{u}_2, & \ldots, & \boldsymbol{u}_1'\wedge\boldsymbol{u}_n \\ \boldsymbol{u}_2'\wedge\boldsymbol{u}_1, & \boldsymbol{u}_2'\wedge\boldsymbol{u}_2, & \ldots, & \boldsymbol{u}_2'\wedge\boldsymbol{u}_n \\ \vdots, & \vdots, & \ddots, & \vdots \\ \boldsymbol{u}_n'\wedge\boldsymbol{u}_1, & \boldsymbol{u}_n'\wedge\boldsymbol{u}_2, & \ldots, & \boldsymbol{u}_n'\wedge\boldsymbol{u}_n \end{pmatrix}. \end{aligned}$$

Let $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = (u_{ij})$. Then u_{ij} are iid $N(0,1)$ and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are iid $N_p(\mathbf{0}, I)$. Hence,

$$E(\boldsymbol{u}_1'\wedge\boldsymbol{u}_1) = \operatorname{tr}\wedge, \qquad E(\boldsymbol{u}_1'\wedge\boldsymbol{u}_2) = 0,$$

and

$$E(Y'Y) = E(U' \wedge U) = pa_1 I_n$$

where

$$a_1 = (\operatorname{tr}\wedge/p), \qquad \text{and} \qquad \lim_{p\to\infty} a_1 = a_{10}.$$

We also note that

$$E(u_{ij}^2) = 1, \qquad \operatorname{var}(u_{ij}^2) = 2.$$

Hence, from Chebyshev's inequality

$$P\left\{\left|\frac{\boldsymbol{u}_1'\wedge\boldsymbol{u}_1}{p} - a_1\right| > \epsilon\right\} = P\left\{\left|\frac{\sum_{i=1}^p \lambda_i(u_{1i}^2 - 1)}{p}\right| > \epsilon\right\}$$

$$\leq \frac{E[\sum_{i=1}^{p}\lambda_i(u_{1i}^2-1)]^2}{p^2\epsilon^2}$$
$$= \frac{E[\sum_{i=1}^{p}\lambda_i^2(u_{1i}^2-1)^2]}{p^2\epsilon^2}$$
$$= \frac{2\sum_{i=1}^{p}\lambda_i^2}{p^2\epsilon^2}.$$

Since $0 < \lim_{p\to 0}(\mathrm{tr}\wedge^2/p) < \infty$, it follows that

$$\lim_{p\to\infty}\frac{\sum_{i=1}^{p}\lambda_i^2}{p^2} = 0.$$

Hence,

$$\lim_{p\to\infty}\frac{\boldsymbol{u}_i'\wedge\boldsymbol{u}_i}{p} \to a_{10}, \qquad i=1,\ldots,n$$

in probability. Similarly, it can be shown that in probability

$$\lim_{p\to\infty}\frac{\boldsymbol{u}_i'\wedge\boldsymbol{u}_j}{p} = 0, \qquad i\neq j,$$

and

$$\lim_{p\to\infty}\frac{Y'Y}{p} = a_{10}I_n \qquad \text{in probability.}$$

This proves (a). Also, if $l_1,\ldots,l_n$ denote the non-zero eigenvalues of YY' then, from the above result, it follows that

$$\lim_{p\to\infty}\left(\frac{1}{p}\right)L = a_{10}I_n \qquad \text{in probability.}$$

This proves (b).

It follows from the above results that if $d_1,\ldots,d_n$ are the non-zero eigenvalues of $\wedge^{-1/2}V\wedge^{-1/2}$ and $D=\mathrm{diag}(d_1,\ldots,d_n)$, then in probability

$$\lim_{p\to\infty}\frac{1}{p}D = I_n, \tag{A.6}$$

since $d_1,\ldots,d_n$ are the eigenvalues of $W\sim W_n(p,I)$. This proves (c). We note that

$$YY' = H'L^{1/2}GG'L^{1/2}H,$$

for any $n\times n$ orthogonal matrix G, $GG'=I_n$ depending on Y. Choosing $G = L^{1/2}HY(Y'Y)^{-1}$, we find that in distribution,

$$Y = H'L^{1/2}G \sim N_{p,n}(0,\wedge,I_n).$$

Thus, in distribution

$$GY'\wedge YG' = GU'\wedge^2 UG' = L^{1/2}H\wedge H'L^{1/2},$$

where $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$. We note that

$$E\left(\frac{\boldsymbol{u}_i' \wedge^2 \boldsymbol{u}_j}{p}\right) = \frac{\operatorname{tr}\wedge^2}{p}\delta_{ij},$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$, $i \neq j$, $i, j = 1, \ldots, n$, the Kronecker symbol. Similarly,

$$\begin{aligned}\operatorname{Var}\left(\frac{\boldsymbol{u}_i' \wedge^2 \boldsymbol{u}_j}{p}\right) &= \frac{2\operatorname{tr}\wedge^4}{p^2}, \qquad i = j,\\ &= \frac{\operatorname{tr}\wedge^4}{p^2}, \qquad i \neq j.\end{aligned}$$

Since, $\lim_{p\to\infty} \operatorname{tr}\wedge^4/p = a_{40}$, and $0 < a_{40} < \infty$, it follows that

$$\lim_{p\to\infty} \frac{\operatorname{tr}\wedge^4}{p^2} = 0.$$

Hence, in probability,

$$\lim_{p\to\infty}\left[G\left(\frac{Y' \wedge Y}{p}\right)G'\right] = \left(\lim_{p\to\infty}\frac{\operatorname{tr}\wedge^2}{p}\right) I_n = a_{20}I_n.$$

Thus, in probability

$$\lim_{p\to\infty}\left(\frac{L^{1/2}H \wedge H'L^{1/2}}{p}\right) = a_{20}I_n.$$

Since, $\lim_{p\to\infty}(L/p) = a_{10}I_n$ it follows that in probability

$$\lim_{p\to\infty}(H \wedge H') = (a_{20}/a_{10})I_n.$$

This proves (d).

To prove (e), consider a non-null p-vector $\boldsymbol{a} = (a_1, \ldots, a_p)'$. Then, since $YY' = H'LH$, $HH' = I_n$, we get

$$\frac{\boldsymbol{a}'YY'\boldsymbol{a}}{pn} = \frac{\boldsymbol{a}'H'LH\boldsymbol{a}}{pn}.$$

With $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})'$, the left side

$$\begin{aligned}&= \frac{1}{pn}\sum_{i=1}^{n}(\boldsymbol{a}'\boldsymbol{y}_i)^2\\ &= \frac{1}{pn}\left[\sum_{i=1}^{n}\sum_{j=1}^{p}a_j^2y_{ij}^2\right] + \frac{2}{pn}\sum_{i=1}^{n}\sum_{j<k}^{p}a_ja_ky_{ij}y_{ik},\end{aligned}$$

of which the second term goes to zero in probability.

Hence, in probability

$$\lim_{n,p\to\infty}\frac{1}{pn}\sum_{i=1}^{n}\sum_{j=1}^{p}a_j^2y_{ij}^2=\lim_{n,p\to\infty}\frac{\boldsymbol{a}'H'LH\boldsymbol{a}}{pn}.$$

From the law of large numbers, the left side goes to $\lim_{n\to\infty}\lim_{p\to\infty}(\frac{\boldsymbol{a}'\wedge\boldsymbol{a}}{p})$, and from the results in (a), we have in probability $\lim_{p\to\infty}p^{-1}L=a_{10}$. Hence, in probability

$$\lim_{n,p\to\infty}\left(\frac{\boldsymbol{a}'H'H\boldsymbol{a}}{n}\right)=\lim_{n,p\to\infty}\left(\frac{\boldsymbol{a}'\wedge\boldsymbol{a}}{p}\right)/a_{10}.$$

Proof of Theorem 2.3. We note that $A=(H\wedge H')^{-1/2}$ and from the proof of Lemma A.2 (d),

$$(H\wedge H)'\stackrel{d}{=}\left(\frac{L}{p}\right)^{-1/2}\left(\frac{GU'\wedge^2UG'}{p}\right)\left(\frac{L}{p}\right)^{-1/2},$$

where $GG'=I_n$, $U=(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)$, and $\boldsymbol{u}_i$ are iid $N_p(\boldsymbol{0},I)$. Since the distribution of T^{+2} is invariant under orthogonal transformations, we get from (2.15)

$$\begin{aligned}p\hat{b}\frac{T^{+2}}{n}&=\hat{b}\boldsymbol{z}'\left(\frac{ALA}{p}\right)^{-1}\boldsymbol{z}\\&\stackrel{d}{=}\hat{b}\boldsymbol{z}'\left(\frac{L}{p}\right)^{-1/2}A^{-2}\left(\frac{L}{p}\right)^{-1/2}\boldsymbol{z}\\&\stackrel{d}{=}\hat{b}\boldsymbol{z}'\left(\frac{L}{p}\right)^{-1}G\left(\frac{U'\wedge^2U}{p}\right)G'\left(\frac{L}{p}\right)^{-1}\boldsymbol{z}.\end{aligned}$$

Hence,

$$\begin{aligned}\lim_{n,p\to\infty}P_0\left[\frac{(p/n)\hat{b}T^{+2}-n}{(2n)^{1/2}}<z_{1-\alpha}\right]&=\lim_{n,p\to\infty}P_0\left[\frac{\boldsymbol{z}'G\frac{U'\wedge^2U}{pa_{20}}G'\boldsymbol{z}-n}{(2n)^{1/2}}<z_{1-\alpha}\right]\\&=\lim_{n,p\to\infty}P_0\left[\frac{\boldsymbol{z}'\frac{U'\wedge^2U}{paa_{20}}\boldsymbol{z}-n}{(2n)^{1/2}}<z_{1-\alpha}\right].\end{aligned}$$

Let r_i be the eigenvalues of $(U'\wedge^2U/pa_{20})$. Then

$$r_i=1+O_p(p^{-1/2}),$$

and

$$\frac{1}{n}\left(\sum_{i=1}^{n}r_i\right)=\frac{\operatorname{tr}\wedge^2UU'}{npa_{20}}=1+O_p(np)^{-1/2}.$$

Hence,

$$\lim_{n,p\to\infty}P_0\left[\frac{(p/n)\hat{b}T^{+2}-n}{(2n)^{1/2}}<z_{1-\alpha}\right]$$

$$= \lim_{n,p\to\infty} P_0\left[\frac{\sum_{i=1}^n r_i z_i^2 - n}{(2n)^{1/2}} < z_{1-\alpha}\right]$$
$$= \lim_{n,p\to\infty} P_0\left[\frac{\sum_{i=1}^n r_i(z_i^2-1)}{(2n)^{1/2}} + \frac{\sum_{i=1}^n r_i - n}{(2n)^{1/2}} < z_{1-\alpha}\right].$$

Now

$$\frac{\sum_{i=1}^n r_i - n}{(2n)^{1/2}} = \left(\frac{n}{2}\right)^{1/2}\left[\frac{\sum_{i=1}^n r_i}{n} - 1\right]$$
$$= \left(\frac{n}{2}\right)^{1/2}\left(\frac{1}{np}\right)^{1/2} O_p(1) \to 0 \quad \text{as} \quad (n,p)\to\infty.$$

Hence, from Lemma 2.1,

$$\lim_{n,p\to\infty} P_0\left[\frac{(p/n)\hat{b}T^{+2} - n}{(2n)^{1/2}} < z_{1-\alpha}\right] = \Phi(z_{1-\alpha}).$$

Proof of Theorem 2.4. The ALA occuring in (2.15) can be rewritten as

$$ALA = (H\wedge H')^{-1/2}(HVH')(H\wedge H')^{-1/2}$$
$$= (H\wedge H')^{-1/2} H \wedge^{1/2} W \wedge^{1/2} H'(H\wedge H')^{-1/2}$$
$$= PWP',$$

where $W \sim W_p(I,n)$ and $P = (H\wedge H')^{-1/2}H\wedge^{1/2}$, $PP' = I_n$. Let G be an $n\times n$ orthogonal matrix such that

$$PWP' = GMG'$$

where $M = \operatorname{diag}(m_1,\ldots,m_n)$, $m_1 > \cdots > m_n$ are the ordered eigenvalues of PWP'. Then, if $d_1 > \cdots > d_n$ are the ordered nonzero eigenvalues of W, we get from Poincare Separation Theorem, see Rao (1973, p. 64), $m_i \le d_i$, $i = 1,\ldots,n$. Hence in distribution

$$\frac{T^{+2}}{n} = \boldsymbol{z}'GM^{-1}G'\boldsymbol{z} = \boldsymbol{v}'M^{-1}\boldsymbol{v}$$
$$\ge \boldsymbol{v}'D^{-1}\boldsymbol{v},$$

where $\boldsymbol{v} \sim N_n(\boldsymbol{0}, I)$ is independently distributed of D. Thus from Corollary A.1, $F^+ \ge F_{n,p-n+1}$.

Acknowledgements

I am grateful to Dr. Ekkehard Glimm of AICOS Technologies, Basel, Switzerland for kindly reading this article and offering many suggestions that greatly improved the presentation. Thanks are also due to Professors Y. Fujikoshi, J. Láuter, R. Pincus, M. Genton, and two anonymous referees for their helpful comments. The calculation for the example in Section 8 was carried out by M. Shakhatreh, that of Tables 1–3 by Meng Du, and that of Tables 4–6 by Yan Liu; sincerest thanks to each of them. The research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Alon, U., Notterman, D. A., Gish, K., Yhurra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of Gene expression revealed by clustering analysis of Tumor and Normal colon tissues probed by Oligonucleotide Assays, *Proceedings of the National Academy of Sciences*, **96**, 6745–6750.

Alter, O., Brown, P. O. and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modelling, *Proceedings of the National Academy of Sciences*, **97**, 10101–10106.

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

Benjamini, Y. and Hockberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. Ser. B*, **57**, 289–300.

Dempster, A. P. (1958). A high dimensional two sample significance test, *Ann. Math. Statist.*, **29**, 995–1010.

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

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of Tumors using gene expression data, *J. Amer. Statist. Assoc.*, **97**, 77–87.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *J. Amer. Statist. Assoc.*, **96**, 1151–1160.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression-patterns, *Proceeding of the National Academy of Sciences*, **95**, 14863–14868.

Fujikoshi, Y., Himeno, T. and Wakaki, H. (2004). Asymptotic results of a high dimensional MANOVA test and power comparison when the dimension is large compared to the sample size, *J. Japan. Statist. Soc.*, **34**, 19–26.

Ibrahim, J., Chen, M. and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data, *J. Amer. Statist. Assoc.*, **97**, 88–99.

Johnston, I. M. (2001). On the distribution of the largest eigenvalue in principle component analysis, *Ann. Statist.*, **29**, 295–327.

Läuter, J., Glimm, E. and Kropf, S. (1998). Multivariate tests based on left-spherically distributed linear scores, *Ann. Statist.*, **26**, 1972–1988.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size, *Ann. Statist.*, **30**, 1081–1102.

Lehmann, E. L. (1959). *Testing Statistical Hypothesis*, Wiley, New York.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York.

Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and Its Applications*, Wiley, New York.

Schott, J. R. (1997). *Matrix Analysis for Statistics*, Wiley, New York.

Simaika, J. B. (1941). On an optimum property of two important statistical tests, *Biometrika*, **32**, 70–80.

Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis*: *A Graduate Course and Handbook*, American Sciences Press, Inc., Columbus, Ohio, U.S.A.

Srivastava, M. S. (1970). On a class of non-parametric tests for regression parameters, *J. Statist. Res.*, **4**, 117–131.

Srivastava, M. S. (1972). Asymptotically most powerful rank tests for regression parameters in MANOVA, *Ann. Inst. Statist. Math.*, **24**, 285–297.

Srivastava, M. S. (2002). *Methods of Multivariate Statistics*, Wiley, New York.

Srivastava, M. S. (2003). Singular Wishart and Multivariate beta distributions, *Ann. Statist.*, **31**, 1537–1560.

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in High-Dimensional data, *J. Japan. Statist. Soc.*, **35**, 251–272.

Srivastava, M. S. and Fujikoshi, Y. (2006). Multivariate analysis of variance with fewer observations than the dimension, *J. Multivariate Anal.*, **97**, 1927–1940.

Srivastava, M. S. and Khatri, C. G. (1979). *An Introduction to Multivariate Statistics*, North-Holland, New York.

Srivastava, M. S. and von Rosen, D. (2002). Regression Models with unknown singular covariance matrix, *Linear Algebra and its Applications*, **354**, 255–273.

Srivastava, M. S., Hirotsu, C., Aoki, S. and Glimm, E. (2001). Multivariate one-sided tests, *Data Analysis from Statistical Foundations*, (eds. Mohammad, A. K. and Saleh, E.), 387–401, Nova Science Publishers Inc., New York.