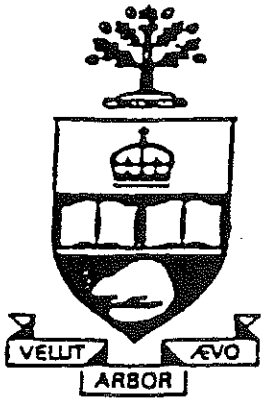

Some Tests Criteria For the Covariance Matrix With Fewer Observations Than the Dimension

by

M.S. Srivastava
Department of Statistics
University of Toronto

Technical Report No. 0406 March 22, 2006

TECHNICAL REPORT SERIES

University of Toronto

Department of Statistics

Some Tests Criteria For the Covariance Matrix With Fewer Observations Than the Dimension

M. S. Srivastava

Department of Statistics

University of Toronto

100, St. George Street

Toronto, Ontario, M5S 3G3.

CANADA

Phone: 416-978-4450, Fax: 416-978-5133

srivasta@utstat.toronto.edu

Abstract

In this article, we consider testing certain hypotheses concerning the covariance matrix Σ when the number of observations $N = n + 1$ on the p-dimensional random vector $\boldsymbol{x}$, distributed as normal, is less than p, $n < p$, and (n/p) goes to zero. Specifically, we consider testing $\Sigma = \sigma^2 I$, $\Sigma = I$, and $\Sigma = \wedge$, a diagonal matrix, where I is the $p \times p$ identity matrix. The first two tests are the adapted versions of the likelihood ratio tests when $n > p$, and (p/n) goes to zero and the third test is the normalized version of the Fisher's z-transformation. These tests are compared with some recently proposed tests.

Key words and phrases: covariance matrix, hypothesis testing, multivariate normal, power comparison, spherecity hypothesis.

1 Introduction

Recent advances in techonology to obtain DNA mciroarrays have made it possible to measure quantitatively the expressions of thousands of genes. These observations are, however, corelated to each other as the genes are from the same subject. Since the number of subjects available for taking the observations are so few as compared to the number of genes expressions, multivariate theory needs to be developed. Alternatively, if it can be verified that the covariance matrix for the p gene expressions is either an identity matrix or a constant times the identity matrix, then the usual univariate theories can be applied.

Indeed, univariate theories have recently been used to analyze microarray datasets without verifying the sphericity assumption, see for example, Efron, Tibshirani, Storey and Tusher (2001). On the other hand, Dudoit, Fridlyand and Speed (2002) assumed that the covariance matrix is a diagonal matrix and applied Fisher's linear discriminant rule with estimated diagonal elements and the mean vectors. But the assumption of the diagonality of the covariance matrix was not verified and tested for the data. In fact, we applied our tests presented in this paper to check for the spherecity and the diagonality of the covariance matrix of the Colon datasets of Alon et al. (1999) and of Leukemia datasets of Golub et al. (1999) and found that the covariance matrices of these two datasets are neither sphereical nor diagonal with a p-value of zero.

In this article, we consider the problem of testing the following three hypothesis:

$$1.\quad H_1 \quad \Sigma = \sigma^2 I_p,\ \sigma^2 > 0,\ \text{vs}\ A_1 \neq H_1 \qquad (1.1)$$

$$2.\quad H_2 \quad \Sigma = I_p, \text{ vs } A_2 \neq H_2 \tag{1.2}$$

$$3.\quad H_3 \quad \Sigma = \wedge = \text{ diag}(\lambda_1, \ldots, \lambda_p) \text{ vs } A_3 \neq H_3 \tag{1.3}$$

For the first two hypothesis H_1, and H_2, we propose adapted versions of the likelihood ratio tests available for $n > p$ and (p/n) going to zero. The advantage of such tests is that the same tests can be used for both situations when $n < p$ or $n > p$ by simply switching n and p. For the diagonality hypothesis H_3, however, such an adapted version is not available, and we propose a test based on the normalized version of the Fisher's z-transformations of the pairwise correlation coefficients.

To describe the three proposed tests, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be independently and identically distributed (iid) as multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ denoted as $N_p(\boldsymbol{\mu}, \Sigma)$. We shall assume that the $p \times p$ covariance matrix Σ is positive definite and often denoted as $\Sigma > 0$. The sample mean vector $\overline{\boldsymbol{x}}$ and the sample covariance matrix S are respectively given by

$$\overline{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \ , \tag{1.4}$$

and

$$nS = V = \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})' \ . \tag{1.5}$$

When $n < p$, the sample covariance matrix S is singular, and there are only n non-zero eigenvalues of S of V. Let $\tilde{l}_1, \ldots, \tilde{l}_n$ denote the non-zero eigenvalue of V, and let $l_1, \ldots, l_n$ denote the corresponding eigenvalues of S. That is

$$\tilde{l}_i = n l_i \ . \tag{1.6}$$

Then, the proposed test for the testing problem (1.1) is given by

$$L_1 = \frac{\prod_{i=1}^{n} l_i}{\left(\sum_{i=1}^{n} \frac{l_i}{n}\right)^n} = \frac{\prod_{i=1}^{n} \tilde{l}_i}{\left(\sum_{i=1}^{n} \frac{\tilde{l}_i}{n}\right)^n} \tag{1.7}$$

Similarly, an adapted version of the likelihood ratio test for the testing problem in (1.2) is given by

$$L_2 = \left(\frac{e}{p}\right)^{\frac{1}{2}pn} \left(\prod_{i=1}^{n} \tilde{l}_i\right)^{\frac{1}{2}p} e^{-\frac{1}{2}\sum_{i=1}^{n} \tilde{l}_i} \tag{1.8}$$

For testing the diagonality, we consider the statistic

$$Q_3 = \frac{(n-2)\sum_{i<j} z_{ij}^2 - \frac{1}{2}p(p-1)}{\sqrt{p(p-1)}}\,, \tag{1.9}$$

where

$$z_{ij} = \log\left(\frac{1+r_{ij}}{1-r_{ij}}\right)\,,\ r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}\,,\ \text{and } S = (s_{ij}). \tag{1.10}$$

It may be noted that the testing problems given in (1.1) and (1.2) remain invariant under the transformation by an element of the group of $p \times p$ orthogonal matrices. In addition, the testing problem in (1.1) is also invariant under a scalar transformation. Thus, without any loss of generality, we may assume that the covariance matrix Σ is a diagonal matrix given by

$$\Sigma = \ \text{diag}(\lambda_1, \ldots, \lambda_p) = \Lambda\ . \tag{1.11}$$

Thus, the testing problem stated in (1.1) becomes

$$H_1 :\ \lambda_1 = \ldots = \lambda_p = \lambda,\ \text{against } A_1 : \lambda_1 \neq \lambda \tag{1.12}$$

for at least one $i = 1, \ldots, p$, where λ in unknown. Similarly, the testing problem stated in (1.2) becomes

$$H_2 :\ \lambda_1 = \ldots = \lambda_p = 1,\ \text{against } A_2 \neq H_2\ . \tag{1.13}$$

The testing problem (1.3) remains invariant under the transformation $C\boldsymbol{x}$, where $C = \text{ diag}(c_1, \ldots, c_p),\ c_i \neq 0,\ i = 1, \ldots, p$. Hence, if

$$\Sigma = (\sigma_{ij}) \text{ and } \rho_y = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\ ,\ i \neq j\ , \tag{1.14}$$

then for testing the independence of the p characteristics of the p-random vector $\boldsymbol{x}$, we may test the hypothesis that $\sigma_{ij} = 0$ or equivalently $\rho_{ij} = 0,\ i \neq j$, against the alternative $A_3 :\ \sigma_{ij} \neq 0$ for at least one pair $(i, j),\ i \neq j$. The organization of the paper is as follows.

The test for testing the sphericity of covariance matrix is described in Section 2 and that for testing the hypothesis that the covariance matrix is an identity matrix is described in Section 3. The test for the diagonality hypothesis is described in Section 4. In Section 5, we test for the spherecity and the diagonality of two microarray datasets, analyzed by Dudoit et al. (2002) under the assumption that the covariance matrices are diagonal matrices. We compare, by simulation, the attained significance level (ASL) with those obtained from the asymptotic distributions in Section 6. A comparison of powers with some recently proposed tests by simulation is presented in Section 7. The paper concludes in Section 8.

2 Testing the Sphericity Hypothesis

For testing the sphericity hypothesis that $\Sigma = \sigma^2 I$ against the alternative that $\Sigma \neq \sigma^2 I$, the modified likelihood ratio test when $n \geq p$, is equivalent to a test based on the sufficient statistic S, ignoring the information available on the mean vector $\boldsymbol{\mu}$. We shall therefore also consider a test based on the

sufficient statistic S, which is distributed as $W_p(\Lambda, n)$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$. The modified likelihood ratio test has been shown by Carter and Srivastava (1977) to have a monotone power function. But when $n < p$, the likelihood ration test does not exist. To construct a test that minics the above property, we first consider the case when $p = kn$, k is an integer. We shall now assume that the $p \times p$ diagonal matrix Λ is given by

$$\Lambda = D_d \otimes I_k \ , \tag{2.1}$$

where $D_d = \text{diag}(d_1, \ldots, d_n)$ is an $n \times n$ diagonal matrix, and $A \otimes B$ is the Kronecker product of two matrices $A = (a_{ij}) : \ m \times n$ and $B : p \times q$ given by $(a_{ij}B) : mp \times nq$. Then, the sphericity hypothesis is equivalent to testing that $d_i = d$, d unknown for all $i, 1 = 1, \ldots, n$. It is known, see, for example, Srivastava and Khatri (1979) that

$$nS = YY' \ , \tag{2.2}$$

where $Y = (y_1, \ldots, y_n)$ and $\boldsymbol{y}_i$'s are iid. $N_p(\mathbf{0}, \Lambda)$, and $\Lambda = D_d \otimes I_k$. Writing

$$\boldsymbol{Y}' = \left(\boldsymbol{y}_{(1)}, \ldots, \boldsymbol{y}_{(p)}\right) \ , \tag{2.3}$$

We find that $\boldsymbol{y}_{(i)}$'s are iid. $N_n(\mathbf{0}, D_d)$, $i = 1, \ldots, p$. Hence, $\boldsymbol{Y}'\boldsymbol{Y} \sim W_n(D_d, p)$, a Wishart distribution with mean pD_d and degrees of freedom p. Hence, the likelihood ratio test based on the observations $\boldsymbol{y}_{(1)}, \ldots, \boldsymbol{y}_{(p)}$ for testing the hypothesis that $d_i = d$ for all $i = 1, \ldots, n$ against the alternatively that $d_i \neq d$ for at least one $i = 1, \ldots, n$ is given by

$$L_1 = \frac{\prod_{i=1}^{n} \tilde{l}_i}{\left(\sum_{i=1}^{n} \frac{\tilde{l}_i}{n}\right)^n} \tag{2.4}$$

where $\tilde{l}_1, \ldots, \tilde{l}_n$, are the non-zero eigenvalues of $\boldsymbol{Y}'\boldsymbol{Y}$, and thus equivalently of $V = \boldsymbol{Y}\boldsymbol{Y}'$. The test L_1 is the ratio of the geometric mean and the arithmetic mean of the n non-zero eigenvalues of V or equivalently of S.

Although the above test has been derived assuming that $p = kn$, k an integer and $\Lambda = D_d \otimes I_k$, we shall propose the test L_1 for all the cases. The test given in (2.4) may thus be considered as an adapted version of the likelihood ratio test when $n > p$ to the case $n < p$, obtained by simply interchanging n and p. The distribution of L_1, can also be obtained in the same manner. To write it explicitly, let

$$c_1 = \frac{(n+1)(n-1)(n+2)(2n^3+6n^2+3n+2)}{288n^2} , \tag{2.5}$$

$$m_1 = p - \frac{2n^2+n+2}{6n} , \tag{2.6}$$

$$g_1 = \frac{1}{2}n(n+1) - 1 , \tag{2.7}$$

$$Q_1 = -m \log L_1 . \tag{2.8}$$

Then, under the hypothesis that $\Lambda = \sigma^2 I$,

$$P(Q_1 \geq z) = P(\chi^2_{g_1} \geq z) + c_1 m_1^{-2} \left[P(\chi^2_{g_1+4} \geq z) - P(\chi^2_{g_1} \geq z) \right] + O(m_1^{-3}) , \tag{2.9}$$

which is obtained by interchanging n and p in the asymptotic expression of the cdf of the corresponding likelihood ratio test when $n > p$, see Srivastava (2002, p482.).

In Section 4, we compare the power of the Q_1 test with a recently proposed test by Srivastava (2005), given by

$$T_1 = \left(\frac{n}{2}\right)(\hat{\gamma}_1 - 1) , \tag{2.10}$$

where

$$\hat{\gamma}_1 = \frac{\hat{a}_2}{\hat{a}_1^2} , \tag{2.11}$$

$$\hat{a}_2 = \frac{n^2}{(n-1)(n+2)} \left(\frac{1}{p}\right) \left[\mathrm{tr}S^2 - \frac{1}{n}(\mathrm{tr}S)^2\right] , \tag{2.12}$$

$$\hat{a}_1 = \frac{\mathrm{tr}S}{p} , \tag{2.13}$$

Asymptotically as $(n,p) \to \infty$,

$$\left(\frac{n}{2}\right)(\hat{\gamma}_1 - \gamma_1) \sim N(0, \tau_1^2) , \tag{2.14}$$

with

$$\tau_1^2 = \frac{2n(a_4 a_1^2 - 2a_1 a_2 a_3 + a_2^3)}{p a_1^6} + \frac{a_2^2}{a_1^4} , \tag{2.15}$$

$$a_i = \frac{\mathrm{tr}\Sigma_i}{p} , \; i = 1, \ldots, 4 , \; \gamma_1 = a_2/a_1^2 . \tag{2.16}$$

It is assumed that

$$0 < a_{i0} = \lim_{p \to \infty} a_i < \infty . \tag{2.17}$$

Since, under the hypothesis $\gamma_1 = 1$, it follows that as $(n,p) \to \infty$, asymptotically

$$T_1 \sim N(0,1) , \tag{2.18}$$

The asymptotic power of T_1 is given by

$$\lim_{(n,p)\to\infty} P\left\{\frac{n}{2}\left[(\hat{\gamma}_1 - 1) - (\gamma_1 - 1)\right] > z_\alpha - \frac{n}{2}(\gamma_1 - 1) | \gamma_1 > 1\right\} = \Phi\left[\frac{\frac{n}{2}(\gamma_1 - 1) - z_\alpha}{\tau_1}\right] \tag{2.19}$$

where z_α is the upper $100\alpha\%$ point of the standard normal distribution, and Φ denotes the cdf of a standard normal random variable.

The simulation result (not included) shows that the power given in (2.19) provides a very good approximation for large p. In the above results, there is no restriction as to how n and p go to infinity.

3 Testing That the Covariance Matrix Is an Identity Matrix

Again, as in Section 2, because of the invariance under the orthogonal transformation, we shall assume that Σ is a diagonal matrix given in (1.12). It has been shown by Nagao (1967) that the modified likelihood ratio test has a monotone power function. Thus, motivated by this fact, we propose a test for testing that $\lambda_1 = 1$ for all $i = 1, 2 \ldots, p$, against the alternative that $\lambda_i \neq 1$ for at least one $i = 1, 2, \ldots, p$, which mimics the likelihood ratio test except that here $n < p$ and we have only n non-zero eigenvalues $\tilde{l}_1, \ldots, \tilde{l}_n$ of $V = nS$. Thus, we propose a test statistic

$$L_2 = \left(\frac{e}{p}\right)^{\frac{1}{2}pn} \left(\prod_{i=1}^{n} \tilde{l}_i\right)^{\frac{1}{2}p} e^{-\frac{1}{2}\sum_{i=1}^{n} \tilde{l}_i} . \tag{3.1}$$

Let

$$\begin{aligned} g_2 &= \frac{1}{2}n(n+1) . \\ m_2 &= p - \frac{2n^2+3n+1}{6(n+1)} . \\ c_2 &= \frac{n}{288(n+1)}(2n^4+6n^3+n^2-12n-13) . \\ Q_2 &= -\left(\frac{2m_2}{p}\right)\log L_2 . \end{aligned} \tag{3.2}$$

Then

$$P(Q_2 \geq z) = P(\chi^2_{g_2} \geq z) + c_2 m_2^{-2} \left[P(\chi^2_{g_2+4} \geq z) - P(\chi^2_{g_2} \geq z)\right] + O(m_2^{-3}) \ . \tag{3.3}$$

When $n \geq p$, we replace n by p and p by n in L_2 as well as in all the formulas given above.

The test given in (3.1) is the likelihood ratio test when $\Lambda = D_d \otimes I_k$, $p = kn$, and $D_d = \text{diag}(d_1, \ldots, d_n)$ and we wish to test the hypothesis that $d_i = 1$ for all $i = 1, 2, \ldots, n$, against the alternative that $d_i \neq 1$ for at least one $i = 1, 2, \ldots, n$.

The test proposed by Srivastava (2005) using a consistent estimator of the parametric function that seperates the null hypothesis from the alternative hypothesis is given by

$$T_2 = \left(\frac{n}{2}\right)(\hat{\gamma}_2 + 1) \ , \tag{3.4}$$

where $\hat{\gamma}_2 = \hat{a}_2 - 2\hat{a}_1$, $\hat{a}_2$ and $\hat{a}_1$ have been defined in (2.12) and (2.13) respectively. Asymptotically as $(n, p) \to \infty$,

$$\left(\frac{n}{2}\right)(\hat{\gamma}_2 - \gamma_2) \sim N(0, \tau_2^2) \ , \tag{3.5}$$

with

$$\gamma_2 = a_2 - 2a_1 \ ,$$

and

$$\tau_2^2 = \left(\frac{2n}{p}\right)(a_2 - 2a_3^3 + a_4) + a_2^2 \ ,$$

where a_i has been defined in (2.16), and it is assumed that (2.17) holds.

Under the hypothesis that $\lambda_i = 1$ (or $d_i = 1$), $\gamma_2 = -1$ and $\tau^2 = 1$. Hence,

the asymptotic null distribution as $(n,p) \to \infty$ is given by

$$T_2 \sim N(0,1) \ . \tag{3.6}$$

Thus, the asymptotic power of the T_2 test is given by

$$\lim_{(n,p)\to\infty} P\left\{T_2 > z_\alpha \mid \gamma_2 + 1 > 0\right\} = \Phi\left[\frac{\frac{n}{2}(\gamma_2 - 1) - z_\alpha}{\tau_2}\right]$$

4 Testing That the Covariance Matrix Is a Diagonal Matrix

In this section, we consider the problem of testing the hypothesis described in (1.3), namely, that the covariance matrix Σ is a diagonal matrix against the alternative that it is not a diagonal matrix when the sample size $N \leq p$. Unfortunately, the device used to obtain adapted versions of the likelihood ratio tests for $N > p$ in Sections 2 and 3 cannot be applied here as Λ cannot be assumed to be equal to $D_d \otimes I_k$. Thus, we shall consider tests based on covariances or correlations since the problem remains invariant under the transformation $\boldsymbol{x} \to C\boldsymbol{x}$, where $C = \text{diag}(c_1, \ldots, c_p)$. Thus, let

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \ , \ i \neq j \ , \tag{4.1}$$

where $S = (s_{ij})$. Under the hypothesis of independence of the p characteristics of the random vector that is $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Lambda)$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $r_{ij}\sqrt{n}$, $i \neq j$ are asymptotically independently distributed with mean 0 and variance 1, see Srivastava and Khatri (1979, p. 103). However, as is well known, its convergence to normality is slow. Thus we consider the Fisher's

z-transformation defined by

$$z_{ij} = \frac{1}{2}\log\frac{1+r_{ij}}{1-r_{ij}}\ ,\ i \neq j\ , \tag{4.2}$$

and propose the test statistic

$$Q_3 = \frac{(n-2)\sum_{i<j} z_{ij}^2 - \frac{1}{2}p(p-1)}{\sqrt{p(p-1)}}\ , \tag{4.3}$$

which is asymptotically distributed as $N(0,1)$ under the hypothesis as n and p go to infinity.

Another test based on the covariances have recently been proposed by Srivastava (2005). it is given by

$$T_3 = \left(\frac{n}{2}\right)\frac{(\hat{\gamma}_3 - 1)}{\left[1 - \frac{1}{p}\left(\frac{\hat{a}_{40}}{\hat{a}_{20}^2}\right)\right]^{\frac{1}{2}}}\ , \tag{4.4}$$

where

$$\hat{\gamma}_3 = \frac{\hat{a}_2}{\hat{a}_{20}}\ , \tag{4.5}$$

$$\hat{a}_{20} = \frac{n}{p(n+2)}\sum_{i=1}^{p} s_{ii}^2\ , \tag{4.6}$$

$$\hat{a}_{40} = \frac{1}{p}\sum_{i=1}^{p} s_{ii}^4\ , \tag{4.7}$$

and $\hat{a}_2$ has been defined in (2.12). It has been shown by Srivastava (2005) that under the hypothesis, as n and p go to infinity, T_3 is asymptotically $N(0,1)$.

Under the alternative hypothesis of $\sigma_{ij} \neq 0,\ i \neq j,\ i,\ j = 1,\ldots,p$, for at least one pair of (i,j) asymptotically as $(n,p) \to \infty$.

$$T_3 \sim N(\delta, \tau_3^2)\ , \tag{4.8}$$

where

$$\delta = \frac{n}{2}(\gamma_3 - 1)\left[1 - \left(\frac{1}{p}\right)\frac{a_{40}}{a_{20}^2}\right]^{-\frac{1}{2}} \tag{4.9}$$

$$\tau_3^2 = \frac{a_2^2 - p^{-1}a_4}{a_{20}^2 - p^{-1}a_{40}}\ ,\ a_{20} = \frac{\sum_{i=1}^{p}\sigma_{ii}^2}{p}\ ,\ a_{40} = \frac{\sum_{i=1}^{p}\sigma_{ii}^4}{p}\ , \tag{4.10}$$

and a_i's satisfying (2.17) have been defined in (2.16).

5 Two Examples

In this section, we test the hypotheses of spherecity and diagonality of the following two data sets.

Colon Datasets

In this dataset, expression levels of 40 tumors and 22 normal colon tissues for 6500 human genes are measured using the Affymetrix technology. A selection of 2000 genes with highest minimal intensity across the samples has been made by Alon, Barkai, Motterman, Gish, Mack, and Levine (1999). Thus $p = 2000$, and the degrees of freedom available to estimate the covariance matrix is only 60.

Leukemia Datasets

This dataset contains gene expression level of 72 patients either suffering from acute lymphoblastic leukemia (ALL, 47 cases) or acute myeloid leukemia (AML, 25 cases) and was obtained from Affymetrix oligonucleotide microarrays. More information can be found in Golub, Slonim, Huard, Gassenbeek, Mesirov, Coller, Loh, and Downing (1999); following the protocol in Dudoit, Fridlyand, and Speed (2002), Dettling and Buhlman (2002) preprocess them

by thresholding, filtering, a logarithmic transformation and standardization, so that the data finally comprise the expression values of $p = 3571$ genes, and the degrees of freedom available for estimating the covariance is only 70.

These data are publicly available at "http://www.molbio.princeton.edu/colondata". A base 10 logarithmic transformation is applied.

The description of the above datasets and preprocessing are due to Dettling and Buhlman (2002), except that we do not process the datasets such that each tissue sample has zero mean and unit variance across genes, which is not explainable in our framework. We roughly check the normality by QQ-plotting around 50 genes selected randomly. The results are nearly satisfactory.

For testing spherecity of the colon data, the value of the test statistic $Q_1 = 82086.322$ and that of $T_1 = 2771.654$, see (2.8) and (2.10). Thus the hypothesis of spherecity is rejected by both tests with p-value $= 0$ in each case. For testing the diagonality of the colon data, the value of the test statistic $Q_2 = \infty$, and $T_2 = 2005.894$, see (3.2) and (3.4). Thus, the hypothesis of diagonality is also rejected with p-value $= 0$ in each case.

For the Leukemia data, the value of the two statistics for testing spherecity is given by $Q_1 = 86210.830$, and $T_1 = 2294.918$ respectively. Hence, the spherecity hypothesis is rejected by both tests with p-value zero in each case. For testing diagonality, the value of the test statistics are $Q_2 = 2669.243$, and $T_2 = 1275.528$. Thus both tests reject the diagonality hypothesis of the Leukemia data at p-value $= 0$ in each case.

It would thus appear that the assumption of diagonality made by Dudoit et al. (2002) in not supported by the data in both the examples.

6 Attained Significant Level

In order to check as to how good the normal approximations are for the six statistics Q_1, Q_2, Q_3 and T_1, T_2, T_3, we carry out simulation. A random sample of size n+1 is drawn from $N_p(\mathbf{0}, I)$ and replicated 1000 times. All the six statistics are calculated for a sample and the percentage of times they exceed z_α is recorded, where z_α is the upper $100\alpha\%$ point of the normal cdf. Tables 1-6 present these percentages for $\alpha = 0.05$. We call it "attained significance level (ASL)". For the test Q_1, Q_2, Q_3, the attained significance level is close to α when n and p are not close to each other.

For the tests T_1, T_2, T_3, no such restriction on n and p is needed and attained significance is close to α, usually much better when $n \geq 20$ and or $p \geq 20$.

Tables 1-6 around here

7 Power Comparison

In this section, we compare the power of T_1 with Q_1, T_2 with Q_2 and T_3 with Q_3. The power comparison is based on simulation. We first carry out the simulation to obtain the significance points for all six statistics as done in Section 5. However, here we calculate $T_{i\alpha}$ and $Q_{i\alpha}$, $i = 1, 2, 3$ such that under the hypothesis

$$P\{T_i > T_{i\alpha}\} = \alpha, \text{ and } P\{Q_i > Q_{i\alpha}\} = \alpha, \ i = 1, 2, 3 \ .$$

We have chosen $\alpha = 0.05$. We simulate again. A sample of size $n + 1$ is drawn from $N_p(\mathbf{0}, \wedge)$, where $\wedge = \text{diag}(\lambda_1, \ldots, \lambda_p)$. The values of λ_i are

obtained by taking p iid observations from the uniform distribution over the range (0.5, 1.5). The sample is replicated 1000 times. The percentages of times that the values of statistics T_i and Q_i, $i = 1, 2$, exceed $T_{i\alpha}$ and $Q_{i\alpha}$ respectively are recorded. These are the stimulated power of the four statistics T_1, T_2, Q_1 and Q_2, given in Tables 7-10.

For the statistics T_3 and Q_3, $n+1$ samples are cdrawn from $N_p(\mathbf{0}, DRD)$ where $R = (r_{ij})$ with

$$r_{ij} = \frac{1}{6}^{\frac{1}{2}|i-j|},$$

and $D = \text{diag}(d_1, \ldots, d_p)$, $d_i \sim U(0.5, 1.5)$. The sample is replicated 1000 times and the power of the tests T_3 and Q_3 are obtained. The percentages of the values of the statistics T_3 and Q_3 exceeding $T_{3\alpha}$ and $Q_{3\alpha}$ respectively are the power of these tests shown in Tables 11-12.

Tables 7-12 around here

8 Conclusion

In this paper, we have proposed adapted versions of the likelihood ratio tests available when $n > p$ to the case when $n < p$ for testing sphericity and testing that the covariance matrix is an identity matrix. The advantage of such tests is that the program and the formula for the distribution can be obtained by just interchanging n and p in the available formula for the case $n > p$, except that all the test statistics must be expressed in terms of the eigenvalues of nS. The performance of these two tests is comparable to the proposed tests of Srivastava (2005). For testing the independency of the p variables or

equivalently the diagonality of the covariance matrix, both tests perform well.

Acknowledgement

This research was supported by the Natural Science and Engineering Research Council of Canada. The computation was carried out by Yau Liu to whom I express my sincerest thanks.

REFERENCES

Alon, U., N. Barkai, D Motterman, K. Gish, S. Mack, and J. Levine. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS,* 6745-6750.

Carter, E. M. and Srivastava, M.S. (1977). Monotonicity of the power functions of the modified likelihood ratio criteria for the homogeneity of variances and of the sphericity test. *J. Multiv. Analy.*, **7**, 229-233.

Dettling, M., and P. Buhlman (2002). Boosting for tumor classification with gene expression data. *Bioinformatics,* 1-9.

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

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

Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeck, J. Mesirov, H. Coller, M. Loh, and J. Dowing (1999). Molecular classification of cancer:class discovery and class prediction by gene expression monitoring *Science,* 531-537.

Nagao, H. (1967). Monotonicity of the modified likelihood ratio test for the covariance matrix *J. Sci. Hiroshima Univ.,* **A131**, 147-150.

Srivastava, M.S. (2005). Some tests concerning the covariance matrix in high-dimensional data. *J. Japan Stat. Soc.,* **35**. [redacted]. 251-272.

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

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

Table 1 ASL^* of T_1 Test Under H_1
Sample from N(0,1), n=N-1

	n=20	n=30	n=60	n=100
p=60	0.053	0.050	0.052	0.048
p=100	0.050	0.045	0.049	0.041
p=150	0.050	0.058	0.053	0.048
p=200	0.046	0.058	0.053	0.048
p=250	0.070	0.051	0.046	0.048
p=300	0.043	0.058	0.055	0.059
p=400	0.048	0.055	0.049	0.047

*ASL-Attained Significance Level

Table 2 ASL^* of Q_1 Test Under H_1
Sample from N(0,1), n=N-1

	n=20	n=30	n=60	n=100
p=60	0.060	0.062	**NA**	0.147
p=100	0.053	0.045	0.056	**NA**
p=150	0.057	0.045	0.056	0.571
p=200	0.051	0.052	0.052	0.157
p=250	0.045	0.059	0.058	0.064
p=300	0.032	0.045	0.052	0.066
p=400	0.049	0.046	0.054	0.056

*ASL-Attained Significance Level

Table 3 ASL^* of T_2 Test Under H_2
Sample from N(0,1), n=N-1

	n=20	n=30	n=60	n=100
p=60	0.057	0.051	0.042	0.064
p=100	0.067	0.046	0.047	0.056
p=150	0.043	0.060	0.056	0.049
p=200	0.049	0.055	0.050	0.046
p=250	0.048	0.054	0.045	0.047
p=300	0.063	0.061	0.049	0.065
p=400	0.058	0.055	0.053	0.047

*ASL-Attained Significance Level

Table 4 ASL^* of Q_2 Test Under H_2
Sample from N(0,1), n=N-1

	n=20	n=30	n=60	n=100
p=60	0.057	0.065	**NA**	0.178
p=100	0.050	0.060	0.163	**NA**
p=150	0.052	0.056	0.066	0.536
p=200	0.049	0.040	0.056	0.169
p=250	0.066	0.044	0.060	0.088
p=300	0.048	0.049	0.053	0.066
p=400	0.054	0.051	0.055	0.057

*ASL-Attained Significance Level

Table 5 ASL^* of T_3 Test Under H_3
Sample from N(0,1), n=N-1

	n=20	n=30	n=60	n=100
p=60	0.054	0.050	0.044	0.037
p=100	0.050	0.051	0.049	0.049
p=150	0.061	0.037	0.050	0.055
p=200	0.048	0.056	0.059	0.054
p=250	0.055	0.055	0.060	0.049
p=300	0.057	0.049	0.045	0.048
p=400	0.044	0.054	0.051	0.051

*ASL-Attained Significance Level

Table 6 ASL^* of Q_3 Test Under H_3
Sample from N(0,1), n=N-1

	n=20	n=30	n=60	n=100
p=60	0.061	0.067	0.055	0.052
p=100	0.051	0.056	0.044	0.061
p=150	0.055	0.053	0.057	0.044
p=200	0.043	0.059	0.055	0.052
p=250	0.060	0.054	0.044	0.050
p=300	0.038	0.046	0.052	0.060
p=400	0.042	0.049	0.067	0.045

*ASL-Attained Significance Level

Table 7 Power of T_1 Test Under A_1
(1) 1000 sample from N(0,1) to simulate T_{1a}, a=0.05, for each pair of (p,n).
(2) 1000 sample from N(0,D) to obtain $T_1, P(T_1 > T_{1a})$=power.
(3) D=diag$(d_1, \ldots, d_p)$, where $d_i \sim U(0.5, 1.5)$.

	n=20	n=30	n=60	n=100
p=60	0.284	0.350	0.495	0.979
p=100	0.207	0.436	0.887	0.940
p=150	0.165	0.400	0.725	0.986
p=200	0.194	0.362	0.730	0.993
p=250	0.177	0.334	0.839	0.992
p=300	0.221	0.352	0.771	0.984
p=400	0.182	0.329	0.721	0.989

Table 8 Power of Q_1 Test Under A_1
(1) 1000 sample from N(0,1) to simulate Q_{1a}, a=0.05, for each pair of (p,n).
(2) 1000 sample from N(0,D) to obtain $Q_1, P(Q_1 > Q_{1a})$=power.
(3) D=diag$(d_1, \ldots, d_p)$, where $d_i \sim U(0.5, 1.5)$.

	n=20	n=30	n=60	n=100
p=60	0.162	0.352	0.169	0.786
p=100	0.208	0.258	0.605	0.285
p=150	0.194	0.344	0.693	0.923
p=200	0.234	0.290	0.755	0.953
p=250	0.188	0.336	0.818	0.963
p=300	0.218	0.337	0.735	0.985
p=400	0.199	0.340	0.752	0.988

Table 9 Power of T_2 Test Under A_2
(1) 1000 sample from N(0,1) to simulate T_{2a}, a=0.05, for each pair of (p,n).
(2) 1000 sample from N(0,D) to obtain $T_2, P(T_2 > T_{2a})$=power.
(3) D=diag$(d_1, \ldots, d_p)$,where $d_i \sim U(0.5, 1.5)$.

	n=20	n=30	n=60	n=100
p=60	0.159	0.389	0.748	0.909
p=100	0.275	0.360	0.682	0.992
p=150	0.261	0.300	0.727	0.985
p=200	0.244	0.339	0.675	0.984
p=250	0.165	0.287	0.714	0.994
p=300	0.193	0.361	0.724	0.992
p=400	0.196	0.376	0.735	0.991

Table 10 Power of Q_2 Test Under A_2
(1) 1000 sample from N(0,1) to simulate Q_{2a}, a=0.05, for each pair of (p,n).
(2) 1000 sample from N(0,D) to obtain $Q_2, P(Q_2 > Q_{2a})$=power.
(3) D=diag$(d_1, \ldots, d_p)$,where $d_i \sim U(0.5, 1.5)$.

	n=20	n=30	n=60	n=100
p=60	0.189	0.237	0.148	0.896
p=100	0.168	0.300	0.618	0.215
p=150	0.209	0.283	0.689	0.899
p=200	0.186	0.354	0.626	0.981
p=250	0.256	0.348	0.671	0.946
p=300	0.276	0.386	0.780	0.977
p=400	0.179	0.307	0.760	0.987

Table 11 Power of T_3 Test Under A_3
(1) 1000 sample from N(0,1) to simulate T_{3a}, a=0.05, for each pair of (p,n).
(2) 1000 sample from N(0,D) to obtain $T_3, P(T_3 > T_{3a})$=power.
(3) $\Sigma = D_1 * R_1 * D_1$ which is the variance matrix for sample generating.
(4) D=diag$(d_1, \ldots, d_p)$,where $d_i \sim U(0.5, 1.5)$. R_1 is correlation matrix with $e_{ij} = (k/6 * *(|i-j|/2)$, where k=1.

	n=20	n=30	n=60	n=100
p=60	0.861	0.987	1.000	1.000
p=100	0.807	0.988	1.000	1.000
p=150	0.849	0.992	1.000	1.000
p=200	0.849	0.989	1.000	1.000
p=250	0.842	0.989	1.000	1.000
p=300	0.833	0.993	1.000	1.000
p=400	0.867	0.988	1.000	1.000

Table 12 Power of Q3 Test Under A3
(1) 1000 sample from N(0,1) to simulate Q_{3a}, a=0.05, for each pair of (p,n).
(2) 1000 sample from N(0,D) to obtain $Q_3, P(Q_3 > Q_{3a})$=power.
(3) $\Sigma = D_1 * R_1 * D_1$ which is the variance matrix for sample generating.
(4) D=diag$(d_1, \ldots, d_p)$,where $d_i \sim U(0.5, 1.5)$. R_1 is correlation matrix with $e_{ij} = (k/6 * *(|i-j|/2)$, where k=1.

	n=20	n=30	n=60	n=100
p=60	0.933	1.000	1.000	1.000
p=100	0.960	1.000	1.000	1.000
p=150	0.963	1.000	1.000	1.000
p=200	0.966	1.000	1.000	1.000
p=250	0.952	1.000	1.000	1.000
p=300	0.962	1.000	1.000	1.000
p=400	0.953	1.000	1.000	1.000