ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 8, 2004

Two new multivariate tests, in particular for a high dimension

JÜRGEN LÄUTER

ABSTRACT. Two new test statistics for the multivariate one-sample problem are introduced that are applicable to normally distributed data in all cases with a sample size $n \geq 2$ and a dimension $p \geq 1$. The dimension may also be greater than the sample size. The tests are based on the theory of spherical distributions. They utilize the principal components of the total sums of products matrix of the given data. Under the null hypothesis, the statistics are exactly beta distributed. The performance of the tests is investigated by simulations. Finally, the methods are applied to high-dimensional data from gene expression analysis (dimension $p = 12625$).

1. Introduction

Given n independent p-dimensional normally distributed data vectors

$$\mathbf{x}'_{(j)} = \begin{pmatrix} x_{j1} & \dots & x_{jp} \end{pmatrix} \sim \mathrm{N}_p(\boldsymbol{\mu}', \boldsymbol{\Sigma}) \qquad (j = 1, \dots, n;\ n \geq 2;\ p \geq 1), \quad (1)$$

the one-sample mean-value null hypothesis is

$$\boldsymbol{\mu} = \mathbf{0}. \quad (2)$$

The covariance matrix $\boldsymbol{\Sigma}$ is supposed to be an unknown positive definite $p \times p$ matrix. The n data vectors are represented as an $n \times p$ data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}'_{(1)} \\ \vdots \\ \mathbf{x}'_{(n)} \end{pmatrix} \sim \mathrm{N}_{n \times p}(\mathbf{1}_n \boldsymbol{\mu}', \mathbf{I}_n \otimes \boldsymbol{\Sigma}). \quad (3)$$

Here, $\mathbf{1}_n$ is the $n \times 1$ vector consisting of ones; $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Received October 24, 2003.
2000 *Mathematics Subject Classification.* 62F03, 62H15, 62H25.
Key words and phrases. Multivariate test, exact test, spherical distribution, principal components.

In classical multivariate analysis, Hotelling's well-known T^2 test is available for such a testing problem. However, this test requires to maintain the condition $p < n$. Furthermore, difficulties of overfitting arise which impair the results.

Another class of multivariate tests, the class of so-called spherical tests, has been developed in the last years [Läuter (1996), Läuter, Glimm and Kropf (1998)]. These tests

- are not based on the method of least squares and the maximum likelihood method,
- are not invariant under affine transformations of the p variables,
- are based on the theory of spherical distributions,
- are applicable also for situations with $p \gg n$.

In this paper, a special type of spherical multivariate tests is proposed. The new tests are based on the principal components of the total sums of products matrix $\mathbf{X}'\mathbf{X}$. They have the remarkable property that the rank of $\mathbf{X}$ can be fully exhausted by the principal components.

2. Derivation of the tests

We start from the $p \times p$ eigenvalue problem

$$(\mathbf{X}'\mathbf{X})\mathbf{D} = \mathbf{D}\mathbf{\Lambda}, \qquad \mathbf{D}'\mathbf{D} = \mathbf{I}_q \tag{4}$$

or the corresponding scale-adjusted eigenvalue problem

$$(\mathbf{X}'\mathbf{X})\mathbf{D} = \mathrm{Diag}\,(\mathbf{X}'\mathbf{X})\mathbf{D}\mathbf{\Lambda}, \qquad \mathbf{D}'\,\mathrm{Diag}\,(\mathbf{X}'\mathbf{X})\mathbf{D} = \mathbf{I}_q. \tag{5}$$

Here, $\mathbf{D}$ is the matrix consisting of the eigenvectors pertaining to the q largest eigenvalues λ_h of $\mathbf{X}'\mathbf{X}$, and $\mathbf{\Lambda}$ is the $q \times q$ diagonal matrix of these eigenvalues. As the eigenvectors of $\mathbf{X}'\mathbf{X}$ are unique only up to the signs of reversal in the direction, we additionally need a rule to determine these signs from $\mathbf{X}'\mathbf{X}$. This topic is addressed below. Thus, the data matrix $\mathbf{X}$ of the p given variables is replaced by a matrix $\mathbf{Z}$ of q principal components:

$$\mathbf{Z} = \mathbf{X}\mathbf{D} \quad \text{with} \quad \mathbf{Z}'\mathbf{Z} = \mathbf{\Lambda}. \tag{6}$$

The number q may take any value with

$$1 \leq q \leq \min(n, p). \tag{7}$$

In the case of $q = \min(n, p)$, the information of $\mathbf{X}$ is completely exhausted by the principal components. Then for the eigenvalue problem (4), the spectral decomposition

$$\mathbf{X}'\mathbf{X} = \mathbf{D}\mathbf{\Lambda}\mathbf{D}' \tag{8}$$

is valid, the n columns of $\mathbf{X}'$ lie in the space spanned by the eigenvectors:

$$\mathbf{X}' = \mathbf{D}\mathbf{D}'\mathbf{X}', \qquad \mathrm{rank}(\mathbf{X}) = q \quad \text{with probability one}. \tag{9}$$

For the second eigenvalue problem (5), the corresponding equations

$$\mathbf{X'X} = \mathrm{Diag}(\mathbf{X'X})\mathbf{D}\boldsymbol{\Lambda}\mathbf{D'}\,\mathrm{Diag}(\mathbf{X'X}), \qquad \mathbf{X'} = \mathrm{Diag}(\mathbf{X'X})\mathbf{D}\mathbf{D'}\mathbf{X'} \tag{10}$$

are fullfilled in this special case.

To construct a new test statistic, we note that

$$\mathrm{B} = \frac{1}{n}(\mathbf{1}_n'\mathbf{v})^2 = \frac{1}{n}\mathbf{1}_n'\mathbf{v}\mathbf{v}'\mathbf{1}_n \;\sim\; \mathrm{B}(\frac{1}{2}, \frac{n-1}{2}), \tag{11}$$

where $\mathbf{v}$ denotes an $n \times 1$ random vector with the standard spherical distribution satisfying the condition $\mathbf{v'v} = 1$ (uniform distribution on the n-dimensional unit sphere). On the right-hand side, the well-known beta distribution with the parameters $\frac{1}{2}$ and $\frac{n-1}{2}$ occurs. Given the data matrix $\mathbf{X} \sim \mathrm{N}_{n\times p}(\mathbf{1}_n\boldsymbol{\mu}', \mathbf{I}_n \otimes \boldsymbol{\Sigma})$ under the null hypothesis $\boldsymbol{\mu} = \mathbf{0}$, the random matrix $\mathbf{U} = \mathbf{Z}(\mathbf{Z'Z})^{-1/2} = \mathbf{Z}\boldsymbol{\Lambda}^{-1/2}$ has the $n \times q$ standard left-spherical distribution with $\mathbf{U'U} = \mathbf{I}_q$ [Fang and Zhang (1990)]. Thus, $\mathbf{u} = \mathbf{U}\mathbf{1}_q\frac{1}{q^{1/2}}$ follows the $n \times 1$ standard spherical distribution and, therefore,

$$\mathrm{B}_1 = \frac{1}{n}(\mathbf{1}_n'\mathbf{u})^2 = \frac{1}{nq}(\mathbf{1}_n'\mathbf{U}\mathbf{1}_q)^2 = \frac{1}{nq}(\mathbf{1}_n'\mathbf{Z}\boldsymbol{\Lambda}^{-1/2}\mathbf{1}_q)^2 = \frac{1}{nq}(\mathbf{1}_n'\mathbf{X}\mathbf{D}\boldsymbol{\Lambda}^{-1/2}\mathbf{1}_q)^2 \tag{12}$$

becomes $\mathrm{B}(\frac{1}{2}, \frac{n-1}{2})$ distributed. If we define the mean vector of the data matrix $\mathbf{X}$,

$$\bar{\mathbf{x}}' = \frac{1}{n}\mathbf{1}_n'\mathbf{X}, \tag{13}$$

the final expression of B_1 is obtained:

$$\mathrm{B}_1 = \frac{n}{q}(\bar{\mathbf{x}}'\mathbf{D}\boldsymbol{\Lambda}^{-1/2}\mathbf{1}_q)^2 \;\sim\; \mathrm{B}(\frac{1}{2}, \frac{n-1}{2}). \tag{14}$$

In the elementwise notation we have

$$\mathrm{B}_1 = \frac{n}{q}(\sum_{i=1}^{p}\sum_{h=1}^{q}\bar{x}_i d_{ih}\lambda_h^{-1/2})^2. \tag{15}$$

Possibly, it is a disadvantage of this statistic that the lambda term appears with a negative exponent. Thus, the eigenvectors corresponding to the smallest eigenvalues λ_h can get large weight in this expression. Therefore, a second beta statistic will be introduced in the following which avoids this deficiency.

We start again from the eigenvalue problem (4) or (5) and the corresponding $p \times q$ score matrix $\mathbf{Z} = \mathbf{XD}$ in (6). Recall that $\mathbf{Z}$ is left-spherically distributed under the null hypothesis $\boldsymbol{\mu} = \mathbf{0}$. Then, the $n \times 1$ vectors

$$\mathbf{z} = \mathbf{Z}\mathbf{1}_q \tag{16}$$

and

$$\mathbf{u} = \mathbf{z}(\mathbf{z}'\mathbf{z})^{-1/2} = \mathbf{X}\mathbf{D}\mathbf{1}_q(\mathbf{1}_q'\mathbf{\Lambda}\mathbf{1}_q)^{-1/2} = \mathbf{X}\mathbf{D}\mathbf{1}_q\frac{1}{(\lambda_1+\ldots+\lambda_q)^{1/2}} \qquad (17)$$

are also spherically distributed. Because of $\mathbf{u}'\mathbf{u}=1$, the second beta statistic is obtained:

$$\mathrm{B}_2 = \frac{(\mathbf{1}_n'\mathbf{u})^2}{n} = \frac{(\mathbf{1}_n'\mathbf{X}\mathbf{D}\mathbf{1}_q)^2}{n(\lambda_1+\ldots+\lambda_q)} = \frac{n(\overline{\mathbf{x}}'\mathbf{D}\mathbf{1}_q)^2}{\lambda_1+\ldots+\lambda_q} \sim \mathrm{B}(\frac{1}{2},\frac{n-1}{2}). \qquad (18)$$

In the elementwise notation we have

$$\mathrm{B}_2 = \frac{n}{\lambda_1+\ldots+\lambda_q}(\sum_{i=1}^{p}\sum_{h=1}^{q}\overline{x}_i d_{ih})^2. \qquad (19)$$

For the test statistics B_1 and B_2, the orientation of the eigenvectors $\mathbf{d}_1,\ldots,\mathbf{d}_q$, which are the columns of $\mathbf{D}$, is important. The expressions of B_1 and B_2 change if the sign of an eigenvector is reversed. To attain the exact null distribution, the orientation must be defined in a unique way as a function of $\mathbf{X}'\mathbf{X}$. In the case of the eigenvalue problem (4), we propose the following approach: Select the variable $i_{\max}$ with the largest absolute weight in the first eigenvector $\mathbf{d}_1$. Then, determine the directions of all eigenvectors $\mathbf{d}_1,\ldots,\mathbf{d}_q$ in such a way that $d_{i_{\max}h}$ $(h=1,\ldots,q)$ become all positive. Thus, the directions are uniquely derived from $\mathbf{X}'\mathbf{X}$. This strategy implies that the factorial effects reinforce each other mutually, at least in the dominant variable $i_{\max}$. In the case of the scale-adjusted eigenvalue problem (5), the variable $i_{\max}$ is used which has the largest absolute value in the vector $[\mathrm{Diag}(\mathbf{X}'\mathbf{X})]^{1/2}\mathbf{d}_1$.

Of course, the orientation of eigenvectors may also be defined in another way if there are specific reasons.

For $q=1$, B_1 and B_2 yield the same results.

3. Simulations

In this section we will present simulation results of the beta tests treated above. At first, the power of the B_2 test is compared with the power of Hotelling's well-known test. We consider different cases of the one-factor parameter structure

$$\mathbf{\Sigma} = \kappa\mathbf{I}_p + \boldsymbol{\vartheta}\boldsymbol{\vartheta}', \quad \boldsymbol{\mu}' = \delta\boldsymbol{\vartheta}', \qquad (20)$$

where $\boldsymbol{\vartheta}$ is a $p\times 1$ scale vector, κ is the error variance of the p single variables, δ is a non-centrality coefficient. The principal components are determined from the eigenvalue problem (4). The level of significance is $\alpha = 0.05$. Simulations are performed with 10^4 replications. The following 4 models were examined by simulation.

1. A symmetric parameter structure with $p < n$.
Suppose $p = 4$, $n = 6$, $q = 4$, $\Delta^2 = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = 4$, $\boldsymbol{\mu}' = (\ \mu \quad \mu \quad \mu \quad \mu\ )$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}.$$

The power values for the correlation coefficients $\rho = 0.20, 0.40, 0.60, 0.90$ are given by the table

correlation ρ	0.20	0.40	0.60	0.90
power of B_2	0.6870	0.8170	0.9009	0.9547
power of Hotelling's test	0.3008	0.2986	0.2945	0.2970

.

It can be seen that test B_2 has much higher power than Hotelling's test. All four principal components are utilized in this example, according to the rank 4 of $\mathbf{X}$.

2. A non-symmetric parameter structure with $p < n$.
Set again $p = 4$, $n = 6$, $q = 4$, $\Delta^2 = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = 4$, but $\boldsymbol{\mu}' = (\ \mu \quad \mu \quad 0 \quad 0\ )$,

$$\boldsymbol{\Sigma} = (1-\rho)\mathbf{I}_4 + \rho \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} (\ 1 \quad 1 \quad 0 \quad 0\ ) = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1-\rho & 0 \\ 0 & 0 & 0 & 1-\rho \end{pmatrix}.$$

The two first variables contribute genuine one-factor information, and the other two represent "independent noise". The following table provides the power values:

correlation ρ	0.20	0.40	0.60	0.90
power of B_2	0.6357	0.7434	0.8451	0.9467
power of Hotelling's test	0.3003	0.2958	0.2951	0.3052

.

3. A symmetric parameter structure with $p > n$.
In such a case, Hotelling's test cannot be applied. Suppose $p = 10$, $n = 6$, $q = 6$, $\Delta^2 = 4$, $\boldsymbol{\mu}' = (\ \mu \quad \mu \quad \ldots \quad \mu\ )$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots \\ \vdots & \rho & 1 & \rho \\ \rho & \vdots & \rho & 1 \end{pmatrix}.$$

Once more, the number of the principal components coincides with the rank of $\mathbf{X}$. The following power values are obtained:

correlation ρ	0.20	0.40	0.60	0.90
power of B_2	0.5432	0.7665	0.8804	0.9591

4. A non-symmetric parameter structure with $p > n$.
Set again $p = 10$, $n = 6$, $q = 6$, $\Delta^2 = 4$, but $\boldsymbol{\mu}' = \mu \left(\begin{array}{cc} \mathbf{1}_5' & \mathbf{0}_5' \end{array} \right)$,

$$\boldsymbol{\Sigma} = (1-\rho)\mathbf{I}_{10} + \rho \begin{pmatrix} \mathbf{1}_5 \\ \mathbf{0}_5 \end{pmatrix} \left(\begin{array}{cc} \mathbf{1}_5' & \mathbf{0}_5' \end{array} \right).$$

In this example, the first five variables yield the substantial information. The power values are

correlation ρ	0.20	0.40	0.60	0.90
power of B_2	0.4043	0.6365	0.8016	0.9450

The B_2 test is superior to the B_1 test in many applications. However, there are also situations, where B_1 attains higher power than B_2. For example, for $p = q = 2$, $n = 12$,

$$\boldsymbol{\mu}' = \left(\begin{array}{cc} 1 & 0 \end{array} \right), \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1.6667 & 1 \\ 1 & 1.6667 \end{pmatrix},$$

the B_1 test has the power 0.7310, but the B_2 test only 0.6870, if $\alpha = 0.05$. Hotelling's test gives the power value 0.7304 in the simulation.

4. Application to gene expression data

We consider gene expression data by M. Eszlinger, K. Krohn and R. Paschke (University of Leipzig) in patients with cold nodules in the thyroid gland. Tissue samples of the nodule and of the surrounding have been taken from 15 patients. The data of the tissue samples have been recorded on Affymetrix gene chips referring to 12625 genes. In this example we use the logarithmic expression values, and we consider their differences between nodule and surrounding for each of the 12625 genes in the 15 patients. These differences are approximately normally distributed.

In this case, $p = 12625$ and $n = 15$. We want to apply the B_1 and B_2 test. The eigenvalue problem (4) is used. However, there are some difficulties in the implementation because the total sums of products matrix $\mathbf{X}'\mathbf{X}$ has the order 12625×12625. To avoid handling this matrix, we can exploit the duality property of the eigenvalue problem

$$(\mathbf{X}\mathbf{X}')\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}, \qquad \mathbf{U}'\mathbf{U} = \mathbf{I}_q, \tag{21}$$

where the matrix $\mathbf{XX}'$ is only of the order 15×15. The eigenvalues of (4) and (21) are the same, and the eigenvectors are related via

$$\mathbf{D} = \mathbf{X}'\mathbf{U}\mathbf{\Lambda}^{-1/2}. \tag{22}$$

Therefore,

$$\mathbf{Z} = \mathbf{XD} = \mathbf{XX}'\mathbf{U}\mathbf{\Lambda}^{-1/2} = \mathbf{U}\mathbf{\Lambda}^{1/2}, \tag{23}$$

that is, the matrix $\mathbf{U}$ is the same as in Section 2. Thus, the beta tests can be written as

$$\mathrm{B}_1 = \frac{n}{q}(\overline{\mathbf{u}}'\mathbf{1}_q)^2, \qquad \mathrm{B}_2 = \frac{n(\overline{\mathbf{u}}'\mathbf{\Lambda}^{1/2}\mathbf{1}_q)^2}{\lambda_1 + \ldots + \lambda_q}, \tag{24}$$

with the "mean vector" $\overline{\mathbf{u}}' = \frac{1}{n}\mathbf{1}_n'\mathbf{U}$.

The orientation of the eigenvectors $\mathbf{d}_h$ and $\mathbf{u}_h$ $(h = 1, \ldots, q)$ according to the sign regulation at the end of Section 2 is attained over (22). We find $i_{\max} = 6746$, the Affimetrix notation of this gene is 36681_at. The corresponding elements $\overline{u}_h$ of the vectors $\overline{\mathbf{u}}'$ (for $q = 1, \ldots, 15$) are

$$\overline{u}_{1,\ldots,15} = 10^{-2}(-21, -9, -3, -2, -6, -1, -1, -2, -4, +3, -1, -5, +2, -5, -6).$$

The most mean values $\overline{u}_h$ are negative, in particular, the largest by absolute value. Therefore, all cumulative sums $\sum_{h=1}^{q} \overline{u}_h$ and $\sum_{h=1}^{q} \overline{u}_h \lambda_h^{1/2}$ $(q = 1, \ldots, 15)$ are negative, and the factorial effects are supporting each other (in the sense of the absolute value). This is a confirmation of our rule for the orientation of the eigenvectors. A special conclusion from this consideration is that, in total, the cold thyroid nodules must have a reduced gene expression level in comparison to the surrounding tissue.

We obtain the following P values in the B_1 test for $q = 1, \ldots, 15$:
P = .0002, .0002, .002, .007, .004, .009, .01, .02, .02, .04, .04, .03, .04, .03, .02 .
The corresponding P values of the B_2 test are:
P = .0002, .00008, .0005, .002, .0009, .002, .004, .005, .004, .01, .01, .007, .01, .008, .004 .

All P values are smaller than 0.05. Therefore, a significant difference between the nodules and the surrounding is proved for all values of q, if $\alpha = 0.05$. The best result, $P = 0.00008$, arises at the B_2 test in the case of two principal components $(q = 2)$.

References

Fang, K.-T. and Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*. Springer, Berlin.

Läuter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* **52**, 964–970.

Läuter, J., Glimm, E. and Kropf, S. (1998). Multivariate tests based on left-spherically distributed linear scores. *Ann. Statist.* **26**, 1972–1988. Correction: **27**, 1441.

Otto von Guericke University Magdeburg, Mittelstr. 2/151, 39114 Magdeburg, Germany

E-mail address: Juergen.Laeuter@medizin.uni-magdeburg.de