The Annals of Statistics
1998, Vol. 26, No. 5, 1972–1988

MULTIVARIATE TESTS BASED ON LEFT-SPHERICALLY DISTRIBUTED LINEAR SCORES[1]

BY JÜRGEN LÄUTER, EKKEHARD GLIMM AND SIEGFRIED KROPF

Otto von Guericke University

In this paper, a method for multivariate testing based on low-dimensional, data-dependent, linear scores is proposed. The new approach reduces the dimensionality of observations and increases the stability of the solutions. The method is reliable, even if there are many redundant variables. As a key feature, the score coefficients are chosen such that a left-spherical distribution of the scores is reached under the null hypothesis. Therefore, well-known tests become applicable in high-dimensional situations, too. The presented strategy is an alternative to least squares and maximum likelihood approaches. In a natural way, standard problems of multivariate analysis thus induce the occurrence of left-spherical, nonnormal distributions. Hence, new fields of application are opened up to the generalized multivariate analysis. The proposed methodology is not restricted to normally distributed data, but can also be extended to any left-spherically distributed observations.

1. Introduction. In recent years, it has been proved [Hsu (1990a, b), Anderson, Fang and Hsu (1986), Kariya and Sinha (1989), Fang and Zhang (1990), Anderson (1993)] that the classical multivariate linear model tests are valid not only for normally distributed data, but remain exact also in the wider class of spherical and elliptically contoured distributions. This research has shown the robustness, especially the so-called null robustness, of the classical methods. However, it also has revealed the limitations of generalized multivariate analysis because if n independent p-dimensional data vectors are given, exact tests of this type are available only in the special case of the normal distribution. Although the rows of an $n \times p$ left-spherically distributed matrix are uncorrelated, they are generally not stochastically independent.

In 1996, the authors of this paper [Läuter (1996), Läuter, Glimm and Kropf (1996)] proposed a new class of tests for independent, p-dimensional normally distributed observations. These tests are based on linear scores with coefficients ensuring a left-spherical score distribution under the null hypothesis. The score coefficients are determined from the observations via well-defined sums of products matrices. This approach compresses high-dimensional observations into low-dimensional scores which are then analyzed instead of the original data.

Thus, the standard problems of applied multivariate analysis naturally lead to left-spherical, nonnormal matrix distributions. Hence, the existing theory of

Received February 1997; revised May 1998.

[1]Supported by German Governmental Grant BMBF 01ZZ9510.

AMS 1991 *subject classifications*. 62F35, 62H15, 62H20, 62H25, 62J10, 62J15.

Key words and phrases. Multivariate test, linear scores, spherical distribution, generalized multivariate analysis, exact test, null robustness.

spherical and elliptically contoured distributions attains relevance for a broad scope of practical purposes. Moreover, any left-spherically distributed data can also be analyzed by these tests, which initially were derived for normal data only.

The new tests offer surprising opportunities. In contrast to the classical multivariate procedures, exploratory steps of data preprocessing and model choice can be incorporated into a confirmatory analysis without producing a bias. Data preprocessing can be tailored for special applications. If, for example, the data are presumed to have an underlying factorial structure, the data reduction should be done using principal component analysis or factor analysis. In case of a time series, methods of smoothing should be used. If there is reason to believe that valuable and useless variables are both present in the data, selection of variables based on correlations or covariances is the method of choice.

To give an introductory example, the comparison of the mean vectors of two p-dimensional normal distributions $N_p(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$ is considered, where $\boldsymbol{\Sigma}$ is an unknown covariance matrix. The null hypothesis to be tested is $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$. Two independent samples $\mathbf{x}^{(1)}_{(1)}, \ldots, \mathbf{x}^{(1)}_{(n^{(1)})}$ and $\mathbf{x}^{(2)}_{(1)}, \ldots, \mathbf{x}^{(2)}_{(n^{(2)})}$ of the sizes $n^{(1)}$ and $n^{(2)}$, respectively, are assumed. Now, if the p-dimensional coefficient vector $\mathbf{d}$ is defined depending on these samples via a given function $\mathbf{d}(\mathbf{W})$ in which the total sums of products matrix

$$\mathbf{W} = \sum_{k=1}^{2} \sum_{j=1}^{n^{(k)}} \left(\mathbf{x}^{(k)}_{(j)} - \bar{\mathbf{x}}\right)\left(\mathbf{x}^{(k)}_{(j)} - \bar{\mathbf{x}}\right)' \quad \text{with } \bar{\mathbf{x}} = \frac{1}{n^{(1)} + n^{(2)}} \sum_{k=1}^{2} \sum_{j=1}^{n^{(k)}} \mathbf{x}^{(k)}_{(j)} \tag{1}$$

occurs as the argument, then the statistic of the usual univariate two-sample t test

$$t = \sqrt{n^{(1)} + n^{(2)} - 2}\, \frac{\sqrt{n^{(1)} n^{(2)} / (n^{(1)} + n^{(2)})}\,(\bar{z}^{(1)} - \bar{z}^{(2)})}{\sqrt{\sum_{k=1}^{2} \sum_{j=1}^{n^{(k)}} (z^{(k)}_{(j)} - \bar{z}^{(k)})^2}} \quad \text{with}$$

$$\bar{z}^{(k)} = \frac{1}{n^{(k)}} \sum_{j=1}^{n^{(k)}} z^{(k)}_{(j)}, \qquad k = 1, 2 \tag{2}$$

can be calculated for the linear score values

$$z^{(k)}_{(j)} = \mathbf{d}' \mathbf{x}^{(k)}_{(j)}, \qquad k = 1, 2;\ j = 1, \ldots, n^{(k)}. \tag{3}$$

In this case, one can prove that the statistic t has exactly Student's t distribution with $n^{(1)} + n^{(2)} - 2$ degrees of freedom provided the hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ is valid. Therefore the univariate statistic t is suitable for testing the multivariate hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$. The used function $\mathbf{d} = \mathbf{d}(\mathbf{W})$ of the score coefficient vector [or the more general matrix function $\mathbf{D} = \mathbf{D}(\mathbf{W})$] determines a special method of data preprocessing. It should be chosen from the huge number of possibilities according to practical points of view, to attain a high power of the test.

In contrast to classical multivariate theory, the newly proposed methods no longer require the sample size n to be larger than the number p of variables. Redundance in the observed variables will no longer pose a substantial problem. A certain degree of homogeneity of the data is even a prerequisite for the sensible application of the presented theory. Our strategy makes it possible to compensate for small sample sizes by exploiting the large number of variables, thus avoiding numerical and statistical instability of inference.

As a decisive difference between our approach and the classical analysis, the method of least squares and the maximum likelihood approach are abandoned. These latter methods are based on an optimal fit of a model to the data. Thus, they produce instability in the case of a small sample size n and a large number of variables p. Instead, we use more equalizing and smoothing strategies, and we recommend exploiting prior information on models and parameters as far as possible.

The presented theory refers to $n \times p$ left-spherically distributed matrices $\mathbf{X}$. Such random matrices are characterized by the fact that

$$\mathbf{X} =_d \mathbf{CX} \quad \text{for every fixed } n \times n \text{ orthogonal matrix } \mathbf{C} \tag{4}$$

and by a characteristic function of the form $\phi(\mathbf{T}'\mathbf{T})$. The symbol $=_d$ denotes the equality of two distributions [Fang and Zhang (1990)]. This class of distributions is too wide for the construction of optimal tests according to the likelihood ratio criterion. Therefore, other authors [Fang and Zhang (1990), Anderson (1993), Gupta and Varga (1993)] prefer a restricted class of distributions $\mathbf{X}$ with a characteristic function $\psi(\mathrm{tr}(\mathbf{T}'\mathbf{TA})) = \psi(\mathrm{tr}(\mathbf{TAT}'))$, where $\mathbf{A}$ is a $p \times p$ positive definite symmetric matrix. The matrices $\mathbf{X}$ of this class have the representation $\mathbf{X} =_d \mathbf{YA}^{1/2}$ with a vector-spherically distributed matrix $\mathbf{Y}$ [Fang and Zhang (1990), page 96]. The notation $\mathbf{A}^{1/2}$ indicates the positive definite symmetric matrix that satisfies $(\mathbf{A}^{1/2})^2 = \mathbf{A}$. In this paper, we do not intend a mathematically rigorous treatment of the power and optimality of the tests. Therefore we can admit all left-spherical distributions. This paper is primarily focussed on the problem of invariance of the null distribution against different definitions of the score coefficients. However, we would like to emphasize that the test statistics developed here in general will not have the property that their values are invariant under arbitrary p-dimensional affine transformations.

2. The special case of the normal distribution. Consider the $n \times p$ data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}'_{(1)} \\ \vdots \\ \mathbf{x}'_{(n)} \end{pmatrix} \sim N_{n \times p}(\mathbf{M},\ \mathbf{I}_n \otimes \mathbf{\Sigma}) \tag{5}$$

with n independent p-dimensional normally distributed row vectors $\mathbf{x}'_{(j)}$ $(j = 1, \ldots, n)$. Here $\mathbf{I}_n$ is the $n \times n$ identity matrix, the symbol $\otimes$ represents the

Kronecker product. Classical linear multivariate tests concerning the mean structure $\mathbf{M}$ are based on the two stochastically independent $p \times p$ matrices

$$\mathbf{H} = \mathbf{X}'\mathbf{Q}_H\mathbf{X}, \qquad \mathbf{G} = \mathbf{X}'\mathbf{Q}_G\mathbf{X}. \tag{6}$$

$\mathbf{H}$ is the so-called hypothesis sums of products matrix, $\mathbf{G}$ is the residual sums of products matrix, $\mathbf{Q}_H$ and $\mathbf{Q}_G$ are mutually orthogonal $n \times n$ projection matrices, that is,

$$\begin{gathered}\mathbf{Q}_H' = \mathbf{Q}_H = \mathbf{Q}_H^2, \qquad rk(\mathbf{Q}_H) = f_H, \qquad \mathbf{Q}_G' = \mathbf{Q}_G = \mathbf{Q}_G^2, \qquad rk(\mathbf{Q}_G) = f_G, \\ \mathbf{Q}_H\mathbf{Q}_G = \mathbf{0}, \quad f_H + f_G \le n.\end{gathered} \tag{7}$$

The null hypothesis is characterized by $\mathbf{Q}_H\mathbf{M} = \mathbf{0}$, $\mathbf{Q}_G\mathbf{M} = \mathbf{0}$, and hence

$$\mathbf{H} \sim W_p(\mathbf{\Sigma}, f_H), \qquad \mathbf{G} \sim W_p(\mathbf{\Sigma}, f_G) \tag{8}$$

holds under the null hypothesis, where W_p denotes the Wishart distribution. Under an alternative hypothesis, $\mathbf{Q}_H\mathbf{M} = \mathbf{0}$ would be violated.

These sums of products matrices are the starting point for the development of score-based multivariate tests. In addition, the following Theorem 1 includes the $p \times p$ matrix $\mathbf{L}$ which is independent of $\mathbf{H}$ and $\mathbf{G}$. This matrix allows the incorporation of "neutral information" into the test. The score coefficients for the dimension reduction are given by a $p \times q$ random matrix $\mathbf{D}$ which is a fixed function of the argument $\mathbf{H} + \mathbf{G} + \mathbf{L}$. In principle, this function $\mathbf{D} = \mathbf{D}(\mathbf{H} + \mathbf{G} + \mathbf{L})$ may be chosen arbitrarily. However, it should be suitable to the parameter structure conjectured in the data, as far as possible. A more general function $\mathbf{D} = \mathbf{D}(\mathbf{H} + \mathbf{G}, \mathbf{L})$ also could be assumed in the following theorem, where $\mathbf{L}$ is any random variable which is independent of $\mathbf{H}$ and $\mathbf{G}$ but does not necessarily have to be a $p \times p$ matrix. For most applications, though, the restricted argument $\mathbf{H} + \mathbf{G} + \mathbf{L}$ consisting of three independent $p \times p$ matrices $\mathbf{H}, \mathbf{G}$, and $\mathbf{L}$ is adequate (see application (4) of Theorem 1).

THEOREM 1. *Assume* $1 \le f_H$, $1 \le f_G$ *and* $1 \le q \le f_H + f_G$. *Assume a test statistic* $F = F(\mathbf{H}_Z, \mathbf{G}_Z)$ *as a Borel function defined for all* $q \times q$ *positive semidefinite symmetric matrices* $\mathbf{H}_Z$ *and* $\mathbf{G}_Z$ *and satisfying the invariance condition*

$$F(\mathbf{A}\mathbf{H}_Z\mathbf{A}, \mathbf{A}\mathbf{G}_Z\mathbf{A}) = F(\mathbf{H}_Z, \mathbf{G}_Z) \tag{9}$$

for every $q \times q$ *positive definite symmetric matrix* $\mathbf{A}$.

Now, suppose a dimension p *with* $p \ge q$. *Let* $\mathbf{H}, \mathbf{G}$ *and* $\mathbf{L}$ *be three* $p \times p$ *random positive semidefinite symmetric matrices that are mutually stochastically independent, where* $\mathbf{H}$ *and* $\mathbf{G}$ *have the Wishart distributions*

$$\mathbf{H} \sim W_p(\mathbf{\Sigma}, f_H), \qquad \mathbf{G} \sim W_p(\mathbf{\Sigma}, f_G) \tag{10}$$

for a positive definite $\mathbf{\Sigma}$. *Let* $\mathbf{D}$ *be a* $p \times q$ *random matrix defined as a Borel function of* $\mathbf{H} + \mathbf{G} + \mathbf{L}$ *and having rank* q *with probability* 1.

Then the distribution of $F = F(\mathbf{D}'\mathbf{H}\mathbf{D}, \mathbf{D}'\mathbf{G}\mathbf{D})$ *is the same for each* p, *each* $\mathbf{\Sigma}$, *each* $\mathbf{D}$ *function and each suitable* $\mathbf{L}$ *distribution*.

HINTS CONCERNING THE PROOF. The basic idea of the proof has been published in Läuter, Glimm and Kropf (1996). The justification can also be attained by using the more general Theorem 2 of this paper. One has to take into account that the matrices $\mathbf{H}$ and $\mathbf{G}$ have representations

$$\mathbf{H} =_d \sum_{j=1}^{f_H} \mathbf{h}_j \mathbf{h}_j', \qquad \mathbf{G} =_d \sum_{j=1}^{f_G} \mathbf{g}_j \mathbf{g}_j', \tag{11}$$

consisting of $f = f_H + f_G$ independent vectors $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{f_H}, \mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{f_G}$ distributed each according to $N_p(\mathbf{0}, \mathbf{\Sigma})$. Setting

$$\mathbf{X}' = \begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_{f_H} & \mathbf{g}_1 & \mathbf{g}_2 & \cdots & \mathbf{g}_{f_G} \end{pmatrix}, \tag{12}$$

then $\mathbf{X} \sim N_{f \times p}(\mathbf{0}, \mathbf{I}_f \otimes \mathbf{\Sigma})$. In this situation, application of Theorem 2 with $n = f$, $\mathbf{E} = \mathbf{I}_f$, $\mathbf{Q} = \mathbf{Q}_0 = \mathbf{I}_f$, and with the test statistic

$$F\left(\mathbf{Z}'_{f\times q}\begin{pmatrix} \mathbf{I}_{f_H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Z}_{f\times q},\ \mathbf{Z}'_{f\times q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{f_G} \end{pmatrix}\mathbf{Z}_{f\times q}\right) \tag{13}$$

defined for an arbitrary $\mathbf{Z}_{f\times q}$ yields the desired result. Indeed, the assumptions of Theorem 2 are here fulfilled; the test statistic (13) keeps its value if $\mathbf{Z}_{f\times q}$ is replaced by $\mathbf{Z}_{f\times q}\mathbf{A}$, where $\mathbf{A}$ is any $q \times q$ positive definite symmetric matrix; for every fixed value of $\mathbf{L}$, the given coefficient matrix $\mathbf{D}$ can be considered as a Borel function of $\mathbf{X}'\mathbf{X} =_d \mathbf{H} + \mathbf{G}$. □

If the assumptions (5) to (7) and the null hypothesis are true, Theorem 1 provides the distribution of $F = F(\mathbf{Z}'\mathbf{Q}_H\mathbf{Z}, \mathbf{Z}'\mathbf{Q}_G\mathbf{Z})$, where $\mathbf{Z} = \mathbf{X}\mathbf{D}$ is the $n \times q$ score matrix. In general, $\mathbf{Z}$ will no longer be normally distributed, and its row vectors will not be independent. However, Theorem 2 will show that certain important sphericity properties of $\mathbf{Z}$ are secured.

In any case, F is null distributed as if it were defined by $F(\mathbf{X}'_H\mathbf{X}_H,\ \mathbf{X}'_G\mathbf{X}_G)$ in the setting

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_H \\ \mathbf{X}_G \end{pmatrix} \sim N_{(f_H+f_G)\times q}(\mathbf{0}, \mathbf{I}_{f_H+f_G} \otimes \mathbf{I}_q), \tag{14}$$

$$\mathbf{Q}_H = \begin{pmatrix} \mathbf{I}_{f_H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{f_G} \end{pmatrix}, \qquad \mathbf{Q}_G = \begin{pmatrix} \mathbf{0}_{f_H} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{f_G} \end{pmatrix}, \qquad \mathbf{D} = \mathbf{I}_q. \tag{15}$$

The null distribution is the same as in the special case of normally distributed scores. We emphasize that this is true for arbitrarily large dimension p.

For power considerations, Theorem 1 is of no use. In any concrete application, it is necessary to achieve a high power by a suitable definition of the $\mathbf{D}$ function. It is self-evident that the definition must not depend on peculiarities of the given data except $\mathbf{H} + \mathbf{G} + \mathbf{L}$. At best, the method of score building and calculation of coefficients is fixed before beginning the measurements. The multivariate tests described here are "adaptive" in the sense that different strategies of data analysis can be chosen for every surface $\mathbf{H} + \mathbf{G} + \mathbf{L} = \text{const}$.

The multivariate analysis thus offers opportunities which are not given for the univariate parametric inference. Up to now, exact adaptive tests have mostly been considered in the field of nonparametric inference [Büning (1991)].

APPLICATIONS (1) *One-sample test*. The so-called standardized sum test (SS test) in its one-sample version can be used for testing the hypothesis H: $\boldsymbol{\mu} = \mathbf{0}$, if $\mathbf{X} \sim N_{n\times p}(\mathbf{1}_n\boldsymbol{\mu}', \mathbf{I}_n \otimes \boldsymbol{\Sigma})$. Here $\mathbf{1}_n$ is the $n \times 1$ vector consisting of ones only. The SS test for this situation is characterized by the "univariate" F statistic

$$F = \frac{(n-1)\mathbf{H}_z}{\mathbf{G}_z} = \frac{(n-1)\mathbf{d}'\mathbf{X}'\mathbf{Q}_H\mathbf{X}\mathbf{d}}{\mathbf{d}'\mathbf{X}'\mathbf{Q}_G\mathbf{X}\mathbf{d}} = \frac{(n-1)n(\bar{\mathbf{x}}'\mathbf{d})^2}{\mathbf{d}'(\mathbf{X}-\mathbf{1}_n\bar{\mathbf{x}}')'(\mathbf{X}-\mathbf{1}_n\bar{\mathbf{x}}')\mathbf{d}} = \frac{n(\bar{\mathbf{x}}'\mathbf{d})^2}{\mathbf{d}'\mathbf{S}\mathbf{d}} \tag{16}$$

with $\mathbf{Q}_H = (1/n)\mathbf{1}_n\mathbf{1}_n'$, $\mathbf{Q}_G = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n'$ and $\mathbf{d} = [\mathrm{Diag}(\mathbf{X}'\mathbf{X})]^{-1/2}\mathbf{1}_p$. Additionally, $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the estimators of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$, respectively. This special statistic is sensible if the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be expected to be nearly symmetric with respect to the p variables. Here symmetry is defined as identity of all means, identity of all variances and identity of all correlation coefficients. In each case, the statistic (16) follows an exact F distribution with 1 and $n-1$ degrees of freedom under H. Note that $\mathbf{H} + \mathbf{G} = \mathbf{X}'\mathbf{X}$, $\mathbf{L} = \mathbf{0}$, and $q = 1$. This multivariate one-sample test is applicable for arbitrary p and $n \geq 2$. For practical purposes, it is convenient to determine the score vector $\mathbf{z} = \mathbf{X}\mathbf{d}$ first and to do an ordinary univariate F or t test subsequently.

An elementary proof that the statistic (16) has exactly the F distribution with 1 and $n-1$ degrees of freedom can also be given in the following way: if $\mathbf{C}$ is a fixed $n \times n$ orthogonal matrix, the rotated matrix $\mathbf{X}_C = \mathbf{C}\mathbf{X}$ has the same distribution as the original matrix $\mathbf{X}$ under the hypothesis H. Then the corresponding score vectors $\mathbf{z}_C = \mathbf{X}_C\mathbf{d}_C$ and $\mathbf{z} = \mathbf{X}\mathbf{d}$ also possess the same distributions because their coefficients $\mathbf{d}_C$ and $\mathbf{d}$ are uniquely determined by $\mathbf{X}_C'\mathbf{X}_C$ and $\mathbf{X}'\mathbf{X}$, respectively, both following one and the same function. As $\mathbf{X}_C'\mathbf{X}_C = \mathbf{X}'\mathbf{X}$, then $\mathbf{d}_C = \mathbf{d}$ and $\mathbf{z}_C = \mathbf{C}\mathbf{X}\mathbf{d} = \mathbf{C}\mathbf{z}$, that is, $\mathbf{z} = (z_1 \cdots z_n)'$ is spherically distributed and, therefore, $F = (n\bar{z}^2/s_z^2) = (n(\bar{\mathbf{x}}'\mathbf{d})^2/\mathbf{d}'\mathbf{S}\mathbf{d})$ is F distributed according to Fang and Zhang (1990), page 63. The rejection area of the test for a given level of significance α is a cone in the $\mathbf{z}$ space around the equiangular line $z_1 = z_2 = \cdots = z_n$.

Table 1 provides some power values of the SS test for $\alpha = 0.05$ and for symmetric parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a diagonal matrix and $\Delta^2 = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = 16$, $n = 3$. The table has been calculated by simulations of 10^5 replications.

For $p = 1$, the SS test is identical with the usual univariate F test. For $p = 2$, the SS test is obviously much better than Hotelling's T^2 test. For the larger values of p, a comparison is impossible. Further power considerations can be found in the papers by Kropf, Hothorn and Läuter (1997a) and Kropf, Läuter and Glimm (1997b).

TABLE 1
Power values of the SS test for $\alpha = 0.05$ *and* $\boldsymbol{\mu}, \boldsymbol{\Sigma}$

Number of variables p	1	2	4	10	20
Power of the SS test	0.909	0.904	0.896	0.880	0.876
Power of Hotelling's test, if possible	0.909	0.274	—	—	—

(2) *Two-sample test.* In this special application, the q-fold principal component test (PC test) is used for testing H: $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ in a two-sample setup, where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix} \sim N_{(n^{(1)}+n^{(2)})\times p}\left(\begin{pmatrix} \mathbf{1}_{n^{(1)}}\boldsymbol{\mu}^{(1)\prime} \\ \mathbf{1}_{n^{(2)}}\boldsymbol{\mu}^{(2)\prime} \end{pmatrix}, \ \mathbf{I}_{n^{(1)}+n^{(2)}} \otimes \boldsymbol{\Sigma} \right). \tag{17}$$

This test should be applied when a factorial structure is conjectured in the data. Here,

$$\begin{gathered} \mathbf{H} = a(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})', \qquad n = n^{(1)} + n^{(2)}, \\ a = \frac{n^{(1)}n^{(2)}}{n}, \qquad f_H = 1, \\ \mathbf{G} = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}}) - \mathbf{H} = f_G\mathbf{S}, \\ \bar{\mathbf{x}} = \frac{1}{n}(n^{(1)}\bar{\mathbf{x}}^{(1)} + n^{(2)}\bar{\mathbf{x}}^{(2)}), \qquad \bar{\mathbf{X}} = \mathbf{1}_n\bar{\mathbf{x}}', \qquad f_G = n - 2, \end{gathered} \tag{18}$$

with $\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$ and $\mathbf{S}$ being the usual estimators of $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ and $\boldsymbol{\Sigma}$, respectively. For the PC test, the $p \times q$ coefficient matrix $\mathbf{D} = (\mathbf{d}_1 \cdots \mathbf{d}_q)$ is determined as the solution of the eigenvalue problem

$$(\mathbf{H} + \mathbf{G})\mathbf{D} = \mathrm{Diag}(\mathbf{H} + \mathbf{G})\mathbf{D}\boldsymbol{\Lambda}, \tag{19}$$

where $\mathbf{H}+\mathbf{G} = (\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})$, the vectors $\mathbf{d}_j$ $(j = 1, \ldots, q)$ are the eigenvectors corresponding to the q largest eigenvalues, and $\boldsymbol{\Lambda}$ is the diagonal matrix of these eigenvalues. The score matrix is given by $\mathbf{Z} = \mathbf{XD}$. The test is performed using Hotelling's "q-dimensional" T^2,

$$\begin{aligned} T^2 &= \mathrm{tr}(\mathbf{H}_Z\mathbf{G}_Z^{-1}) = \mathrm{tr}((\mathbf{D}'\mathbf{H}\mathbf{D})(\mathbf{D}'\mathbf{G}\mathbf{D})^{-1}) \\ &= \frac{a}{n-2}(\bar{\mathbf{z}}^{(1)} - \bar{\mathbf{z}}^{(2)})'\mathbf{S}_Z^{-1}(\bar{\mathbf{z}}^{(1)} - \bar{\mathbf{z}}^{(2)}) \end{aligned} \tag{20}$$

with $\bar{\mathbf{z}}^{(1)} = \mathbf{D}'\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{z}}^{(2)} = \mathbf{D}'\bar{\mathbf{x}}^{(2)}$, $\mathbf{S}_Z = \mathbf{D}'\mathbf{S}\mathbf{D}$. The expression $((n-q-1)/q)T^2$ is exactly F distributed with q and $n-q-1$ degrees of freedom if the hypothesis H is true. This multivariate two-sample test can be used if $p \geq q \geq 1$ and $n^{(1)} \geq 1$, $n^{(2)} \geq 1$, $n \geq q+2$. According to Theorem 1, the number q of principal components is fixed. However, a generalization of Theorem 1 is possible, where q is determined from $\mathbf{H} + \mathbf{G}$.

(3) *Model choice, selection of variables.* A combination of the PC or the SS test with a selection of variables is suitable under certain conditions. In case

of the one-factor structure of data [Läuter (1992), Läuter, Glimm and Kropf (1996)], for example, the means and covariances are related to each other. This fact may be exploited by selecting the most highly correlated and thus most informative variables from the "correlation" matrix

$$[\mathrm{Diag}(\mathbf{H}+\mathbf{G})]^{-1/2}(\mathbf{H}+\mathbf{G})[\mathrm{Diag}(\mathbf{H}+\mathbf{G})]^{-1/2} \tag{21}$$

in the situation of application (2). The PC test then is performed using only the selected variables. Due to Theorem 1, this procedure does not affect the test's α level. Further practical recommendations concerning the selection of variables can be found in Kropf, Läuter and Glimm (1997b).

(4) *Application of the additional matrix* $\mathbf{L}$. The incorporation of the additional matrix $\mathbf{L}$ with neutral information in Theorem 1 allows using the same scores $\mathbf{Z}=\mathbf{XD}$ for testing different hypotheses in a multivariate model. For example, consider a multivariate two-way classification with orthogonal design. In the classical MANOVA approach, tests of hypotheses are performed by means of various $p\times p$ sums of products matrices $\mathbf{H}_A$, $\mathbf{H}_B$, and $\mathbf{H}_{A\times B}$, say, associated with main effects and interactions. Using Theorem 1, a coefficient vector $\mathbf{d}$ for weighting the p variables may be determined as a function of $\mathbf{H}_A+\mathbf{H}_B+\mathbf{H}_{A\times B}+\mathbf{G}$, where $\mathbf{G}$ is the residual sum of products matrix. The main effect A may be tested by

$$F=\frac{\mathbf{d}'\mathbf{H}_A\mathbf{d}/f_A}{\mathbf{d}'\mathbf{G}\mathbf{d}/f_G}. \tag{22}$$

According to Theorem 1, this statistic is F distributed with f_A and f_G degrees of freedom under H_A, because $\mathbf{L}=\mathbf{H}_B+\mathbf{H}_{A\times B}$ is stochastically independent of $\mathbf{H}_A$ and $\mathbf{G}$, even if the hypotheses H_B and $\mathrm{H}_{A\times B}$ are not true. The same weight vector $\mathbf{d}$ can also be used for testing the effects B and $A\times B$. Of course, in any practical application, one would also have to consider the power of the resulting tests.

(5) *Tests for correlation of variables*. Consider $\mathbf{X}\sim N_{n\times p}(\mathbf{M},\mathbf{I}_n\otimes\mathbf{\Sigma})$ and the null hypothesis H: $\mathbf{M}=\mathbf{1}_n\boldsymbol{\mu}'$, where $\boldsymbol{\mu}$ is a p-dimensional vector. Given a fixed n-dimensional vector $\mathbf{k}$ with $\mathbf{k}'\mathbf{k}=1$ and $\mathbf{k}'\mathbf{1}_n=0$, the contrast $\mathbf{k}'\mathbf{M}$ is investigated. In analogy to application (2), set

$$\mathbf{H}=\mathbf{X}'\mathbf{k}\mathbf{k}'\mathbf{X}=(\mathbf{X}-\bar{\mathbf{X}})'\mathbf{k}\mathbf{k}'(\mathbf{X}-\bar{\mathbf{X}}),\ \ \mathbf{G}=(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})-\mathbf{H}, \tag{23}$$

and then use the beta statistic

$$\mathrm{B}=\frac{\mathbf{H}_z}{\mathbf{H}_z+\mathbf{G}_z}=\frac{\mathbf{d}'\mathbf{H}\mathbf{d}}{\mathbf{d}'(\mathbf{H}+\mathbf{G})\mathbf{d}}=\frac{(\mathbf{k}'(\mathbf{X}-\bar{\mathbf{X}})\mathbf{d})^2}{\mathbf{d}'(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})\mathbf{d}}. \tag{24}$$

Under the null hypothesis, B has a $\mathrm{B}(1/2,((n-2)/2))$ distribution. The coefficient vector $\mathbf{d}$ is defined as a function of $(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})$.

In correlation analysis, these statements may be used to derive a test of independence between a block of $\mathbf{Y}$ and a block of $\mathbf{X}$ variables. Consider

the $n \times (m + p)$ matrix $(\mathbf{Y} \quad \mathbf{X})$. The matrix $\mathbf{X}$ is assumed to have the np-dimensional normal distribution given above. We do not require any distributional properties for $\mathbf{Y}$, except that it is independent of $\mathbf{X}$. Now let

$$\mathbf{k} = \frac{1}{\sqrt{\mathbf{e}'(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})\mathbf{e}}}(\mathbf{Y} - \bar{\mathbf{Y}})\mathbf{e}, \tag{25}$$

where $\mathbf{e}$ is an m-dimensional vector of coefficients uniquely determined by $(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})$. For $\mathbf{Y}$ fixed, a conditional B test of the type (24) is given by

$$\mathrm{B} = \frac{(\mathbf{e}'(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{X} - \bar{\mathbf{X}})\mathbf{d})^2}{\mathbf{e}'(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})\mathbf{e} \cdot \mathbf{d}'(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})\mathbf{d}}. \tag{26}$$

This, of course, is also an unconditional test, because B has the same distribution for all possible choices of $\mathbf{Y}$.

In the special case where $\mathbf{e}$ and $\mathbf{d}$ are the square roots of the inverse diagonals of sums of products matrices, that is,

$$\begin{aligned} \mathbf{e} &= [\mathrm{Diag}((\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}))]^{-1/2}\mathbf{1}_m, \\ \mathbf{d} &= [\mathrm{Diag}((\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}}))]^{-1/2}\mathbf{1}_p, \end{aligned} \tag{27}$$

the beta statistic takes the form

$$\mathrm{B} = \frac{(\mathbf{1}_m'\mathbf{R}_{YX}\mathbf{1}_p)^2}{\mathbf{1}_m'\mathbf{R}_{YY}\mathbf{1}_m \cdot \mathbf{1}_p'\mathbf{R}_{XX}\mathbf{1}_p} = \frac{\bar{r}_{YX}^2}{\bar{r}_{YY}\bar{r}_{XX}}. \tag{28}$$

Here,

$$\begin{pmatrix} \mathbf{R}_{YY} & \mathbf{R}_{YX} \\ \mathbf{R}_{XY} & \mathbf{R}_{XX} \end{pmatrix} \tag{29}$$

denotes the $(m + p) \times (m + p)$ correlation matrix of $\mathbf{Y}$ and $\mathbf{X}$, and $\bar{r}_{YX}$, $\bar{r}_{YY}$, $\bar{r}_{XX}$ denote the averages of the elements of the submatrices $\mathbf{R}_{YX}$, $\mathbf{R}_{YY}$, $\mathbf{R}_{XX}$. This beta statistic, a ratio of averages of correlation coefficients, follows a $\mathrm{B}(1/2, ((n - 2)/2))$ distribution if $\mathbf{Y}$ and $\mathbf{X}$ are independent, regardless of the dimensions m and p and for every $n \geq 3$.

This interesting result motivates the definition of the so-called summary correlation coefficient of $\mathbf{Y}$ and $\mathbf{X}$ [Läuter, Glimm and Kropf (1996)],

$$r = \frac{\bar{r}_{YX}}{\sqrt{\bar{r}_{YY}\bar{r}_{XX}}}\ . \tag{30}$$

Since $\mathbf{Y}$ and $\mathbf{X}$ are independent, it possesses the same distribution as the ordinary correlation coefficient of two independent, normally distributed variables and can therefore be tested with the usual t test for zero correlation

$$t = \sqrt{n - 2}\frac{r}{\sqrt{1 - r^2}}. \tag{31}$$

Calculation of the summary correlation coefficient can most effectively be done by forming the "univariate" scores $\mathbf{Ye}$ and $\mathbf{Xd}$ first and then determining their mutual bivariate correlation.

The summary correlation coefficient offers an alternative to the widespread multiple correlation coefficient. For certain parameter structures and for appropriate directions of the variables, it can be a more effective measure of association than the multiple correlation coefficient.

The possibilities for the random variable $\mathbf{Y}$ are manifold. The matrix $\mathbf{Y}$ does not have to fulfill any distributional conditions except independence of $\mathbf{X}$, and hence it can be the result of some arbitrarily chosen data transformation. Of course, the transformations have to be based solely on the $\mathbf{Y}$ information. Two such transformations are of special interest: first, a matrix of $\mathbf{Y}$ ranks may be used instead of the original $\mathbf{Y}$ data matrix and second, one may impute missing values in $\mathbf{Y}$ and still test for independence using the proposed procedure.

(6) *The invariance and the unbiasedness of the tests.* Theorem 1 facilitates both the construction of scale-invariant and of scale-dependent multivariate tests. For example, the eigenvalue problem $(\mathbf{H}+\mathbf{G})\mathbf{D} = \mathbf{D}\mathbf{\Lambda}$ could be used instead of (19) in application (2). This eigenvalue problem is no longer scale-invariant. Note that this remark does not contradict the invariance requirement (9) for the function F in Theorem 1. Furthermore, the tests constructed in this paper will not generally be unbiased.

3. The case of more general left-spherical distributions. In this section, the considerations of the former section are extended to nonnormal spherical distributions. As a generalization of (5), suppose now an $n \times p$ left-spherically distributed matrix $\mathbf{X} - \mathbf{M}$ with $\mathbf{M}$ being a constant matrix. Then $\mathbf{X}$ is centered around $\mathbf{M}$. We are interested in testing hypotheses of the form

$$\text{H: } \mathbf{E}'\mathbf{M} = \mathbf{0}, \tag{32}$$

where $\mathbf{E}$ is a fixed $n \times f$ matrix with $\mathbf{E}'\mathbf{E} = \mathbf{I}_f$, and $\mathbf{Q} = \mathbf{E}\mathbf{E}'$ is the corresponding projection matrix of rank f. The setting in Section 2 is a special case of this situation with $f = f_H + f_G$, $\mathbf{Q} = \mathbf{Q}_H + \mathbf{Q}_G$. Condition (32) contains f restrictions on the means for both the hypothesis and the residuals.

Since the columns of $\mathbf{E}$ are mutually orthogonal, there is an $n \times (n-f)$ orthogonal complement $\mathbf{E}_*$ such that $(\mathbf{E} \quad \mathbf{E}_*)$ becomes an $n \times n$ orthogonal matrix. Under the hypothesis, we thus have

$$\begin{pmatrix} \mathbf{E}' \\ \mathbf{E}_*' \end{pmatrix}(\mathbf{X} - \mathbf{M}) = \begin{pmatrix} \mathbf{E}'\mathbf{X} - \mathbf{E}'\mathbf{M} \\ \mathbf{E}_*'\mathbf{X} - \mathbf{E}_*'\mathbf{M} \end{pmatrix} = \begin{pmatrix} \mathbf{E}'\mathbf{X} \\ \mathbf{E}_*'\mathbf{X} - \mathbf{E}_*'\mathbf{M} \end{pmatrix}. \tag{33}$$

This matrix is left-spherically distributed. Furthermore, the conditional distribution of the submatrix $\mathbf{E}'\mathbf{X}$ for given $\mathbf{E}_*'\mathbf{X} - \mathbf{E}_*'\mathbf{M}$ or $\mathbf{E}_*'\mathbf{X}$ is left-spherical.

The following theorem is useful for such situations.

THEOREM 2. *Assume* $1 \le q \le f \le n$, *and let* $(\mathbf{E} \quad \mathbf{E}_*)$ *be an* $n \times n$ *orthogonal matrix consisting of the* $n \times f$ *matrix* $\mathbf{E}$ *and the* $n \times (n-f)$ *complement* $\mathbf{E}_*$. *Assume a test statistic* $F = F(\mathbf{Z}_{n\times q})$ *as a Borel function defined for all* $n \times q$

matrices $\mathbf{Z}_{n\times q}$ *and satisfying the invariance condition*

$$F(\mathbf{EUA}+\mathbf{B}) =_d F(\mathbf{EU}) \tag{34}$$

for every fixed $q \times q$ *positive definite symmetric matrix* $\mathbf{A}$ *and for every fixed* $n \times q$ *matrix* $\mathbf{B}$ *with* $\mathbf{E}'\mathbf{B} = \mathbf{0}$, *where* $\mathbf{U}$ *is an* $f \times q$ *left-spherically, uniformly distributed matrix* [*for the definition, see Fang and Zhang* (1990), *page* 93].

Now, suppose any dimension p *with* $p \geq q$. *Let* $\mathbf{X}$ *be an* $n \times p$ *random matrix such that* $\mathbf{W} = \mathbf{E}'\mathbf{X}$ *is conditionally left-spherically distributed for given* $\mathbf{W}_* = \mathbf{E}_*'\mathbf{X}$, *and let* $\mathbf{D}$ *be a* $p \times q$ *random matrix determined as a Borel function of* $\mathbf{X}'\mathbf{Q}_0\mathbf{X}$, *where* $\mathbf{Q}_0$ *is a projection matrix with* $\mathbf{Q}_0 \geq \mathbf{Q} = \mathbf{EE}'$. *Assume that* $\mathbf{E}'\mathbf{XD}$ *has rank* q *with probability* 1.

Then $\mathbf{E}'\mathbf{XD}$ *is conditionally left-spherically distributed for given* $\mathbf{E}_*'\mathbf{XD}$. *The distribution of* $F(\mathbf{XD})$ *is the same for each* p, *each suitable* $\mathbf{X}$ *distribution, each* $\mathbf{D}$ *function and each projection matrix* $\mathbf{Q}_0$.

PROOF. The inequality $\mathbf{Q}_0 \geq \mathbf{Q}$ for the idempotent symmetric matrices $\mathbf{Q}_0$ and $\mathbf{Q}$ denotes that $\mathbf{Q}_0 - \mathbf{Q}$ is positive semidefinite. This implies that $\mathbf{QQ}_0 = \mathbf{Q}$ and that $\mathbf{Q}_0 - \mathbf{Q}$ is also idempotent and symmetric. Hence

$$\mathbf{E}'(\mathbf{Q}_0 - \mathbf{Q}) = \mathbf{E}'\mathbf{EE}'(\mathbf{Q}_0 - \mathbf{Q}) = \mathbf{E}'\mathbf{Q}(\mathbf{Q}_0 - \mathbf{Q}) = \mathbf{0}. \tag{35}$$

Therefore $\mathbf{Q}_0 - \mathbf{Q}$ is in the column space of $\mathbf{E}_*$, that is, $\mathbf{Q}_0 - \mathbf{Q} = \mathbf{E}_*\mathbf{C}$ for some $\mathbf{C}$. As a consequence,

$$\begin{aligned} \mathbf{X}'\mathbf{Q}_0\mathbf{X} &= \mathbf{X}'\mathbf{QX} + \mathbf{X}'(\mathbf{Q}_0 - \mathbf{Q})\mathbf{X} \\ &= (\mathbf{E}'\mathbf{X})'(\mathbf{E}'\mathbf{X}) + (\mathbf{C}'\mathbf{E}_*'\mathbf{X})'(\mathbf{C}'\mathbf{E}_*'\mathbf{X}) \\ &= \mathbf{W}'\mathbf{W} + (\mathbf{C}'\mathbf{W}_*)'(\mathbf{C}'\mathbf{W}_*). \end{aligned} \tag{36}$$

Now, consider the conditional distributions of $\mathbf{W}$ and $\mathbf{Y} = \mathbf{WD}$ given $\mathbf{W}'\mathbf{W}$ and $\mathbf{W}_*$. By definition, $\mathbf{W}$ given $\mathbf{W}_*$ is conditionally left-spherically distributed. Fixing $\mathbf{W}'\mathbf{W}$ additionally preserves the left-sphericity of the conditional distribution. Hence, the random $f \times q$ matrix $\mathbf{Y} = \mathbf{WD}$ is also conditionally left-spherically distributed, because $\mathbf{D}$ is a constant for fixed $\mathbf{X}'\mathbf{Q}_0\mathbf{X}$ and therefore also for fixed values $\mathbf{W}'\mathbf{W}$ and $\mathbf{W}_*$.

Modifying the terms in the condition from $\mathbf{W}'\mathbf{W}$ to $\mathbf{Y}'\mathbf{Y} = \mathbf{D}'\mathbf{W}'\mathbf{WD}$ and from $\mathbf{W}_*$ to $\mathbf{Y}_* = \mathbf{W}_*\mathbf{D}$ still retains the property of left-sphericity of $\mathbf{Y} = \mathbf{WD}$, because the distribution of $\mathbf{Y}$ given $\mathbf{Y}'\mathbf{Y}$ and $\mathbf{Y}_*$ is a mixture of the more specialized conditional distribution of $\mathbf{Y}$ given $\mathbf{W}'\mathbf{W}$ and $\mathbf{W}_*$. Finally, by the same argument, the unconditional distribution of $\mathbf{Y} = \mathbf{WD}$ is also left-spherical. $\mathbf{Y}$ has rank q with probability 1 and therefore, there is a representation

$$\mathbf{Y} =_d \mathbf{U}(\mathbf{Y}'\mathbf{Y})^{1/2}, \tag{37}$$

where $\mathbf{U}$ has the $f \times q$ left-spherical uniform distribution and is independent of $\mathbf{Y}$ [Fang and Zhang (1990), page 93].

Concerning the test statistic F and its conditional distribution for given $\mathbf{Y}'\mathbf{Y}$ and $\mathbf{Y}_*$, we have

$$\begin{aligned} F(\mathbf{XD}) &= F((\mathbf{EE}' + \mathbf{E}_*\mathbf{E}_*')\mathbf{XD}) \\ &= F(\mathbf{EY} + \mathbf{E}_*\mathbf{Y}_*) =_d F(\mathbf{EU}(\mathbf{Y}'\mathbf{Y})^{1/2} + \mathbf{E}_*\mathbf{Y}_*). \end{aligned} \tag{38}$$

Applying the invariance condition (34) yields

$$F(\mathbf{XD}) =_d F(\mathbf{EU}), \tag{39}$$

because of $\mathbf{E}'\mathbf{E}_* = \mathbf{0}$. Obviously, the obtained distribution no longer depends on the particular values of $\mathbf{Y}'\mathbf{Y}$ and $\mathbf{Y}_*$, and hence, it is also the unconditional distribution of F. The distribution is the same for each p, each possible $\mathbf{X}$ distribution, each appropriate $\mathbf{D}$ function and each projection matrix $\mathbf{Q}_0$. □

REMARKS. (i) In the special case of $f = n$ in Theorem 2, the matrix $\mathbf{E}_*$ is missing, and $\mathbf{E}$ is an $n \times n$ orthogonal matrix. Then the conditional distributions of $\mathbf{E}'\mathbf{X}$ given $\mathbf{E}_*'\mathbf{X}$ and $\mathbf{E}'\mathbf{XD}$ given $\mathbf{E}_*'\mathbf{XD}$ should be replaced by the unconditional distributions of $\mathbf{E}'\mathbf{X}$ and $\mathbf{E}'\mathbf{XD}$, respectively.

(ii) As a consequence of Theorem 2, all applications of Theorem 1 in Section 2 can be extended to random variables $\mathbf{X}$, where $\mathbf{E}'\mathbf{X}$ has a conditional left-spherical distribution given the complement $\mathbf{E}_*'\mathbf{X}$. In particular, this is true if $\mathbf{X} - \mathbf{M}$ is left-spherical with $\mathbf{E}'\mathbf{M} = \mathbf{0}$.

(iii) The invariance condition (34) for the test statistic F is fulfilled in the special case

$$F(\mathbf{Z}_{n\times q}) = F_0(\mathbf{E}'\mathbf{Z}_{n\times q}), \tag{40}$$

where the function $F_0 = F_0(\mathbf{Y}_{f\times q})$ satisfies the equation

$$F_0(\mathbf{U}_{f\times q}\mathbf{A}) = F_0(\mathbf{U}_{f\times q}) \tag{41}$$

for every $f \times q$ matrix $\mathbf{U}_{f\times q}$ with $\mathbf{U}_{f\times q}'\mathbf{U}_{f\times q} = \mathbf{I}_q$ and for every $q \times q$ positive definite symmetric matrix $\mathbf{A}$. If $F_0(\mathbf{U}_{f\times q})$ has already been defined for all $\mathbf{U}_{f\times q}$ with $\mathbf{U}_{f\times q}'\mathbf{U}_{f\times q} = \mathbf{I}_q$, then the value of such a function F_0 for an arbitrary $\mathbf{Y}_{f\times q}$ of rank q is obtained by $F_0(\mathbf{Y}_{f\times q}) = F_0(\mathbf{Y}_{f\times q}(\mathbf{Y}_{f\times q}'\mathbf{Y}_{f\times q})^{-1/2})$. This principle will be used in application (4) of Theorem 2 [see (48) and (49)].

(iv) The data compression into scores according to Theorem 2 may also be performed in a multistage procedure. The first step consists of calculating $\mathbf{Z} = \mathbf{XD}$ from $\mathbf{X}$, where $\mathbf{D}$ is a function of $\mathbf{X}'\mathbf{Q}_0\mathbf{X}$ with $\mathbf{Q}_0 \geq \mathbf{Q} = \mathbf{EE}'$. In the next step, one may compute scores $\mathbf{Z}_1 = \mathbf{ZD}_1$ by means of a coefficient matrix $\mathbf{D}_1$ derived from $\mathbf{Z}'\mathbf{Q}_1\mathbf{Z}$ with $\mathbf{Q}_1 \geq \mathbf{Q} = \mathbf{EE}'$. In the same way, the data compression could be continued. In the end, it is imperative that a hypothesis $\mathbf{E}'\mathbf{M} = \mathbf{0}$ must only be tested if $\mathbf{EE}' \leq \mathbf{Q}_0$, $\mathbf{EE}' \leq \mathbf{Q}_1, \ldots$ are fulfilled. The test that finally will be performed does not have to be fixed in advance, that is, $\mathbf{E}$ can be chosen after calculation of the scores. However, the matrices $\mathbf{Q}_0, \mathbf{Q}_1, \ldots$ have to be chosen "big enough" for the final test. The conditional left-sphericity of $\mathbf{E}'\mathbf{X}$ given $\mathbf{E}_*'\mathbf{X}$ implies the conditional left-sphericity of $\mathbf{E}'\mathbf{Z}$ given $\mathbf{E}_*'\mathbf{Z}$; this implies the conditional left-sphericity of $\mathbf{E}'\mathbf{Z}_1$ given $\mathbf{E}_*'\mathbf{Z}_1$ and so on.

FURTHER APPLICATIONS (1) *One-sided tests*. Theorem 2 allows performing one-sided tests. For example, testing H: $\boldsymbol{\mu} = \mathbf{0}$ versus the alternative A: $\mu_i > 0$ for $i = 1, \ldots, p$ with an $n \times p$ left-spherically distributed matrix $\mathbf{X} - \mathbf{1}_n \boldsymbol{\mu}'$ can be done using the t statistic

$$t(\mathbf{z}_{n\times 1}) = \frac{\sqrt{n}\,\bar{z}}{s_z} \tag{42}$$

for the score vector $\mathbf{z} = \mathbf{X}\mathbf{d}$. The coefficient vector $\mathbf{d}$ is determined as a function of $\mathbf{X}'\mathbf{X}$ with positive components, for example, in the form of application (1) of Section 2. Statistic (42) meets requirement (34), because $t(\mathbf{z}_{n\times 1}a) = t(\mathbf{z}_{n\times 1})$ is true for each positive scalar a. Under the null hypothesis, (42) has the t distribution with $n - 1$ degrees of freedom.

(2) *A directed comparison of the mean vectors of several populations.* A directed comparison of the means of K independent, p-dimensional normally distributed populations or of corresponding left-spherically distributed observations, respectively, is possible on the basis of Theorem 2. Consider, for example, the hypothesis H: $\boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(K)}$ and the directed alternative A: $\mu_i^{(1)} < \mu_i^{(2)} < \cdots < \mu_i^{(K)}$ for $i = 1, \ldots, p$. A possible approach is based on the test of correlation between an $n \times 1$ matrix $\mathbf{Y}$ and an $n \times p$ matrix $\mathbf{X}$ (cf. application (5) of Section 2). Let

$$\mathbf{Y}' = (y^{(1)} \cdots y^{(1)}, y^{(2)} \cdots y^{(2)}, \ldots, y^{(K)} \cdots y^{(K)}), \qquad n = n^{(1)} + \cdots + n^{(K)} \tag{43}$$

with $y^{(k)} = (2k - K - 1)/n^{(k)}$ for $k = 1, \ldots, K$ and calculate the weight vector $\mathbf{d}$ for the $\mathbf{X}$ variables according to (27). Formulas (30) and (31) then yield an exact t test with $n - 2$ degrees of freedom. This test is similar to a univariate one proposed by Barlow, Bartholomew, Bremner and Brunk (1972).

(3) *Dunnett test*. The well-known Dunnett procedure for testing K treatments against a control can be transferred to the multivariate case by using Theorem 2. Of course, this procedure is then also valid for left-spherical observations. The one-sided Dunnett closure procedure [Dunnett (1955), Marcus, Peritz and Gabriel (1976), Dunnett and Tamhane (1991)] is based on the t statistics

$$t^{(k)}(\mathbf{z}_{n\times 1}) = \sqrt{\frac{n^{(0)}n^{(k)}}{n^{(0)} + n^{(k)}}}\,\frac{\bar{z}^{(k)} - \bar{z}^{(0)}}{s_z}, \qquad k = 1, \ldots, K, \tag{44}$$

corresponding to the single hypotheses $\mathrm{H}^{(1)}$ to $\mathrm{H}^{(K)}$. These statistics are arranged in descending order of their values and then tested consecutively. As soon as the first nonsignificant result is reached, the procedure stops. The Dunnett procedure requires that the statistic defined by

$$t(\mathbf{z}_{n\times 1}) = \max_{k:\,\mathrm{H}^{(k)}\text{ is valid}} (t^{(k)}(\mathbf{z}_{n\times 1})) \tag{45}$$

does not exceed its critical value with a probability of more than α for arbitrary parameters.

In the multivariate analogue to the Dunnett procedure proposed here, the vector $\mathbf{d}$ of weights for the p variables is calculated as a function of the ma-

trix $(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})$, with $\bar{\mathbf{X}}=\mathbf{1}_n\bar{\mathbf{x}}'$ being the matrix of total means of all $1+K$ populations. Following Theorem 2, the "multivariate" Dunnett procedure performed with the score values $\mathbf{z}=\mathbf{X}\mathbf{d}$ and the statistic $t(\mathbf{z})$ according to (45) exactly keeps the multiple level of significance α in spite of the fact that $\mathbf{d}$ is based on the data from all $1+K$ populations and regardless of whether the single hypotheses $\mathrm{H}^{(k)}$ are true or not. At this point, it is crucial that Theorem 2 allows the incorporation of neutral, noncentered parts of the data into the calculation of $\mathbf{d}$ ($\mathbf{Q}_0 > \mathbf{Q}$ is admitted). Here, we have $\mathbf{Q}_0=\mathbf{I}_n-(1/n)\mathbf{1}_n\mathbf{1}_n'$, while $\mathbf{Q}=\mathbf{E}\mathbf{E}'$ corresponds to the subset of valid hypotheses $\mathrm{H}^{(k)}$. The test statistic (45) fulfills the invariance condition (34) of Theorem 2, because $t(\mathbf{z}_{n\times 1}a+\mathbf{m})=t(\mathbf{z}_{n\times 1})$ for $a>0$, $\mathbf{E}'\mathbf{m}=\mathbf{0}$. This method is described in more detail by Kropf, Hothorn and Läuter (1997a).

(4) *Test statistics with weakened invariance, repeated measurement analysis.* The test statistic

$$F_q(\mathbf{Z}_{n\times q})=\frac{1}{n}\mathbf{1}_n'\mathbf{Z}_{n\times q}(\mathbf{Z}_{n\times q}'\mathbf{Z}_{n\times q})^{-1}\mathbf{Z}_{n\times q}'\mathbf{1}_n,\qquad q<n, \tag{46}$$

defined for all $n\times q$ matrices $\mathbf{Z}_{n\times q}$ of rank q, is available for testing H: $\boldsymbol{\mu}=\mathbf{0}$ in the case of an $n\times p$ left-spherical matrix $\mathbf{X}-\mathbf{1}_n\boldsymbol{\mu}'$. It can be applied to the score matrix $\mathbf{Z}=\mathbf{X}\mathbf{D}$, where the coefficient matrix $\mathbf{D}$ is a function of $\mathbf{X}'\mathbf{X}$. The statistic (46) fulfills the invariance condition $F_q(\mathbf{Z}_{n\times q}\mathbf{A})=F_q(\mathbf{Z}_{n\times q})$ for any $q\times q$ nonsingular matrix $\mathbf{A}$ and hence meets the assumption (34) of Theorem 2. It can also be written as Pillai's trace, that is, $F_q(\mathbf{Z}_{n\times q})=\mathrm{tr}(\mathbf{H}_Z(\mathbf{H}_Z+\mathbf{G}_Z)^{-1})$ with $q\times q$ matrices $\mathbf{H}_Z$ and $\mathbf{G}_Z$ [Seber (1984)]. Therefore $F_q(\mathbf{Z})$ follows a beta distribution if H is true:

$$F_q(\mathbf{Z})\sim \mathrm{B}\left(\frac{q}{2},\frac{n-q}{2}\right). \tag{47}$$

Furthermore, $F_q(\mathbf{Z})$ also has a representation $F_q(\mathbf{Z})=(1/n)\mathbf{1}_n'\mathbf{U}\mathbf{U}'\mathbf{1}_n$ based on the $n\times q$ left-spherical uniform distribution $\mathbf{U}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}$.

Now, consider the modified test statistic

$$F_{q,r}(\mathbf{Z}_{n\times q})=\frac{1}{n}\mathbf{1}_n'\mathbf{Z}_{n\times q}(\mathbf{Z}_{n\times q}'\mathbf{Z}_{n\times q})^{-1/2}\mathbf{R}(\mathbf{Z}_{n\times q}'\mathbf{Z}_{n\times q})^{-1/2}\mathbf{Z}_{n\times q}'\mathbf{1}_n \tag{48}$$

with a fixed $q\times q$ projection matrix $\mathbf{R}$ of rank r, where $r<q\le n$. This statistic no longer fulfills the general affine invariance condition $F_{q,r}(\mathbf{Z}_{n\times q}\mathbf{A})=F_{q,r}(\mathbf{Z}_{n\times q})$, but for $\mathbf{U}_{n\times q}'\mathbf{U}_{n\times q}=\mathbf{I}_q$ and a $q\times q$ positive definite symmetric matrix $\mathbf{A}$, we have

$$\begin{aligned}
&F_{q,r}(\mathbf{U}_{n\times q}\mathbf{A})\\
&\quad=\frac{1}{n}\mathbf{1}_n'\mathbf{U}_{n\times q}\mathbf{A}(\mathbf{A}\mathbf{U}_{n\times q}'\mathbf{U}_{n\times q}\mathbf{A})^{-1/2}\mathbf{R}(\mathbf{A}\mathbf{U}_{n\times q}'\mathbf{U}_{n\times q}\mathbf{A})^{-1/2}\mathbf{A}\mathbf{U}_{n\times q}'\mathbf{1}_n\\
&\quad=\frac{1}{n}\mathbf{1}_n'\mathbf{U}_{n\times q}\mathbf{A}(\mathbf{A}^2)^{-1/2}\mathbf{R}(\mathbf{A}^2)^{-1/2}\mathbf{A}\mathbf{U}_{n\times q}'\mathbf{1}_n=\frac{1}{n}\mathbf{1}_n'\mathbf{U}_{n\times q}\mathbf{R}\mathbf{U}_{n\times q}'\mathbf{1}_n\\
&\quad=F_{q,r}(\mathbf{U}_{n\times q}).
\end{aligned} \tag{49}$$

Hence, the weaker invariance condition (34) of Theorem 2 is still kept. Consequently, the statistic (48) can be used in Theorem 2, even in the case of $q = n$. This is a bit surprising in comparison to classical multivariate tests, since the latter are available for $f_G \geq q$, that is, for $n \geq f > q$, only.

The statistic $F_{q,r}(\mathbf{Z})$ has the same distribution for each $n \times q$ left-spherically distributed matrix $\mathbf{Z}$. Especially, this distribution is reached for an $n \times q$ uniformly distributed matrix $\mathbf{U}$. If the projection matrix $\mathbf{R}$ is represented via a $q \times r$ orthonormal matrix $\mathbf{D}_r$, that is,

$$\mathbf{R} = \mathbf{D}_r\mathbf{D}_r' \quad \text{with } \mathbf{D}_r'\mathbf{D}_r = \mathbf{I}_r, \tag{50}$$

then we have

$$\begin{aligned} F_{q,r}(\mathbf{U}) &= \frac{1}{n}\mathbf{1}_n'\mathbf{U}\mathbf{R}\mathbf{U}'\mathbf{1}_n \\ &= \frac{1}{n}\mathbf{1}_n'\mathbf{U}\mathbf{D}_r\mathbf{D}_r'\mathbf{U}'\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n'(\mathbf{U}\mathbf{D}_r)(\mathbf{D}_r'\mathbf{D}_r)^{-1}(\mathbf{U}\mathbf{D}_r)'\mathbf{1}_n \\ &= \frac{1}{n}\mathbf{1}_n'(\mathbf{U}\mathbf{D}_r)(\mathbf{D}_r'\mathbf{U}'\mathbf{U}\mathbf{D}_r)^{-1}(\mathbf{U}\mathbf{D}_r)'\mathbf{1}_n = F_r(\mathbf{U}\mathbf{D}_r). \end{aligned} \tag{51}$$

Application of (47) then yields

$$F_{q,r}(\mathbf{Z}) \sim \mathrm{B}\left(\frac{r}{2}, \frac{n-r}{2}\right), \tag{52}$$

because $\mathbf{U}\mathbf{D}_r$ is an $n \times r$ left-spherically distributed matrix which has always rank r. Thus, the null distribution of the statistic (48) depends on $\mathbf{R}$ only via the rank r.

Using $\mathbf{U} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}$ and $\mathbf{u}' = (u_1 \quad u_2 \quad \cdots \quad u_q) = \mathbf{1}_n'\mathbf{U}$ yields the computational formula $F_{q,r}(\mathbf{Z}) = (1/n)\mathbf{u}'\mathbf{R}\mathbf{u}$. Such a test statistic is suitable for repeated measurement analysis with q replicated measurements on n subjects.

Some special cases are:

(a) $$\mathbf{R} = \frac{1}{q}\mathbf{1}_q\mathbf{1}_q', \qquad F_{q,1}(\mathbf{Z}) = \frac{q}{n}\bar{u}^2 \sim \mathrm{B}\left(\frac{1}{2}, \frac{n-1}{2}\right)$$

under H, where $\bar{u} = (1/q)\sum_{i=1}^{q} u_i$.

This test is sensitive against departures of the mean level of the q repetitions from zero.

(b) $$\mathbf{R} = \mathbf{I}_q - \frac{1}{q}\mathbf{1}_q\mathbf{1}_q', \qquad F_{q,q-1}(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{q}(u_i - \bar{u})^2 \sim \mathrm{B}\left(\frac{q-1}{2}, \frac{n-q+1}{2}\right)$$

under H. This test is sensitive against differences between the q repetitions.

4. Concluding remarks. This work provides tests for high-dimensional normally distributed and, more generally, for left-spherically distributed data which are based on calculation of low-dimensional linear scores. The procedure

works without any additional bias due to model fit or selection. In each case, the low-dimensional test statistics have the same null distribution, regardless of the original dimension and the special choice of weights. Traditional theoretical optimality criteria (in the sense of the least squares method or the maximum likelihood method) are abandoned for the sake of a gain in stability and efficiency in many applications. Thus, this class of tests provides exact alternatives to the approximate tests for "multiple endpoints" which were proposed for medical studies by O'Brien (1984), Tang, Geller and Pocock (1993) and others. We believe that the methods treated here will open up new opportunities to theoretical and practical statistics.

Acknowledgments. The authors are grateful to the Editors and referees for advice and helpful comments concerning the paper.

REFERENCES

ANDERSON, T. W. (1993). Nonnormal multivariate distributions: inference based on elliptically contoured distributions. In *Multivariate Analysis: Future Directions* (C. R. Rao, ed.) 1–24. North-Holland, Amsterdam.

ANDERSON, T. W., FANG, K.-T. and HSU, H. (1990). Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions. In *Statistical Inference in Elliptically Contoured and Related Distributions* (K.-T. Fang and T. W. Anderson, eds.) 217–223. Allerton Press, New York.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, New York.

BÜNING, H. (1991). *Robust and Adaptive Tests*. de Gruyter, Berlin.

DUNNETT, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* **50** 1096–1121.

DUNNETT, C. W. and TAMHANE, A. C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layout. *Statistics in Medicine* **10** 939–947.

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

GUPTA, A. K. and VARGA, T. (1993). *Elliptically Contoured Models in Statistics*. Kluwer, Dordrecht.

HSU, H. (1990a). Generalized T^2-test for multivariate elliptically contoured distributions. In *Statistical Inference in Elliptically Contoured and Related Distributions* (K.-T. Fang and T. W. Anderson, eds.) 243–256. Allerton Press, New York.

HSU, H. (1990b). Invariant tests for multivariate elliptically contoured distributions. In *Statistical Inference in Elliptically Contoured and Related Distributions* (K.-T. Fang and T. W. Anderson, eds.) 257–274. Allerton Press, New York.

KARIYA, T. and SINHA, B. K. (1989). *Robustness of Statistical Tests*. Academic Press, San Diego.

KROPF, S., HOTHORN, L. A. and LÄUTER, J. (1997a). Multivariate many-to-one procedures with applications to pre-clinical trials. *Drug Information J.* **31** 433–447.

KROPF, S., LÄUTER, J. and GLIMM, E. (1997b). Stabilized multivariate procedures in tests and prediction. Unpublished manuscript.

LÄUTER, J. (1992). *Stabile multivariate Verfahren: Diskriminanzanalyse, Regressionsanalyse, Faktoranalyse*. Akademie, 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. (1996). New multivariate tests for data with an inherent structure. *Biometrical J.* **38** 5–23.

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660.

O'BRIEN, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40** 1079–1087.

SEBER, G. A. F. (1984). *Multivariate Observations*. Wiley, New York.

TANG, D., GELLER, N. L. and POCOCK, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49** 23–30.

INSTITUTE OF BIOMETRICS AND MEDICAL INFORMATICS
OTTO VON GUERICKE UNIVERSITY
LEIPZIGER STRASSE 44
39120 MAGDEBURG
GERMANY
E-MAIL: juergen.laeuter@medizin.uni-magdeburg.de