Comparison of Independent Samples of High-Dimensional Data by Pairwise Distance Measures

Siegfried Kropf[*,1], **Anke Lux**[1], **Markus Eszlinger**[2], **Holger Heuer**[3] and **Kornelia Smalla**[3]

[1] Institute of Biometry and Medical Informatics, Otto von Guericke University, Magdeburg, Germany
[2] Medical Department III, University of Leipzig, Germany
[3] Institute for Plant Virology, Microbiology and Biosafety, Federal Biological Research Centre for Agriculture and Forestry Braunschweig, Germany

Received 8 July 2005, revised 16 December 2005, accepted 24 April 2006

Summary

Pairwise distance or association measures of sample elements are often used as a basis for hierarchical cluster analyses. They can also be used in tests for the comparison of pre-defined subgroups of the total sample. Usually this is done with permutation tests.

In this paper, we compare such a procedure with alternative tests for high-dimensional data based on spherically distributed scores in simulation experiments and with real data. The tests based on the pairwise distance or similarity measures perform quite well in this comparison. As the number of possible permutations is small in very small samples, this might restrict the use of the test. Therefore, we propose an exact parametric small sample version of the test using randomly rotated samples.

Key words: Exact parametric test; Monte Carlo techniques; Multivariate tests; Pairwise distance measures; Permutation test.

1 Introduction

In many branches of medicine and biology one wants to compare independent samples of high-dimensional data where the dimension of the observation vectors is much larger than the available sample sizes. Examples are gene expression analyses with microarrays (Eszlinger et al., 2004, 2005), genetic fingerprinting (Aittokallio et al., 2000) or neurophysiological techniques (Hemmelmann et al., 2004, 2005). Different questions may be of interest. Often one wants to carry out multiple comparisons for the single variables then. In this paper, we are interested in a global assessment of all variables. In an application those might be all available variables of the subjects or a predefined subset. We consider the comparison of independent samples. Other situations can be treated in a similar manner but that is not the focus here.

Thus, the basic statistical model for the K samples is in the parametric case

$$\boldsymbol{x}_{kj} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \quad k = 1, \ldots, K; \quad j = 1, \ldots, n_k\,. \tag{1}$$

In the nonparametric situation, we assume continuous multivariate distributions

$$\boldsymbol{x}_{kj} \sim F_k(\boldsymbol{x}) \quad k = 1, \ldots, K; \quad j = 1, \ldots, n_k\,. \tag{2}$$

The dimension p of the observation vectors is typically much larger than the combined sample size $n = n_1 + \cdots + n_K$, $p \gg n$, though that is no assumption for the procedures. The null hypothesis to be

* Corresponding author: e-mail: Siegfried.Kropf@medizin.uni-magdeburg.de,
Phone: +49 391 671 3524, Fax: +49 391 671 3536

tested is the identity of the K distributions, i.e.,

$$\mathrm{H}_{0\,\mathrm{par}} : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_K = \boldsymbol{\mu} \tag{3}$$

in the parametric case, and

$$\mathrm{H}_{0\,\mathrm{nonpar}} : F_1(\boldsymbol{x}) = \ldots = F_K(\boldsymbol{x}) \quad \forall \boldsymbol{x} \tag{4}$$

in the nonparametric case.

We are going to consider and compare two basic techniques to cope with the high-dimensional data. The first one transforms the high-dimensional data vectors into low-dimensional score vectors and then carries out classical multivariate tests with the scores. Following the rules given in Läuter et al. (1996, 1998), the scores have left-spherical distributions under the null hypothesis and the final tests are exact parametric tests in the parametric model (1)/(3).

The main focus here, however, is on a different test strategy. In this version, pairwise similarity or distance measures are computed for all sample elements. Then the difference of those measures for pairs from the same group and for pairs from different groups is utilized in a permutation test, which primarily is distribution-free. However, this test is restricted in very small samples. Because of a small number of different permutations, the minimal possible p-value might be too large in applications, particularly if post-hoc tests are considered in the case $K > 2$. Therefore, a parametric small sample version is proposed based on random orthogonal transformations of the data (rotation test, cf. Läuter et al., 2005; Langsrund, 2005).

The two basic procedures are described in more detail in the next section. In Section 3 the comparison of both is considered for real and simulated data. The parametric small-sample test based on the pairwise measures is the contents of Section 4 including the application to the examples and simulation experiments for the robustness of the procedure in case of violations of the normality assumption. The paper ends with a short discussion of the investigated methods.

2 Description of the Test Strategies

As we are mainly interested in the situation where the dimension of the variables is much larger than the sample size, classical multivariate tests as Wilks' Λ are no longer applicable. One could use some of the test statistics published in the 1980s under the keyword "multiple endpoints tests" as, e.g., O'Briens (1984) OLS (ordinary least squares) test. In these tests a summary measure is constructed from the multivariate observations and then treated in standard tests. Most of these tests are treated only with their asymptotical distribution or in permutation tests.

2.1 Parametric principal component test

Instead, we are going to use principal component (PC) based tests, particularly the PC_q test which belongs to a class of exact parametric tests proposed by Läuter (1996) and Läuter et al. (1996, 1998). The PC_q test uses a principal component transformation of the variables and allows reducing the high-dimensional data to more than a single score value. In many applications we have found the situation that the first principal component gives some overall assessment of the subjects which may be not the difference looked for in these data. Also in the first example of Section 3.1 below, the first principal component explaining 38.8% of the variance fails to display significant results in the comparison of all groups and in four of the six pairwise comparisons (column "PC_1" in Table 2). The second component explains only 23.9% of the variance but finds significant differences in the global as well in four of the six pairwise comparisons. The same is true for the combination of both first two components (column "PC2" in Table 2).

In more detail, we first collect the sample vectors into the data matrix $\boldsymbol{X} = (\,\boldsymbol{x}_{11} \quad \cdots \quad \boldsymbol{x}_{Kn_K}\,)'$, calculate the mean vector over all samples $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k}\boldsymbol{x}_{kj}$ and the mean reduced sample matrix $\boldsymbol{X}^* = (\,\boldsymbol{x}_{11} - \bar{\boldsymbol{x}} \quad \cdots \quad \boldsymbol{x}_{Kn_K} - \bar{\boldsymbol{x}}\,)'$.

Then a weight matrix $\boldsymbol{D}$ of size $p \times q$ is determined as the matrix of the first q eigenvectors of the eigenvalue problem

$$\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}\boldsymbol{D} = \boldsymbol{D}\boldsymbol{\Lambda}\,,$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of the corresponding q largest eigenvalues. For convenience we assume here that the number q is fixed in advance. It can also be derived from the data (cf. Kropf, 2000). With a large number p of variables the matrix $\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}$ may become huge. Then it is easier to determine the eigenvalues and eigenvectors of the dual eigenvalue problem

$$\boldsymbol{X}^{*}\boldsymbol{X}^{*\prime}\tilde{\boldsymbol{D}} = \tilde{\boldsymbol{D}}\tilde{\boldsymbol{\Lambda}}$$

and to utilize the relationship

$$\boldsymbol{\Lambda} = \tilde{\boldsymbol{\Lambda}}\,, \qquad \boldsymbol{D} = \boldsymbol{X}^{*\prime}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\Lambda}}^{-1/2}\,,$$

which is valid as long as q is small enough such that the eigenvalues contained in $\boldsymbol{\Lambda}$ are all positive.

With the weight matrix $\boldsymbol{D}$, the high-dimensional data vectors $\boldsymbol{x}_{kj}$ are transformed into low-dimensional score vectors $\boldsymbol{z}_{kj}$ $(k = 1, \ldots, K;\ j = 1, \ldots, n_k)$ by $\boldsymbol{z}_{kj} = \boldsymbol{D}'\boldsymbol{x}_{kj}$, and these score vectors are then compared between the K groups in a classical multivariate test as Wilks' Λ test, which is used throughout here. It is shown in the papers by Läuter (1996) and Läuter et al. (1996, 1998) that the final test with the scores exactly maintains the type I error because the scores have a left-spherical null distribution which is sufficient for the Wilks' test to be exact. Scale invariant versions can be found in the same papers. Here we use scale dependent versions in order to be comparable with the tests based on pairwise measures. Subsequently, these PC test versions are denoted as PC_1, PC_2, ..., where the subscript indicates the score dimension q.

2.2 Test based on pairwise similarity or distance measures

The main interest here is a different test strategy based on pairwise distance or similarity measures between the sample vectors, which is described in more detail in Kropf et al. (2004a). We use the same data as in that paper here, to motivate the test.

The aim of the study was to check if different plant cultures influence the bacteria population in the soil. For this, soil samples have been taken from unplanted soil ($n_1 = 4$, denoted as "B"), from the rhizosphere of strawberries ($n_2 = 3$, "S"), of potatoes ($n_3 = 4$, "P") and of oilseed rape ($n_4 = 4$, "R"). Typical parts of the bacterial DNA (deoxyribonucleic acid, the genetic code) in these soil samples were isolated and amplified with polymerase chain reaction (PCR) techniques and all samples have been administered to an electrophoresis gel, where each sample produced a typical electrophoresis lane (the "fingerprint") which have been scanned and transformed into high-dimensional vectors of greyscale values. The image processing was supported by the software GelCompar which also delivers the possibility to calculate several similarity or distance measures for each pair of lanes. We chose Pearson's correlation coefficient as measure and got the correlation matrix $\boldsymbol{R}$ as given in Table 1. The measures for those pairs of lanes from the same group are in the block diagonal and those for pairs from different groups are outside. This matrix is usually used by biologists to construct dendrograms for the similarity structure of the investigated samples in cluster analyses.

We used this matrix to construct a permutation test which is valid in both models (1)/(3) and (2)/(4). It is based on the simple test statistic

$$d = \bar{r}_{\text{within}} - \bar{r}_{\text{between}}\,, \tag{5}$$

where $\bar{r}_{\text{within}}$ denotes the arithmetic mean of all coefficients from the pairs from the same group (block diagonal elements without the ones for the identical pairs in the diagonal) and $\bar{r}_{\text{between}}$ denotes the average for those pairs from different groups. With the above data, the average of the 42 within-group pairs is 0.961, that of the 168 between-group pairs is 0.911, such that $d = 0.050$.

Table 1 Pearson correlation coefficients forming the correlation matrix $\boldsymbol{R}$ to quantify the similarities of the bacterial population in the soil samples of the "fingerprint" example. The denomination of the samples is described in the text. The enlarged distances between some lines and colums separate the four groups to be compared and characterize the block structure of the matrix $\boldsymbol{R}$.

Sample	B_1	B_2	B_3	B_4	S_1	S_2	S_3	P_1	P_2	P_3	P_4	R_1	R_2	R_3	R_4
B_1	1	.965	.982	.977	.904	.948	.845	.954	.957	.945	.915	.898	.949	.855	.866
B_2	.965	1	.964	.948	.890	.933	.833	.932	.942	.921	.903	.884	.927	.840	.854
B_3	.982	.964	1	.982	.916	.960	.862	.958	.962	.952	.935	.902	.946	.873	.876
B_4	.977	.948	.982	1	.907	.953	.849	.957	.958	.954	.935	.902	.943	.864	.864
S_1	.904	.890	.916	.907	1	.954	.907	.898	.904	.883	.891	.858	.900	.825	.832
S_2	.948	.933	.960	.953	.954	1	.920	.959	.962	.946	.945	.922	.946	.895	.893
S_3	.845	.833	.862	.849	.907	.920	1	.861	.870	.847	.874	.840	.858	.827	.832
P_1	.954	.932	.958	.957	.898	.959	.861	1	.991	.982	.967	.959	.972	.923	.918
P_2	.957	.942	.962	.958	.904	.962	.870	.991	1	.983	.967	.960	.976	.929	.926
P_3	.945	.921	.952	.954	.883	.946	.847	.982	.983	1	.962	.955	.977	.939	.929
P_4	.915	.903	.935	.935	.891	.945	.874	.967	.967	.962	1	.948	.950	.947	.930
R_1	.898	.884	.902	.902	.858	.922	.840	.959	.960	.955	.948	1	.962	.955	.953
R_2	.949	.927	.946	.943	.900	.946	.858	.972	.976	.977	.950	.962	1	.932	.945
R_3	.855	.840	.873	.864	.825	.895	.827	.923	.929	.939	.947	.955	.932	1	.974
R_4	.866	.854	.876	.864	.832	.893	.832	.918	.926	.929	.930	.953	.945	.974	1

If we use a similarity measure as the correlation coefficient here and if there are differences among the groups, then the within-group pairs should give larger measures than the between-group pairs in the average. This difference, however, vanishes if sample elements are exchanged between the groups, thus "polluting" the groups. In the matrix $\boldsymbol{R}$, an exchange of two sample elements corresponds to simultaneous exchange of two rows and the corresponding columns. Thus we apply the usual permutation scheme for independent groups by considering all $\dfrac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_K!}$ different allocations of the n sample elements in K groups of sizes n_1, n_2, …, n_K. For each allocation the test statistic is recalculated and the proportion of those allocations with a value larger than or equal to the original allocation is the p-value of the permutation test. When we use a distance measure instead, then the within-group pairs should have smaller values than the between-group pairs with the original grouping and we use the "inverted" difference $d = \bar{r}_{\text{between}} - \bar{r}_{\text{within}}$ instead. The permutation procedure is the same. For large n, it will be difficult to enumerate all possible allocations. One can consider a sufficient number of random permutations because the complete enumeration of all groupings is too time-consuming. If N_{MC} denotes the whole number of random repetitions and m the number of repetitions with a value of the test statistic larger than or equal to the original value then the p-value is calculated as $(m + 1)/(N_{\text{MC}} + 1)$ thus incorporating the original statistic. In the subsequent examples in Sections 3 and 4 we always used this Monte Carlo version for convenience even in the case of small samples.

This test based on pairwise measures can be interpreted as a special case of the so-called Mantel test (Mantel, 1967), which tests the independence of two distance matrices. A similar test is applied – also on electrophoresis data – by Aittokallio et al. (2000). In Kropf et al. (2004a), some extensions are given. One of them considers block designs, which enable, e.g., the combined analysis of electrophoresis data from several gels taking into account a possible confounding effect of the different gels. Approximate permutation tests for more complex designs and applications in ecological studies can be found in Anderson (2001).

2.3 Post-hoc tests for both versions

When stating differences in the comparison of $K > 2$ groups, then the question of pairwise comparisons of groups arises.

In the case of the permutation test based on the statistic d, the pairwise test can be carried out by exchanging only sample elements from the two groups of the pair. This is equivalent to omitting the other groups in the test. An adjustment for multiple comparisons could be done with the Bonferroni method. In this paper, however, we use unadjusted p-values because our main focus is the comparison of the two test strategies, not the multiple comparisons.

For the class of test based on spherically distributed scores comprising the PC_q test, it has been shown (Kropf et al., 1997; Läuter et al., 1998) that one can use the scores derived from all K groups to compare pairs of groups only. Thus the derivation of the scores needs to be done once only utilizing the information of the full data set. In order to be comparable to the above permutation test with d, the tests for pairs of groups have then also been carried out by deleting the scores of the other groups and omitting an α-adjustment.

3 Comparison of Both Strategies in Real Data and in a Simulation Study

3.1 Electrophoretic fingerprints

When we first consider the above example, then the corresponding p-values for the global comparison with all four groups and the unadjusted p-values for the pairwise post-hoc comparisons are given in Table 2 both with the permutation test based on the statistic d and with the PC_q test ($q = 1, \ldots, 5$). Additionally to the permutation test based on Pearson's r ($d(r)_{\text{perm}}$), the test version with the squared Euclidean distance as distance measure ($d(E^2)_{\text{perm}}$) is included. The principal component tests with up to five components included are denoted as PC_1 to PC_5. They use the original 400 greyscale values per lane used as resolution in this example. The last column ($d(E^2)_{\text{rot}}$) in this table and in the following ones is described and referred to in Section 4.

The global comparison with all four groups gives highly significant results for all test versions except for the PC test with one component. In most of the pairwise comparisons of groups, the permutation test yields a p-value of 0.029 which is the minimal possible p-value for these sample sizes. The results with the squared Euclidean distance are similar. The PC test versions have smaller p-values than the permutation tests in some of the comparisons but not in all.

Table 2 Results (p-values) of the comparisons in the fingerprint example with DNA samples from unplanted soil (B) and from rhizosphere from strawberry (S), potato (P) and oilseed rape (R). 9,999 random permutations or rotations (see Section 4) have been used for the tests based on the statistic d.

test plant culture	$d(r)_{\text{perm}}$	$d(E^2)_{\text{perm}}$	PC_1	PC_2	PC_3	PC_4	PC_5	$d(E^2)_{\text{rot}}$
all cultures	<.001	<.001	.058	.001	<.001	<.001	<.001	<.001
B–S	.029	.029	.076	.182	.042	.156	.032	.024
B–P	.029	.029	.015	.004	.002	.002	.018	.001
B–R	.029	.029	.007	.020	.005	.006	.013	.004
S–P	.029	.029	.360	.026	.060	.004	.039	.017
S–R	.029	.057	.969	.098	.076	.036	.152	.043
P–R	.057	.057	.159	.025	.024	.016	.083	.051

Table 3 Results (p-values) of the comparisons in the fingerprint example with DNA samples from soil from four different regions (numbered 1 to 4). 9,999 random permutations or rotations (see Section 4) have been used for the tests based on the statistic d.

Test Region	$d(r)_{\text{perm}}$	$d(\text{E}^2)_{\text{perm}}$	PC_1	PC_2	PC_3	PC_4	PC_5	$d(\text{E}^2)_{\text{rot}}$
All regions	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
1–2	.029	.029	<.001	<.001	<.001	<.001	.004	<.001
1–3	.029	.029	<.001	<.001	<.001	<.001	.010	<.001
1–4	.029	.029	<.001	<.001	<.001	.002	.018	<.001
2–3	.029	.029	.039	.001	.008	.005	.038	<.001
2–4	.029	.029	.561	.446	.019	.010	.060	<.001
3–4	.029	.029	.004	.007	.023	.087	.230	.001

The situation is a bit different in a second example from a similar experiment with bacterial community fingerprints from soils of four different regions (Table 3). Four soil samples have been taken in each region. Here, obviously, the differences in the bacterial populations are larger, which is recognized in the PC tests very well. Also the permutation test (with both pairwise measures) ends up with the minimal possible p-value 0.029 for all pairwise comparisons. Nevertheless, with such a minimal p-value a test could never yield a significant result using the Bonferroni adjustment for multiple comparisons at the common levels for the familywise error rate.

3.2 Gene expression data from microarrays

Stating the good results of the permutation test based on the statistic d in the fingerprinting data, we re-analysed gene expression data of tissue samples from patients with thyroid nodules (Eszlinger et al., 2004, 2005, Kropf et al., 2004b). There were 30 patients, 15 of them with so-called "hot" nodules and 15 with "cold" nodules. The tissue samples have been analyzed with Affymetrix®GeneChips measuring the expression values of 12,625 genes in parallel. As all multivariate test versions prove highly significant differences between hot and cold nodules with the full set of variables (logarithmic expression levels) and the full samples of each 15, we consider 500 randomly generated bootstrap samples of size 3 to 7 for each group. Table 4 gives the results of these tests as rejection rates over the whole set of all 500 bootstrap samples using a 5% error level for all tests.

Table 4 Rejection rates in 500 bootstrap samples of different sizes with tests at the α level 0.05 for the comparison of gene expression data of hot and cold thyroid nodules (Eszlinger et al., 2004, 2005). 999 random permutations or rotations (see Section 4) have been used in the tests based on the statistic d for each bootstrap sample.

Test $n_1 = n_2$	$d(r)_{\text{perm}}$	$d(\text{E}^2)_{\text{perm}}$	PC_1	PC_2	PC_3	PC_4	PC_5	$d(\text{E}^2)_{\text{rot}}$
3	0	0	.27	.25	.22	.19	–	.30
4	.47	.37	.38	.31	.29	.29	.26	.43
5	.81	.60	.49	.55	.49	.44	.46	.60
6	.95	.80	.59	.74	.66	.54	.54	.78
7	.99	.92	.66	.91	.85	.79	.73	.92

The optimal value of included principal components in the PC_q test changes from one to two if the sample sizes exceed 4. The permutation test based on the Pearson correlation coefficient, however, has a larger rate of rejections for all bootstrap sample sizes. The second version based on the squared Euclidean distance is slightly inferior in this example but still superior to the PC test versions.

3.3 Simulation results for samples with multivariate normal observation vectors

In the above data, we do not know the true distributions. Non-normality can influence the results (cf. Section 4.3). We do expect differences among the considered populations but we do not know if they really exist. Therefore, three simulation series for random normal observation vectors are included here.

In all three cases, we consider two independent samples of size 10 with multivariate normal observation vectors of dimension 100. These vectors have been constructed on the basis of q^*-dimensional vectors of independent latent variables ($q^* = 1$ in the first series, $q^* = 5$ in the second and $q^* = 10$ in the third one):

$$z_{k,j} \sim N_{q^*}(\tilde{\boldsymbol{\mu}}_k, \boldsymbol{I}_{q^*}) \,,$$

$k = 1, 2$; $j = 1, \ldots, 10$, where $\tilde{\boldsymbol{\mu}}_1 = (0 \quad \cdots \quad 0)'$, $\tilde{\boldsymbol{\mu}}_2 = (\tilde{\mu}_{2,1} \quad \cdots \quad \tilde{\mu}_{2,q^*})'$ and $\boldsymbol{I}_{q^*}$ is the q^*-dimensional identity matrix. The values $\tilde{\mu}_{2,i}$ have been re-determined for each new sample pair as uniformly distributed random numbers in the interval $[-1.6, 1.6]$ (these bounds being chosen to get fairly good power in the final tests).

With these latent variables, the final observation vectors $\boldsymbol{x}_{k,j}$ have been constructed by

$$\boldsymbol{x}_{k,j} = 0.5(\boldsymbol{z}_{k,j} \otimes \mathbf{1}_{p^*}) + 0.5\, \boldsymbol{\mu}_{k,j} \quad (k = 1, 2;\ j = 1, \ldots, 10) \,,$$

where $\mathbf{1}_{p^*}$ is a vector of p^* ones, p^* chosen such that $p^* \cdot q^* = 100$ and $\boldsymbol{u}_{k,j}$ is a 100-dimensional vector of independent standard normal variables (noise). This way, the p^* variables belonging to the same latent variables have the pairwise correlation coefficient 0.5, variables belonging to different latent variables are uncorrelated.

The results can be found in Table 5. The differences between the three configurations are huge. In the case of only one latent factor all 100 variables have equal expectation. This constant shift in all variables has no influence on the correlation coefficient as pairwise similarity measure. Therefore, the power of the test based on d drops down to the type I error. The version with the squared Euclidean distance, however, works quite well here. The best result for the PC tests is obtained with only one principal component as one would expect with one latent factor. The performance is then the same as with the second version of the permutation test. With increasing number of latent factors, the optimum within the PC test versions is shifted towards a larger number of included principal components (but

Table 5 Rejection rates in simulation experiments with multivariate normal observation vectors as described in the text with 10,000 repetitions for each of the three configurations. All test versions are carried out at the level $\alpha = 0.05$. The number of random permutations and rotations (see Section 4) used in the tests based on the statistic d is 999.

Test q^*, p^*	$D(r)_{\text{perm}}$	$d(\text{E}^2)_{\text{perm}}$	PC_1	PC_2	PC_3	PC_4	PC_5	$d(\text{E}^2)_{\text{rot}}$
1, 100	.05	.42	.42	.34	.30	.26	.23	.42
5, 20	.71	.82	.75	.79	.78	.75	.71	.82
10, 10	.89	.94	.88	.92	.92	.90	.88	.94

much smaller than the real number of latent factors). The power of the permutation test based on Pearson's r increases in relation to the PC tests, but the version based on the squared Euclidean distance remains more powerful and becomes optimal among the considered test versions.

4 Exact Parametric Small Sample Version for the Test Based on Pairwise Measures

4.1 Description of the rotation test

As one could see in the examples in the last section – particularly in the fingerprint data – the power of the permutation test is strongly limited for very small samples, even if the difference between the populations to be compared is large. Therefore we are looking for a parametric test version in the model (1)/(3). Because the distribution of the test statistic d for the various pairwise measures might be rather complicated, we do not try to find a closed solution. Instead, we use a Monte Carlo test similar as in Läuter et al. (2005), called rotation test there. The basic idea is the following:

We trace back the test statistic to a modified version $\tilde{\boldsymbol{X}}$ of the data matrix $\boldsymbol{X}$ which under the null hypothesis has a left-spherical matrix distribution. The distribution of such a matrix is invariant to orthogonal rotations, i.e., $\tilde{\boldsymbol{X}} \overset{d}{\sim} \boldsymbol{B}\tilde{\boldsymbol{X}}$ for each fixed orthogonal matrix $\boldsymbol{B}$ of suitable size. Therefore, we can apply the rotation principle in a similar manner as the permutation principle. The test statistic is recalculated for a large number of rotated matrices, where the matrices used to perform the orthogonal rotations are generated as independent matrices with a spherical standard distribution. In contrast to the permutations, the number of different orthogonal rotations is infinite, such that the "technical" restriction of the permutation tests is no longer present.

In more detail: As in Section 2.1 we start with the data matrix $\boldsymbol{X} = (\boldsymbol{x}_{11} \quad \cdots \quad \boldsymbol{x}_{Kn_K})'$, which under the null hypothesis of no group differences has the multivariate normal distribution

$$\boldsymbol{X} \sim N_{n\times p}(\mathbf{M}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}) \quad \text{with} \quad \mathbf{M} = (\boldsymbol{\mu} \quad \ldots \quad \boldsymbol{\mu})'.$$

In order to eliminate the unknown expectation, we turn to the mean-reduced matrix

$$\boldsymbol{X}^* = (\boldsymbol{x}_{11} - \bar{\boldsymbol{x}} \quad \cdots \quad \boldsymbol{x}_{Kn_K} - \bar{\boldsymbol{x}})' \sim N_{n\times p}(\mathbf{0}, \boldsymbol{\Gamma}_n \otimes \boldsymbol{\Sigma}),$$

where the matrix $\boldsymbol{\Gamma}_n = \begin{pmatrix} 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \ddots & \vdots \\ -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix}$ describes the dependency between the reduced sample elements induced by the subtraction of the same mean vector from all rows of $\boldsymbol{X}$. In this matrix $\boldsymbol{X}^*$, one row is redundant because the sum over all rows is zero. As a consequence, $\boldsymbol{\Gamma}_n$ is singular. Thus we omit the last row in $\boldsymbol{X}^*$, yielding the matrix

$$\boldsymbol{X}^{**} = (\boldsymbol{x}_{11} - \bar{\boldsymbol{x}} \quad \cdots \quad \boldsymbol{x}_{Kn_K-1} - \bar{\boldsymbol{x}})' \sim N_{(n-1)\times p}(\mathbf{0}, \tilde{\boldsymbol{\Gamma}}_n \otimes \boldsymbol{\Sigma}),$$

($\tilde{\boldsymbol{\Gamma}}_n$ results from $\boldsymbol{\Gamma}_n$ by deleting the last row and column).

Finally, we consider the matrix

$$\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{\Gamma}}_n^{-1/2}\boldsymbol{X}^{**} \sim N_{(n-1)\times p}(\mathbf{0}, \boldsymbol{I}_{n-1} \otimes \boldsymbol{\Sigma})$$

($\boldsymbol{A}^{-1/2}$ denotes the inverse of a root $\boldsymbol{A}^{1/2}$ of the symmetric positive definite matrix $\boldsymbol{A}$ such that $\boldsymbol{A}^{1/2'}\boldsymbol{A}^{1/2} = \boldsymbol{A}$ and can be derived, e.g., via the Cholesky factorization or the eigenvalue decomposition of $\boldsymbol{A}$). As matrix of independent identically distributed multivariate normal rows with expectation zero, $\tilde{\boldsymbol{X}}$ is left-spherically distributed (cf. Fang, Zhang, 1990) and hence distributional invariant with respect to orthogonal rotations.

The rotations of $\tilde{\boldsymbol{X}}$ are derived as product $\boldsymbol{\Delta}\tilde{\boldsymbol{X}}$. The random rotation matrix $\boldsymbol{\Delta}$ can be generated by first producing a (left-spherical) $(n-1)\times(n-1)$-matrix $\boldsymbol{\Delta}^*$ of independent standard normal variables and orthogonalizing it according to $\boldsymbol{\Delta} = \boldsymbol{\Delta}^*(\boldsymbol{\Delta}^{*\prime}\boldsymbol{\Delta}^*)^{-1/2}$.

Now we can state that the transformation of matrices $\boldsymbol{X}^* \rightarrow \boldsymbol{X}^{**} \rightarrow \tilde{\boldsymbol{X}}$ is unique also in the reversed direction, such that we can reconstruct a distributionally equivalent matrix to the original reduced data matrix $\boldsymbol{X}^*$. If additionally we choose a similarity or distance measure for the statistic d which is invariant to a constant shift in all observation vectors, i.e.,

$$r(\boldsymbol{x},\boldsymbol{y}) = r(\boldsymbol{x}+\boldsymbol{a}, \boldsymbol{y}+\boldsymbol{a}), \tag{6}$$

then the test statistic d can be derived from $\boldsymbol{X}^*$ instead of $\boldsymbol{X}$ and we can also calculate the rotated equivalents to the test statistic $d = d(\boldsymbol{R})$. Condition (6) is met with the squared Euclidean distance, but not with Pearson's correlation coefficient r. Therefore, the rotation test is used here only with the squared Euclidean distance.

Thus summarizing, the rotation test works as follows:

1. Use the original data to calculate the matrix $\boldsymbol{R}$ of pairwise similarity or distance measures fulfilling condition (6), e.g., the squared Euclidean distances,
2. calculate the "original" test statistic d using (5),
3. determine the matrix $\tilde{\boldsymbol{X}}$ from the original data matrix $\boldsymbol{X}$ via total mean reduction (yielding $\boldsymbol{X}^*$), omission of the last line (giving $\boldsymbol{X}^{**}$) and removing the correlation between the rows (multiplication with $\tilde{\boldsymbol{\Gamma}}_n^{-1/2}$, yielding $\tilde{\boldsymbol{X}}$),
4. generate a random rotation matrix $\boldsymbol{\Delta}$ as described above and calculate the rotated matrix $\boldsymbol{\Delta}\tilde{\boldsymbol{X}}$,
5. multiply the rotated matrix (from the left) with $\tilde{\boldsymbol{\Gamma}}_n^{1/2}$ yielding the rotated version of $\boldsymbol{X}^{**}$,
6. calculate the sum over all rows of this matrix and add the sum with reversed sign as additional row (rotated version of $\boldsymbol{X}^*$),
7. derive the matrix $\boldsymbol{R}^*$ of pairwise similarity or distance measures of the rows of the rotated version of $\boldsymbol{X}^*$,
8. calculate the test statistic d^* according to (5) from $\boldsymbol{R}^*$,
9. repeat steps 4 to 8 N_{rot} times (N_{rot} sufficiently large),
10. if m denotes the number of repetitions with $d^* \geq d$, then the p-value of the rotation test is $(m+1)/(N_{\text{rot}}+1)$.

4.2 Application of the rotation test to the examples and to the simulated data

The results of the rotation test based on the squared Euclidean distance are already included in Tables 2 to 5 – as last column denoted $d(\text{E}^2)_{\text{rot}}$. In both fingerprint examples (Tables 2 and 3) it gives the best results, now also allowing for a Bonferroni correction in the six pairwise comparisons of groups.

In the bootstrap samples of the microarray data in Table 4 the rotation test yields better results than the five versions of the PC test. The rejection rates of the permutation test and the rotation test based on the squared Euclidean distance are very similar for sample size 5 and above. The permutation test based on Pearson's r, however, still has the highest rejection rates except for the extremely small sample size 3.

No differences between the performance of the permutation test and the rotation test based on the squared Euclidean distance can be found in the simulation experiments with a multivariate normal distribution of the observation vectors and the moderate sample sizes $n_1 = n_2 = 10$.

4.3 Robustness of the rotation test

In order to apply the rotation test instead of the permutation test, we have to accept the parametric model (1)/(3) rather than the more general nonparametric model (2)/(4). Therefore, we carried out

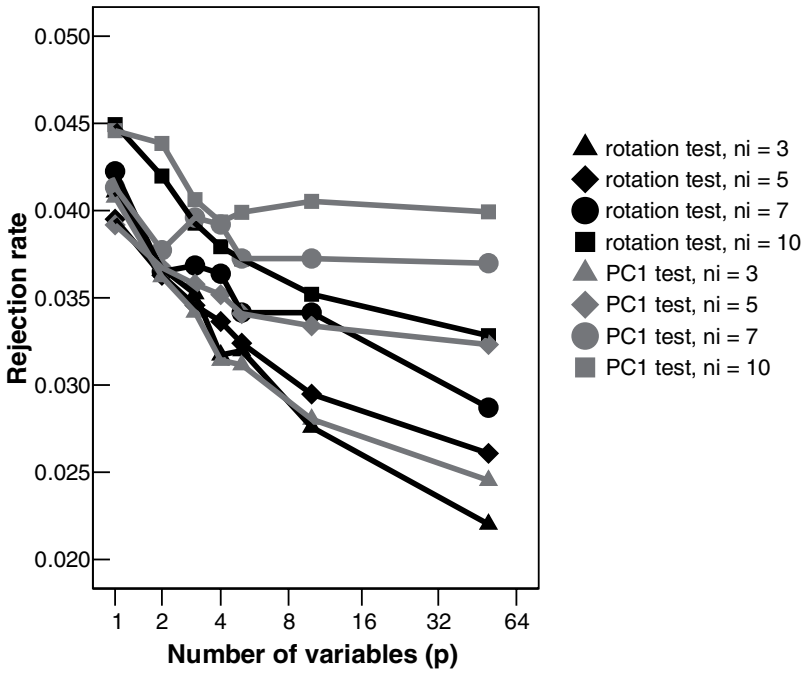

Figure 1 Results of the simulation series under the null hypothesis for an exponential distribution of the p independent variables. The nominal level of the rotation test and of the PC_1 test is always 0.05.

some simulation experiments under the null hypothesis to characterize the behaviour of the test, particularly the control of the type I error, if the normality assumption is violated.

As strong deviation from normality we considered vectors of p independent variables from an exponential distribution. Figure 1 shows the empirical rejection rates of the rotation test based on the squared Euclidean distance and of the PC_1 test for a nominal α of 0.05 for various numbers of variables and various sample sizes. With this extremely skew distribution, the tests are moderately conservative with increasing degree for increasing number of variables. This conservative behaviour is weakened with increasing sample sizes, however only slowly. The results are similar for nominal α of 0.10 or 0.01 (not presented here).

Some analogous experiments with vectors of independent uniformly distributed variables showed that non-normality can also produce an anticonservative behaviour. As an example, with $p = 3$ and sample sizes $n_1 = n_2 = 4$ the rejection rate of the null hypothesis was 0.0596. The deviations from the nominal level are much smaller if the violation of normality is smaller. If, e.g., the variables are generated as the sum of a standard normal and a uniform (in (0,1)) variable, the rejection rate is 0.0502. A similar reduction of the deviations can be stated if the exponential distributions are overlapped with a standard normal distribution (results not presented here).

For some randomly selected configurations, we also checked the rejection rates with normally distributed variables – according to the original parametric model (1)/(3). In all these runs, the nominal error level was kept within the limits of the confidence intervals for the random binomial fluctuations.

5 Discussion

In the paper two strategies for comparisons of independent samples of high-dimensional observation vectors are considered and compared. One of them is based on spherically distributed principal com-

ponent scores (Läuter et al., 1996, 1998). The other one uses pairwise similarity or distance measures between all pairs of sample elements and essentially compares the distances/similarities of pairs from the same group with those from different groups. This second strategy – in the basic version – can be considered as special case of a proposal by Mantel (1967). Though the idea is rather old, it seems not to play an essential role in the current literature.

The examples and simulation experiments presented here show that both types of tests are well applicable in the analysis of high-dimensional data, even with small samples. For the tests based on the pairwise measures, we first used a permutation test which is distribution-free as long as the considered distributions of the populations are identical under the null hypothesis.

For very small samples, however, the small number of possible different permutations restricts the performance of the test. Therefore a so-called rotation test is derived. The test exactly controls the type I error under the usual normal model. For strong deviations from normality, however, the test can be moderately conservative as well as anticonservative. In the present simulation experiments with sample sizes of 2 to 10 per group and the number of variables varying between 1 and 50, the extremely skew exponential distribution yielded decreased rejection rates between 0.02 and 0.05 for a nominal level of 0.05. With uniformly distributed vectors, anticonservative rates up to 0.07 occurred. The deviations from the nominal level were much smaller with less extreme violations of the assumption of multivariate normal data vectors.

The power of the permutation test is essentially dependent on the similarity or distance measure used. With Pearson's correlation coefficient, the test very well detects different patterns in the variables as in the fingerprint data considered here. A global shift over all variables, however, will hardly be found.

Unfortunately, this very usual measure cannot be applied in the parametric rotation test. We propose to use the squared Euclidean distance instead. It is also possible to consider a modified correlation coefficient where the sums of squares are not related to the deviations from the mean over the whole observation vector but to the variablewise means over all sample elements.

Generally, the definition of a suitable similarity or distance measure is easier for variables which have all the same scale. This is often the case with high-dimensional data, as in our examples with greyscale values scanned from the electrophoretic fingerprints or from the microarrays.

As mentioned, some generalizations of the test for pairwise measures are given in Kropf et al. (2004a), particularly tests in a blocked design and tests for dependent measures. Generally, however, the PC-based tests with their background of the general linear models should be more flexible in complex study designs. Approximate permutation tests based on pairwise measures for multi-factorial designs are described in Anderson (2001). Similar methods are also used in the software "CANOCO" by ter Braak and Šmilauer (2002).

Acknowlegdements Parts of the presented work were supported by grants of the German government (BMBF grants 0312629A and 0311295 for the electrophoretic fingerprints and BMBF-NBL3 program for the multivariate tests). The authors thank J. Läuter for helpful discussion of the rotation tests, G. Hambruch, T. Lingner and S. Ribal for assistance in the simulation experiments and the reviewer for constructive advice.

References

Aittokallio, T., Ojala, P., Nevalainen, T. J., and Nevalainen, O. (2000). Analysis of similarity of electrophoretic patterns in mRNA differential display. *Electrophoresis* **21**, 2947–2956.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32–46.

Eszlinger, M., Krohn, K., Frenzel, R., Kropf, S., Tonjes, A., and Paschke, R. (2004). Gene expression analysis reveals evidence for inactivation of the TGF-beta signaling cacade in autonomously functioning thyroid nodules. *Oncogene* **23**, 795–804.

Eszlinger, M., Krohn, K., Berger, K., Läuter, J., Kropf, S., Beck, M., Führer, D., and Paschke, R. (2005). Gene expression analysis reveals evidence for increased expression of cell cycle associated genes and Gq-protein-

protein kinase C signaling in cold thyroid nodules. *Journal of Clinical Endocrinology and Metabolism* **90**, 1163–1170.

Fang, K.-T. and Zhang, Y.-T. (1990). *General Multivariate Analysis*. Science Press Beijing and Springer-Verlag, Berlin Heidelberg.

Hemmelmann, C., Horn, M., Reiterer, S., Schack, B., Süsse, T., and Weiss, S. (2004). Multivariate tests for the evaluation of high-dimensional EEG data. *Journal of Neuroscience Methods* **139**, 111–120.

Hemmelmann, C., Horn, M., Süsse, T., Vollandt, R., and Weiss, S. (2005). New concepts of multiple tests and their use for evaluating high-dimensional EEG data. *Journal of Neuroscience Methods* **142**, 209–217.

Kropf, S. (2000). *Hochdimensionale multivariate Verfahren in der medizinischen Statistik*. Shaker Verlag, Aachen.

Kropf, S., Hothorn, L., and Läuter, J. (1997). Multivariate many-to-one procedures, with application in pre-clinical trials. *Drug Information Journal* **31**, 433–447.

Kropf, S., Heuer, H., Grüning, M., and Smalla, K. (2004a). Significance test for comparing complex microbial community fingerprints using pairwise similarity measures. *Journal of Microbiological Methods* **57**, 187–195.

Kropf, S., Läuter, J., Eszlinger, M., Krohn, K., and Paschke, R. (2004b). Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. *Journal of Statistical Planning and Inference* **125**, 31–47.

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 Journal* **38**, 5–23. Erratum: *Biometrical Journal* **40**, 1015.

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

Läuter, J., Glimm, E., and Eszlinger, M. (2005). Search for relevant sets of variables in a high-dimensional setup keeping the familywise error rate. *Statistica Neerlandica* **59**, 298–312.

Langsrund, Ø. (2005). Rotation tests. *Statistics and Computing* **15**, 53–60.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.

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

ter Braak, C. J. F. and Smilauer, P. (2002). *CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (Version 4.5)*. Microcomputer Power, Ithaca.

