Biometrical Journal **46** (2004) 6, 687–698 DOI: 10.1002/bimj.200410067

Micro Array Based Gene Expression Analysis using Parametric Multivariate Tests per Gene – A Generalized Application of Multiple Procedures with Data-driven Order of Hypotheses

Ernst Schuster[*,1], **Siegfried Kropf**[2], and **Ingo Roeder**[**,1,3]

[1] Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany
[2] Institute of Biometry and Medical Informatics, Otto von Guericke University Magdeburg, Magdeburg, Germany
[3] Interdisciplinary Centre for Bioinformatics, University of Leipzig, Leipzig, Germany

Received 28 May 2004, revised 24 August 2004, accepted 27 August 2004

Summary

Micro array technology allows the simultaneous analysis of ten-thousands of genes. Most often, however, the analysis is based on a few replications only. This causes problems in the application of classical multivariate tests which require sample sizes exceeding the number of observed variables. To overcome these problems, a class of stable, multivariate procedures based on the theory of spherical distributions has been proposed by Läuter, Glimm, and Kropf (1996). These methods allow the use of multivariate information of many genes for testing differential gene expression. Furthermore, multiple testing procedures based on these principles have been constructed (e.g., Kropf, Läuter, 2002), which strictly keep the familywise type I error rate (FWE).

In this paper, these methods have been generalized to allow for the use of full multivariate information on expression intensities of individual genes analysed by the Affymetrix GeneChip technology. In contrast to the usual strategy, which constructs an expression score for each gene, based on averaging of the different oligonucleotide (perfect- and miss-match) information, and then performs some test on these summarized expression values, we suggest using a test procedure based on the complete multivariate perfect match information. We show that a multiple FWE-controlling procedure for normally distributed data proposed by Westfall, Kropf, and Finos (2004), can be generalised to a more powerful procedure based on left-spherically distributed scores derived from the perfect match information, without losing the FWE-controlling property.

To illustrate the proposed test procedures, which have been implemented in the statistical programming environment *R*, we analyse two already published data sets, comparing gene expression of tumour and healthy tissues within identical patients and between two groups of different patients, respectively. Using these examples, we demonstrated that the incorporation of the multivariate perfect match information is superior to classical expression score based methods with respect to the number of identifiable differentially expressed genes.

Key words: Gene expression; Multiple tests; Score-based tests; Data-driven ordered hypotheses.

1 Introduction

Micro array technology allows the simultaneous analysis of tens of thousands of genes. In the simplest case, a paired sample of gene expression values, e.g. tumour versus normal tissue of the same

[*] Corresponding author: e-mail: Ernst.Schuster@imise.uni-leipzig.de

[**] This work has been supported by the Deutsche Forschungsgemeinschaft (DFG): grant BIZ 6/1-2.

individual, is analysed for differential expression. Most often, however, the number of samples n (i.e. hybridised micro arrays) is restricted to a few replicates, which causes problems in the application of classical multivariate test procedures. These procedures require sample sizes exceeding the number p of observed variables (i.e. genes). Facing this problem, the presented work will focus on the application and extension of methods developed by Läuter, Glimm, and Kropf (1998) specifically for this non-classical situation.

Testing of all individual genes accessible by the experimental procedure at a significance level α appropriate for a single test (e.g. $\alpha = 0.05$) would lead to an extremely high number of false positive test results. Therefore, throughout this manuscript, a familywise type I error (FWE) in the strong sense will be considered. Keeping the FWE in the strong sense means, that irrespective of the number of true null-hypothesis, the probability of ending up with at least one false positive test decision is restricted to a given significance level.

Let us first consider the so called *one-sample situation.* Here, one has n independent, identically distributed (iid) p-dimensional sample vectors (which might also be differences from corresponding sample elements of two dependent samples):

$$\boldsymbol{x}_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jp} \end{pmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,, \quad j = 1, \ldots, n$$

with expectation $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}$.

The data matrix subsuming all p-dimensional sample vectors $\boldsymbol{x}_j'$ $(j = 1, \ldots, n)$ is denoted by

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

The multiple test procedures for the local hypotheses $H_i : \mu_i = 0$ $(i = 1, \ldots, p)$ treated in the sequel, are based on a class of global multivariate tests given in Läuter, Glimm, and Kropf (1996). This article also provides the proof of the following theorem.

Theorem 1 Consider the random matrix $\boldsymbol{X}$ as above under the global null hypothesis $H : \boldsymbol{\mu} = \boldsymbol{0}$. Let $\mathbf{d}$ be a p-dimensional weight vector and $\boldsymbol{D}$ a $p \times q$ weight matrix $(q \leq \min(p,\ n-1))$, which depend on $\boldsymbol{X}$ only through the sums of products matrix $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$ (with the additional restriction that rank $(\boldsymbol{Xd}) = 1$ or rank $(\boldsymbol{XD}) = q$, both with probability 1). Then for

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = \boldsymbol{Xd} \quad \text{and} \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_n \end{pmatrix} = \boldsymbol{XD}$$

the following statements hold:

1. The score vector $\boldsymbol{z}$ and the score matrix $\boldsymbol{Z}$ are both left-spherically distributed.
2. The classical t-test for the univariate scores $t = \dfrac{\sqrt{n}\bar{z}}{s_z}$ with $\bar{z} = \dfrac{1}{n}\, \boldsymbol{z}'\boldsymbol{1}_n$ ($\boldsymbol{1}_n$ is a vector consisting of n ones) and $s_z^2 = \dfrac{1}{n-1}\,(\boldsymbol{z}'\boldsymbol{z} - n\bar{z}^2)$ provides an exactly t-distributed statistic with $n-1$ degrees of freedom.
3. Analogously, with $T^2 = n\bar{\boldsymbol{z}}'\boldsymbol{S}_z^{-1}\bar{\boldsymbol{z}}$, where $\bar{\boldsymbol{z}} = \dfrac{1}{n}\,\boldsymbol{Z}'\boldsymbol{1}_n$ and $\boldsymbol{S}_z = \dfrac{1}{n-1}\,(\boldsymbol{Z}'\boldsymbol{Z} - n\bar{\boldsymbol{z}}\bar{\boldsymbol{z}}')$, the statistic $F = \dfrac{n-q}{(n-1)q}\, T^2$ has an exact F-distribution with q and $n-q$ degrees of freedom.

Applying these results, Kropf and Läuter (2002) have proposed the following multiple procedure for testing the univariate local hypotheses H_i, which strictly keeps the familywise type I error level α:

Procedure I:

1. Sort the variables for decreasing values of $w_i = \sum\limits_{j=1}^{n} x_{ji}^2 (i = 1, \ldots, p)$.
2. Carry out, in this order, the usual one-sample t-tests for the variables at the unadjusted error level α as long as significance occurs. Stop at the first non-significant test.

The proof that this procedure keeps the FWE in the strong sense, despite the data-driven ordering of variables, uses Theorem 1 with specific weight vectors. These have weight 1 for the variable with maximum value of w_i and 0 otherwise. This induces a degeneration of the multivariate test to an univariate one using the gene which occupies the key position in Procedure 1. Please note that the w_i are the diagonal elements of the matrix $\boldsymbol{W}$. For more details of the proof we refer to the above cited paper. The basic idea underlying the ordering in Procedure I is the decomposition

$$\sum_{j=1}^{n} x_{ji}^2 = n \cdot \bar{x}_i^2 + \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 = s_i^2 (t_i^2 + n - 1) ,$$

where $\bar{x}_i$ and s_i denote the usual variablewise mean and standard deviation, and $t_i = \frac{\bar{x}_i}{s_i} \sqrt{n}$ the corresponding univariate test statistic. Assuming the variances to be approximately equal, variables with higher absolute means and, therefore, more extreme t-values, will have increased chances to be in the first positions after sorting and to be detected as significant in the test procedure.

The assumption of approximately equal variances in the p variables may appear very restrictive, however, as discussed by Schuster and Kropf (2002) it is rather natural, e.g. in the context of repeated measurements. Furthermore, such an assumption is often achievable by a suitable non-linear transformations of the data. In the context of gene-expression data (see example below), a logarithmic transformation may yield approximately normal variables with roughly equal variances.

If variables with increased variances and no effects in the expectations occur, they may cause an early stopping of the procedure. To reduce this risk, Westfall, Kropf, and Finos (2004) proposed using the same sums of squares w_i (as in the above procedure) to weight the raw P-values in a weighted Bonferroni-Holm procedure (Holm, 1979), which strictly keeps the familywise type I error level α:

Procedure II:

1. Calculate the P-values $P_i (i = 1, \ldots, p)$ for the usual unadjusted t-tests for each of the p variables.
2. For each variable, determine the sums of squares values w_i as above and the weights $g_i = w_i^\eta$ for a fixed value $\eta \geq 0$.
3. Calculate the *weighted P-values* $Q_i = P_i / g_i$ and sort the variables for increasing values: $Q_{i_1} \leq Q_{i_2} \leq \cdots \leq Q_{i_p}$ or $Q_{(1)} \leq Q_{(2)} \leq \cdots \leq Q_{(p)}$, respectively. Define the index sets $S_j = \{i_j, i_{j+1}, \ldots, i_p\} (j = 1, \ldots, p)$.
4. The ordered hypotheses $H_{(j)}$ $(j = 1, 2, \ldots)$ are rejected as long as $Q_{(j)} \leq \frac{\alpha}{\sum\limits_{h \in S_j} g_h}$.
   Stop at the first j yielding a higher value $Q_{(j)}$.

If η equals zero, this weighted Bonferroni-Holm procedure (Procedure II) coincides with the classical Bonferroni-Holm method. Furthermore, as shown previously in Westfall and Krishen (2001) for the case of fixed weights, Procedure II converges to Procedure I if η tends to infinity. Therefore, the choice of a large η is appropriate if the variances are expected to be homogenous to a high degree. In contrast, small values of η are preferable for a considerable expected level of heterogeneity of the p variances. (see also Discussion).

To compare two independent samples $\boldsymbol{x}_j^{(m)} \sim N_p(\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma})$ $(j = 1, \ldots, n_m;\ m = 1, 2)$ (*two-sample situation*) with expectations $\boldsymbol{\mu}^{(m)} = (\mu_1^{(m)}, \ldots, \mu_p^{(m)})'$ and common covariance matrix $\boldsymbol{\Sigma}$, the local null hypotheses $H_i : \mu_i^{(1)} = \mu_i^{(2)}$ $(i = 1, \ldots, p)$ are tested in the corresponding two-sample t-tests and the diagonal elements $w_i = \sum_{m=1}^{2} \sum_{j=1}^{n_m} (x_{ji}^{(m)} - \bar{x}_i)^2$ ($\bar{x}_i$ is the total mean of the i-th variable over all $n = n_1 + n_2$ sample elements) of the sums of products matrix $\boldsymbol{W} = (\boldsymbol{X} - \bar{\boldsymbol{X}})' (\boldsymbol{X} - \bar{\boldsymbol{X}})$ with $\bar{\boldsymbol{X}} = \mathbf{1}_n \bar{\boldsymbol{x}}'$ and $\bar{\boldsymbol{x}} = \frac{1}{n} \mathbf{1}_n' \boldsymbol{X}$ $(n = n_1 + n_2)$ are used. The remaining parts of Procedures I and II are identical to the one-sample case.

To determine the expression level of individual genes, the Affymetrix-type GeneChip® technology uses several different (about 16–20) oligo-nucleotide (oligo) measurements. For each oligo, two different sequences (which differ in one base only), the so called perfect-match (PM) and the miss-match (MM), are determined (Affymetrix, 2002). To estimate the corresponding expression level of the genes, one usually contrasts these two types of matches and thus summarizes in some sense over the oligos belonging to a given gene. Herein, all genes are treated according to one common rule. Subseqently, the obtained expression values are used to ascertain differentially expressed genes.

In contrast to this approach, we propose here a test procedure which incorporates the total information of all individual PMs into the test for differential gene expression, using individual, gene specific rules for the construction of a multivariate expression score. Therefore, Procedures I and II are transferred into another multivariate context. Keeping the control over the FWE, we will use scores from multivariate tests (e.g. principle component test or standardised sum test) proposed by Läuter et al. (1996) as input for the above introduced multiple procedures. Summarizing the above statements, one can say that the essential novelty of our approach is the use of data dependent, gene specific weights rather than one common, gene independent rule in the construction of expression scores.

In the following section the derivation of the amended multiple procedures is described in detail. Moreover, we provide the proof for its strong control of the claimed FWE criteria. Two examples using published data are presented in Section 3, followed by a discussion in the last section. A technical description of two R-routines for the application of the proposed procedures is outlined in the appendix.

2 Multiple Test Procedures with Parametric Multivariate Tests per Gene

Our approach is based on Affymetrix raw data values including the information of all oligos per gene (i.e. *cel*-file level). It is assumed that the PMs follow a common multivariate normal distribution. Let k be the number of genes and p_i the oligos number of gene i. From this, it follows that the total number of PMs is given by $p = \sum_{i=1}^{k} p_i$.

Firstly, we consider the one-sample situation, where the data usually represent difference values from two conditions in the same individual. The column vector containing all p_i PMs of individual j and gene i is denoted by $\boldsymbol{x}_{ji}$ ($j = 1, \ldots, n$; $i = 1, \ldots, k$). Thus, the whole vector for individual j is composed as

$$\boldsymbol{x}_j = \begin{pmatrix} \boldsymbol{x}_{j1} \\ \vdots \\ \boldsymbol{x}_{jk} \end{pmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (j = 1, \ldots, n)$$

with expectation $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_k \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \boldsymbol{\Sigma}_{1k} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{k1} & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix}$.

Again, the n sample vectors are summarized to the $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{X}_1 \cdots \boldsymbol{X}_k)$ with the transposed vectors $\boldsymbol{x}_j$ as rows, and the total sums of products matrix $\boldsymbol{W}$ is given by

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \cdots & \boldsymbol{W}_{1k} \\ \vdots & \ddots & \vdots \\ \boldsymbol{W}_{k1} & \cdots & \boldsymbol{W}_{kk} \end{pmatrix} = (\boldsymbol{X}_1 \cdots \boldsymbol{X}_k)' (\boldsymbol{X}_1 \cdots \boldsymbol{X}_k) = \boldsymbol{X}'\boldsymbol{X}\,.$$

Now, the local hypotheses $H_i : \boldsymbol{\mu}_i = \mathbf{0}$ (i = 1,..., k) are multivariate and are treated according to Theorem 1. The scores derived in these tests are then used in Procedures I and II in the same manner as the single variables previously.

Here, we will consider four different choices for the weight vectors from Theorem 1 (for details see Läuter et al., 1996 or Kropf, 2000). The weight vector for gene i (denoted by $\boldsymbol{c}_i$, $i = 1, \ldots, k$) is obtained for

1. the *non standardized principle component (NPC) test* as the eigenvector $\boldsymbol{c}_i$ corresponding to the largest eigenvalue of the eigenvalue problem $\boldsymbol{W}_{ii}\boldsymbol{c}_i = \boldsymbol{\lambda c}_i$ with $\boldsymbol{c}_i'\boldsymbol{c}_i = 1$;
2. the *standardized principle component (SPC) test* as the eigenvector $\boldsymbol{c}_i$ of the largest eigenvalue of the eigenvalue problems $\boldsymbol{W}_{ii}\boldsymbol{c}_i = \text{Diag}\,(\boldsymbol{W}_{ii})\,\boldsymbol{c}_i$ with $\boldsymbol{c}_i'\,\text{Diag}\,(\boldsymbol{W}_{ii})\,\boldsymbol{c}_i = 1$;
3. the *covariance sum (CS) test* as $\boldsymbol{c}_i = [\text{Diag}\,(\boldsymbol{W}_{ii})]^{-1}\,\boldsymbol{W}_{ii}[\text{Diag}\,(\boldsymbol{W}_{ii})]^{-1/2}\,\mathbf{1}_{p_i}$;
4. the *standardised sum (SS) test* as $\boldsymbol{c}_i = [\text{Diag}\,(\boldsymbol{W}_{ii})]^{-1/2}\,\mathbf{1}_{p_i}$.

To ensure that genes with balanced increased *and* decreased oligo measurements are *not* counted as differentially expressed, the weights $\boldsymbol{c}_i$ are transformed to the modified form $\boldsymbol{d}_i$ by:

$$\boldsymbol{d}_i = (d_{li})_{l=1,\ldots,p_i} \quad \text{with} \quad d_{li} = \frac{|c_{li}|}{\sum\limits_{s=1}^{p_i} |c_{si}|}\,, \qquad l = 1, \ldots, p_i;\ i = 1, \ldots, k\,.$$

Using the transformed weights, the p_i-dimensional data subvectors $\boldsymbol{x}_{ji}$ are subsumed into a score by $z_{ji} = \boldsymbol{d}_i'\boldsymbol{x}_{ji}$. The one-sample t-test with the score values z_{ji} (j = 1, ..., n) now yields the unadjusted P-value P_i for gene i (i = 1, ..., k). Using the following theorem one can formulate generalized versions of the multiple test Procedures I and II.

Theorem 2 Procedures I and II strictly meet a given FWE criteria if applied to left-spherically distributed scores z_{ij}.

To indicate the different assumptions (mulivariate normal and left-spherical), Procedures I and II will be denoted as Procedure I′ and II′, respectively, if applied to left-spherically distrubuted scores.

The subsequent proofs are extensions of the proofs given for Procedures I and II in Kropf and Läuter (2002) and Westfall et al. (2004), respectively.

Proof of Theorem 2, Procedure I′ Let $M_0 = \{i \mid \boldsymbol{\mu}_i = \mathbf{0}, i \in \{1, \ldots, k\}\} = \{i_1, \ldots, i_{k_0}\}$ of size k_0 be the index set of the true local null hypotheses, i.e., of all those genes with expectation zero in all oligos, and let $\boldsymbol{X}_0$ and $\boldsymbol{W}_0$ be the corresponding submatrices of $\boldsymbol{X}$ and $\boldsymbol{W}$, respectively. The case of empty M_0 can be neglected because no type I errors can occur then. All corresponding weight vectors $\boldsymbol{d}_{i_1}, \ldots, \boldsymbol{d}_{i_{k_0}}$ are summarized into the weight matrix

$$\boldsymbol{D}_0 = \begin{pmatrix} \boldsymbol{d}_{i_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{d}_{i_{k_0}} \end{pmatrix}.$$

As the matrix $\boldsymbol{X}_0$ consists of iid multivariate normal row vectors with expectation zero, and all four weight versions (incl. transformation from $\boldsymbol{c}_i$ to $\boldsymbol{d}_i$) use the data only through the corresponding sums of products matrix $\boldsymbol{W}_0$, we are in the framework of Theorem 1. Therefore, the whole matrix $\boldsymbol{Z}_0 = \boldsymbol{X}_0\boldsymbol{D}_0$ is left-spherical as well as each of its columns which correspond to the scores in Procedures I′ and II′. Thus, as shown in Läuter et al. (1998), Theorem 1 can be applied to the scores in the same way as to the raw data before.

Let i_0 denote the index of the score with maximum w_i from all scores from M_0 (uniquely determined with probability 1). Then a test according to Theorem 1 with the scores from M_0 and with the weight vector

$$\boldsymbol{d} = (d_i)_{i=1,\ldots,k_0}\,, \quad \text{where} \quad d_i = \begin{cases} 1 & \text{for} \quad w_i = \max\limits_{l=1,\ldots,k_0} w_l \\ 0 & \text{else} \end{cases}$$

exactly keeps the type I error. That means that the local null hypothesis for the first score after sorting would be accepted with probability $1 - \alpha$ and the procedure would stop before any type I error occurs, if we would confine the procedure to those scores from M_0. Actually, M_0 is unknown, and the sorting includes all scores. Therefore, the order obtained within M_0 is filled up with additional scores. This does not influence the tests with the scores within M_0, but the tests for the additional scores (which cannot produce type I errors per definition) make the procedure more conservative if $k_0 < k$.

Proof of Theorem 2, Procedure II′ With the same notations as above, we consider the conditional distribution of $\mathbf{Z}_0$ for fixed $\mathbf{Z}_0'\mathbf{Z}_0$. According to the theory of spherical distributions, this conditional distribution is again left-spherical. Therefore, the tests with the scores from M_0 (columns of $\mathbf{Z}_0$) are again exact level α tests. If we additionally notice that the sums of squares w_i used in Procedure II′ as well as the weights g_i are fixed for all $i \in M_0$ in this conditional situation, then we can state that

$$P\left(\frac{P_i}{g_i} \le \frac{\alpha}{\sum\limits_{m \in M_0} g_m}\right) = P\left(P_i \le \frac{\alpha g_i}{\sum\limits_{m \in M_0} g_m}\right) = \frac{\alpha g_i}{\sum\limits_{m \in M_0} g_m} \quad \text{for all} \quad i \in M_0 \,.$$

Finally with $g_m \ge 0$ $(m = 1, \ldots, k)$,

$$P\left(\min_{i \in M_0} \frac{P_i}{g_i} \le \frac{\alpha}{\sum\limits_{m=1}^{k} g_m}\right) \le P\left(\min_{i \in M_0} \frac{P_i}{g_i} \le \frac{\alpha}{\sum\limits_{m \in M_0} g_m}\right) = P\left(\bigcup_{i \in M_0}\left(\frac{P_i}{g_i} \le \frac{\alpha}{\sum\limits_{m \in M_0} g_m}\right)\right) \le \sum_{i \in M_0} \frac{\alpha g_i}{\sum\limits_{m \in M_0} g_m} = \alpha \,.$$

As this is true for arbitrary, fixed matrices $\mathbf{Z}_0'\mathbf{Z}_0$, it is true for the unconditioned distribution as well. Therefore, again the test for the first score that could produce a type I error accepts the corresponding local null hypothesis and makes the procedure stop with probability 1-α, so that the multiple type I error is α for $k_0 = k$ or even smaller.

In the comparison of *two independent samples* of sizes n_1 and n_2, respectively, we proceed quite analogously as outlined in the introduction for the original Procedures I and II:

The sample vectors are multivariate normal, i.e. $\boldsymbol{x}_j^{(m)} \sim N_p(\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma})$ $(j = 1, \ldots, n_m; m = 1, 2)$ with expectations $\boldsymbol{\mu}^{(m)} = \begin{pmatrix} \boldsymbol{\mu}_1^{(m)} \\ \vdots \\ \boldsymbol{\mu}_k^{(m)} \end{pmatrix}$ and common covariance matrix Σ.

The local null hypotheses considered here are $H_i : \boldsymbol{\mu}_i^{(1)} = \boldsymbol{\mu}_i^{(2)}$ $(i = 1, \ldots, k)$. Then the sums of products matrix $\boldsymbol{W}$ and its submatrices $\boldsymbol{W}_{ii}$ are defined by

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \cdots & \boldsymbol{W}_{1k} \\ \vdots & \ddots & \vdots \\ \boldsymbol{W}_{k1} & \cdots & \boldsymbol{W}_{kk} \end{pmatrix} = (\boldsymbol{X} - \bar{\boldsymbol{X}})' \, (\boldsymbol{X} - \bar{\boldsymbol{X}})$$

with $n = n_1 + n_2$, $\bar{\boldsymbol{x}}' = \frac{1}{n}\,\mathbf{1}_n'\boldsymbol{X}$, and $\bar{\boldsymbol{X}} = \mathbf{1}_n\bar{\boldsymbol{x}}'$.

Furthermore, the one-sample t-tests are replaced by the two-sample t-tests for the scores. Otherwise, the procedures are identical to the one-sample case. The proofs use straightforward extensions of the corresponding proofs in Kropf and Läuter (2002) and in Westfall et al. (2004), respectively.

3 Worked Examples

3.1 One-sample situation

The first given example will focus on the one-sample situation. The clinical data (Eszlinger, Krohn, and Paschke, 2001) consists of 15 patients with autonomously functioning tyroid nodules (AFTNs).

For each patient the differential expression of 12625 genes between AFTN and surrounding tissue has been determined using Affymetrix GeneChips U95Av2. The raw data is available in cel-file format. Each gene is characterized from between 16 and 20 probe pairs consisting of perfect match (PM) and miss match (MM) information. According to the intent of this work, the following analysis is based on the PM information only and not as usual on a summarized expression score per gene. The applied procedures are implemented in the statistical programming language *R* (Ihaka and Gentleman, 1996; www.R-project.org) using the Bioconductor package (www.bioconductor.org). A detailed technical description of these procedures is given in the Appendix.

Our analysis is based on *AffyBatch* objects obtained from the raw data by application of quantil-normalization without background correction. For the sake of comparison of our results with "classical" methods, summarized expression scores per gene were also calculated using the *R*-method *expresso* with the options *medianpolish* and *pmonly.* To approximately meet the criteria of multivariate normality and variance homogeneity, all expression values (PM's as well as expression scores) have been logarithmized. The expression ratio of trait and surrounding tissue of a specific gene is expected to be one, if its expression is not influenced be the trait. Therefore, the difference of the logarithms will be tested against zero. In order to check the robustness of the results with respect to the choice of raw-data transformation, we repeated the analysis using an arsinh transformation instead of the logarithmic transformation. Because the results are nearly identical, we restricted our presentation to the log-transformed data. The familywise significance level has been set to $\alpha = 0.05$.

Because the test results of Procedure I are contained as a special (limiting) case in Procedure II (see Section 1) we will give only results obtained by the application of Procedures II and II′. Procedure II will be applied for the "classical" case of summarized expression scores per gene (denoted as "control"). In addition, the proposed total-PM-based Procedure II′ has been applied to the four previously described multivariate methods.

Figure 1 shows the number of differentially expressed genes determined as significant for values of η between 0 and 6. It should be noted again that all given results strictly meet the claimed FWE criteria. The results demonstrate that two versions of the proposed total-PM-based procedure (NPC test, CS test) are uniformly superior to the classical *t*-test approach in the sense of finding more

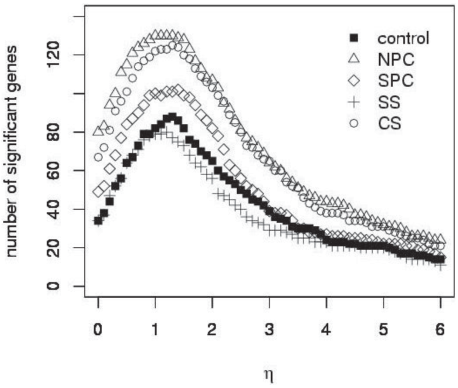

Figure 1 Results of Procedures II and II′ for example data in the one-sample setting. The numbers of significant (according to FWE) differentially expressed genes are given depending on parameter η. Control results have been obtained by application of the classical expression score (method: *medianpolish*) based *t*-test within Procedure II.

significant genes without violating the FWE criteria. It is hypothesised that this superiority can be attributed to more effective utilization of the total multivariate information contained in the individual PMs.

In this example an η-choice of 1.3, giving 88 (Procedure II) and 130 (NPC in Procedure II$'$) significant genes, respectively, would be optimal. Within these two sets of significant genes, there is a common subset of 78 genes. Furthermore, we would like to emphasize that the widely used Bonferroni-Holm procedure ($\eta = 0$) is better than the choice of $\eta \geq 4$ including $\eta = \infty$, i.e. Procedures I and I$'$ (not shown).

Concerning the different versions of Procedure II$'$, one recognizes that in this example the covariance sum test and the non standardized principle component test are qualitatively similar, whereas the standardized principle component test is noticeably worse. The standardized sum test is even inferior to the classical *t*-test within Procedure II and is only included for completeness. Its application is discouraged.

It should be emphasized at this point, that the η must be chosen beforehand to strictly ensure the FWE. If the sample size are not too small, we suggest using $\eta = 1$ (see discussion below).

3.2 Two-sample situation

To demonstrate the application of the proposed test strategy in the two-sample setting we will use data from lung cancer patients. This data set has been analyzed by Bhattacharjee et al. (2001) and the raw data (cel-files) are freely available in the internet (www.genome.wi.mit.edu/MPR/lung). Using gene expression profiles, Bhattacharjee et al. classified the adenomas into subcategories. In the following, we will use the group of 6 small-cell lung cancers (SCLC) and the control group of 17 normal, non-malignant lungs (NL).

To analyze which genes are differentially expressed between these two groups we will apply Procedure II$'$ and for comparison Procedure II, bearing in mind that we must use the versions for two independent samples. The pre-processing of the raw data has been carried out exactly as described in Section 3.1.

The results in Figure 2 show again the superiority of Procedure II$'$ (despite of the unfavourable SS test version) in the region $0 \leq \eta \leq 3$. η-values greater than 3 are not of specific interest because these

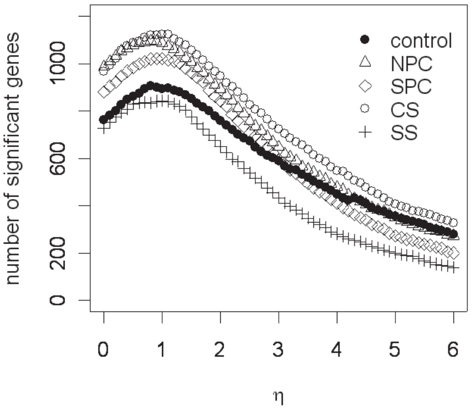

Figure 2 Results of Procedures II and II$'$ for (sample) data in the two-sample setting. The numbers of significant (according to FWE) differently expressed genes are given depending on the parameter η. Control results have been obtained by application of the classical expression score (method: *medianpolish*) based *t*-test within Procedure II.

will give results which are inferior to the classical Bonferroni-Holm procedure. The optimal η-choices for both procedures are 0.8, resulting in 905 (Procedure II, $\eta = 0.8$) and 1096 (Procedure II$'$, NPC test, $\eta = 0.8$) FWE-significant genes, respectively. The set of common genes in both groups contains 815 genes. Please notice, that in this example the NPC test will loose its superiority compared to the classical *t*-test for $\eta \geq 6$. Furthermore, in this example the globally best performance is achieved by the covariance sum test version of Procedure II$'$, which results in the specification of 1122 significant genes at $\eta = 1$.

Again, to strictly meet the claimed significance level, the test method, including the η value, must be chosen beforehand. Similarly, to the one-sample setting, the NPC test and the covariance sum test would be preferential choices.

4 Discussion

As illustrated by the examples in the last section, it can be expected that the proposed total PM-based multivariate test strategies commonly provide more satisfactory results than the "classical" methods based on a summarized gene expression value. This is also affirmed by the application of these methods to further subgroups of the data set described by Bhattacharjee et al. (2001), e.g. the analysis of 21 squamous cell lung carcinomas und 20 pulmonary carcinoids. All five additional group comparisons showed a superiority (with respect to the number of FWE-significant genes) of Procedure II' either using the NPC or the CS test method compared to the expression score (*medianpolish*) based Procedure II (data not shown). Whereas no definite assessment about the preference between the NPC and the CS test methods can be concluded from our results, the application of the SPC and the SS methods are discouraged.

We would like to emphasize that the data compression of the total PM information into scores according to the methods discussed in Section 2 may also be part of a multistage procedure (see Läuter et al., 1998). i.e., the generation of further multivariate scores and, therefore, other analysis procedures within the theory of spherical distributions can be based on these scores.

In contrast to our approach which uses PM information only, some authors have suggested using perfect and miss-match information equally, irrespective of the non-matching base in the MM probes. This approach is justified by the observation that even MM probes (which are supposed to have lower binding affinities for the target sequence than the PM probes) produce signals noticeably above the background level. Although the suggested procedures have been demonstrated for the use of PMs only, they can gererally include both, PM and MM information, into the analyis. A similar test (including PM and MM) has been proposed by Naef et al. (2002). In contrast to his proposal, our approach is able to guarantee a strict control of the FWE criteria. The only assumption needed is that of (approximately) multivariate normal distribution of the probewise data, which may be achieved by suitable transformation. Although we have formally proven the preservation of the FWE criteria under the specified assumptions, we carried out 500 replications for each of the two examples with randomly permuted class labels. The proportion of replications with at least one significant gene was below the nominal level of 5%, the maximum of all runs being $18/500 = 0.036$. Thus, these series empirically support the validity of the theorems and the robustness for deviations from the parametric assumptions.

The proof that the proposed procedures strictly meet a given family-wise error criteria is based on the left-spherical distribution of the used score values. This property, however, is only guarantied (by Theorem 1) for the given linear score. The use of linear scores may appear to be a restriction of the method, however, it covers such important situations as a weighted average of the PMs or the contrast of PMs and MMs. The analysis, whether specific nonlinear scores fulfill the left-spherical property has not yet been performed.

As already mentioned in Section 2, the choice of η remains a critical problem. To strictly meet the claimed FWE significance, it must be chosen before the application of the test procedure. If there is prior knowledge from previously conducted or pilot studies, simulation methods can be applied in order to choose η. Otherwise, the choice of $\eta = 1$, as in the example, may be acceptable for sample sizes that are not too small.

It should be noted that the data pre-treatment (i.e. the applied normalization and background correction methods) can sensitively influence the results of the discussed test procedures. However, to focus on the illustration of the new multivariate oligo-based test approach we restricted ourselves to one fixed, widely used pre-treatment regime (*quantile normalisation without background correction*). A detailed analysis of pre-treatment strategies and their influence on the resulting lists of differentially expressed genes was omitted in the current work for reasons of brevity.

In addition to the advantages of the proposed strategy with respect to the number of detectable differentially expressed genes, our approach has further potential. For example, it is known that the different oligonucleotide probes show variable, sequence-specific hybridisation properties (Hekstra et al., 2003; Binder et al., 2004). This information, which is expected to improve the sensitivity and the correctness of gene expression measurements considerably in the future, could easily be incorporated into the proposed multivariate total PM-based score.

5 Appendix

5.1 Usage of the *R*-implementation of Procedures I′ and II′

In the following the usage of two *R*-routines, which can be applied to realise the proposed test procedures, is described. The implementation has been carried out under *R* version 1.8.1 (www.R-project.org) and depends on the libraries *Biobase* and *affy* both of which can be obtained from www.bioconductor.org.

Whereas *mult.pm.t.test* performs a multivariate version of the classical *t*-test, based on the total information of all oligos per gene, *mt.westfall* realises the multiple testing algorithm described above.

The *R*-code of both routines can be obtained from the authors.

5.2 mult.pm.t.test

This function performs a multivariate version of the classical *t*-test, e.g. for testing differential gene expression in Affymetrix-based micro array experiments using all perfect match information per gene. It is based on the general theory of spherical tests (Läuter et al., 1998).

To apply *mult.pm.t.test* it is necessary to specify an *AffBatch* data object which contains the raw data of all micro array experiments involved in the analysis. Furthermore, two *factor* objects must be defined. One, the *trait*-vector, contains the information, which gene chip belongs to the control and the trait measurements. The second object, the *id*-vector, contains the information which gene chip is assigned to which individual. This is only relevant in the situation of testing dependent pairs of tissue samples in identical individuals (*two sample situation*), however, for reasons of generality it should be specified.

Example 1 Testing of 5 dependent pairs of tissue samples, with array numbers 1 and 2, 3 and 4, . . . , 9 and 10 containing the control and trait pairs of the individuals, respectively.

```
> trait <- factor(c(0,1,0,1,0,1,0,1,0,1))
```

```
> id <- factor(c(1,1,2,2,3,3,4,4,5,5))
```

Example 2 Testing differential gene expression between 4 arrays of control samples (the first four in the Affybatch-object) and 3 arrays of independent trait tissue samples:

```
> trait <- factor(c(0,0,0,0,1,1,1))
```

```
> id <- factor(1:7)
```

The specification, whether a dependent or independent situation is considered, must be done using the Boolean parameter *dependence* which is *FALSE* by default.

The *method* argument specifies the method for determining the score function. It is possible to choose among the following options (for a detailed description of these procedures see Läuter et al. (1996)):

"npc" ... non-standardized principle component test

"spc" ... standardized principle component test

"ss" ... standardized sum test

"cs" ... covariance sum test

The last specific argument for this multivariate *t.test* procedure is *oligo,* which choses the included probe types. Possible options are:

"pm" ... use of perfect-match probes only

"mm"... use of miss-match probles only

"both" ... use of perfect- and miss-match probes.

In addition to these arguments, *mult.pm.t.test* can be run with further *t.test* specific arguments as there are *alternative*, *mu*, and *two.sided* (for details see R-help).

A typical call of *mult.pm.t.test* would, therefore, be

```
> mult.pm.t.test(affybatch.object, factor(c(0,0,0,1,1,1)), factor(1:6),
dependence=FALSE, method="npc", oligo="pm")
```

The return value of the procedure is a list containing the two components *details* and *summary.* Whereas *details* is another list comprising the following components

test ... a list of class "htest" containing identical components as the value of t.test{ctest}

$$\left.\begin{array}{ll} w & \ldots \textit{ sum of squares} \\ z & \ldots \textit{ vector of scores} \end{array}\right\} \textit{(see notation in Section 2, procedure } I' \textit{ and } I''\textit{, respectively)}$$

summary is a three column *data.frame* containing the gene (probe set) names, the raw *P*-values, and the sum of square values, repectively.

5.3 mt.westfall

The function *mt.westfall* performs a multiple procedure to determine significance levels for local hypothesis at a given familywise type I error rate (FWE) based on a weighted FWE-controlling method introduced by Westfall et al. (2004).

There are three arguments to be specified for the application of *mt.westfall:* A data frame, containing a column of gene-wise raw *P*-values (compulsory column name: "*pvalue*") and a column of corresponding weights (e.g. sums of squares as in section 2; compulsory column name: "*weight*"), a parameter vector of η values for which the test procedure will be applied, and the familywise type I error rate α. One possibility to obtain the data frame is the use of the *summary* component of the *mult.oligo.t.test* result.

A typical call of *mt.westfall* using four different choices of η would be

```
> test.result <- mult.pm.t.test(affybatch.object, factor(rep(c(0,1),5)),

factor(c(1,1,2,2,3,3,4,4,5,5)), dependence=TRUE, method="cs")

> mt.westfall(test.result$summary, eta=c(0,1,10,100), alpha=0.05)
```

The return value of *mt.westfall* is a list containing the following components:

eta ... *the actually applied* η *parameter*
nsig ... *the number of significant genes (at the given familywise type I error rate)*
genes ... *a four column data frame containing the names, the raw P-values, the adjusted P-values, and the q-values (the weighted P-values from Procedure II′) of genes determined as significant at the given familywise type I error rate* α.

Acknowledgement The authors are grateful to J. Läuter for his constructive advice regarding the methodology of stable multivariate statistics, and to M. Eszlinger for providing the raw data on autonomously functioning tyroid nodules. They also thank the referees for their helpful comments to an earlier version of this paper.

References

Affymetrix (2002) *Affymetrix Microarray Suite User Guide*. Version 5 edn. Affymetrix Santa Clara, CA.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C. et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America* **98** (24), 1790–1795.

Binder, H., Kirsten, T., Loeffler, M., and Stadler, P. F. (2004). The sensitivity of microarray oligonucleotide probes – variability and the effect of base composition, submitted.

Eszlinger, M., Krohn, K., and Paschke, R. (2001). cDNA expression array analysis suggests a lower expression of signal transduction proteins and receptors in cold and hot thyroid nodules. *Journal of Clinical Endocrinology and Metabolism* **86**, 4834–4842.

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

Hekstra, D., Taussig, A. R., Magnasco, M. and Naef, F. (2003). Absolute mRNA concentration from sequence-specifc calibration of oligonucleotide arrays. *Nucleic acids research* **31**, 1962.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics.* **6**, 65–70.

Ihaka, R. and Gentleman, R. (1996). A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5** (3), 299–314.

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

Kropf, S. and Läuter, J. (2002). Multiple Tests for Different Sets of Variables Using a Data-Driven Ordering of Hypotheses, with an Application to Gene Expression Data. *Biometrical Journal* **44**, 789–800.

Läuter, J., (1996). Exact *t* and *F* Tests for Analysing 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. Correction: *Annals of Statistics* **27**, 1441.

Naef, F., Lim, D. A., Patil, N., and Magnasco (2002). DNA hybridization to mismatched templates: A chip study. *Physical Review E* **65**, 040902.

Schuster, E. and Kropf, S. (2002). A New Proposal for Pairwise Multiple Comparisons with Repeated Measurements. *Pakistan Journal of Statistics* **18** (2), 197–211.

Westfall, P. H. and Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* **99**, 25–40.

Westfall, P. H., Kropf, S., and Finos, L. (2004). Weighted FWE-controlling methods in high-dimensional situations. Accepted for: *New developments in multiple comparison procedures*. IMS Lecture Notes – Monograph Series.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing*. John Wiley & Sons, New York.