Significance Analysis of Microarrays using Rank Scores

M.A. van de Wiel[1]

Abstract

The Significance Analysis of Microarrays (SAM) software is a very practical tool for detecting significantly expressed genes and controlling the proportion of falsely detected genes, the False Discovery Rate (FDR). However, SAM tends to find biased estimates of the FDR. We show that the same method with the data replaced by rank scores does not have this tendency. We discuss the choice of the rank score function in view of the power of this nonparametric multiple testing procedure. Moreover, we introduce a testing formalization of the popular 2-fold rule. This testing procedure is more selective than the basic procedure and it enables the scientist to make a stronger statement about the selected genes than with the 2-fold rule. All procedures are illustrated with the example one-class data available in the SAM software.

1 Introduction

Tusher et al. (2001) introduced Significance Analysis of Microarrays (SAM) as a statistical technique for finding significant genes in microarrays. This technique aims to control the False Discovery Rate (FDR), which is the proportion falsely rejected null hypotheses among all rejected null hypotheses. Within the microarray framework a null hypothesis usually corresponds to a statement like 'the gene is not (differentially) expressed'. Usage of SAM is enhanced by the free SAM Excel add-in that is available via `http://www-stat.stanford.edu/~tibs/SAM`. SAM has the potential of becoming a standard technique and is already used in some medical studies (see e.g. Sørlie et al. (2001)). SAM has, according to Pan et al. (2001) and Efron et al. (2000), one major disadvantage: estimation of the number of significant genes is biased, especially when this number is relatively large. This was our main motivation for developing Significance Analysis of Microarrays using Rank Scores (SAM-RS).

There is a close connection between SAM and the FDR-based approach introduced by Benjamini and Hochberg (1995). In fact, SAM is a version of this approach which controls the critical levels for the multiple testing procedure in a specific way. Other approaches for finding significantly expressed genes are: strong control of the family wise error rate (FWE), which is discussed in Dudoit et al. (2000), and various modelling techniques (e.g. mixture models: Pan et al. (2001), ANOVA: Wolfinger et al. (2001), empirical Bayesian: Efron et al. (2001a)). Control of the FWE implies control of the probability that any gene is falsely called significant under any mixture of expressed and non-expressed genes. While, as opposed to the FDR-based approaches, strong FWE control really controls the type I testing error, it can be too conservative. Hence, there is a trade-off between type I error and lack of power. As for the modelling techniques: obviously, these may give more insight in the structure of the set of genes and dependencies between gene expressions than the direct multiple testing approaches, but they are more complex, often problem specific and usually rely on assumptions on the distributions of the gene expressions.

SAM deals with dependency between gene expressions by assuming that the dependency structure is the same under the alternative hypotheses as under the null hypotheses. A

[1]Department of Mathematics and Computing Science
Eindhoven University of Technology
P. O. Box 513, 5600 MB Eindhoven, The Netherlands
`markvdw@win.tue.nl`

more complex approach to handle dependency was developed by Storey and Tibshirani (2001) in the context of the positive FDR.

Within an empirical Bayesian setting Efron et al. (2001b) used a Wilcoxon rank-sum statistic, because they prefer to use permutation based estimates of the null density above using normal-theory. We show that within the SAM framework, use of rank statistics is not only preferable because of their distribution-freeness property, but even more so to obtain unbiased estimates of the expected number of falsely called genes. Hence, SAM-RS allows for better control of the FDR. The choice for Wilcoxon rank scores may very well be too discrete; using normal rank scores, which are inverse standard normal transformations of the ranks, may be more sensible. It is well-known in nonparametric testing theory that normal rank score statistics are asymptotically as efficient under normality as t-statistics. More importantly, it was demonstrated by Klotz (1963) that even for small sample sizes the efficiency of a normal scores test is high. Using an example data set provided by the SAM software, we observe in section 4 that also in this dependent multiple testing setting the normal rank score statistics do their work well for $n = 8$. A very practical point of SAM-RS is that it easily fits into the SAM software.

In section 3 we propose two procedures that incorporate more selective criteria within the statistical tests as opposed to the often used '2-fold rule', which is applied outside the multiple test as an extra criterion for genes to be called. The proposed procedures are illustrated with example data sets.

We discuss SAM-RS for paired data and one-class data, and discuss how to adapt SAM-RS to other data structures as unpaired two class data, quantitative response and censored data in section 5.

2 SAM-RS for paired data

Microarray data analysis is usually concerned with a huge number of genes, denoted by p, and a small number of experimental objects or conditions (like persons or time since start of treatment), denoted by N. In case of one-class data, Z_{ij} is the gene expression for the ith gene and jth experimental object. For paired data, we define $Z_{ij} = X_{ij} - Y_{ij}$, where X_{ij} and Y_{ij} are gene expressions for the same experimental objects, but under one different condition, for instance before and after treatment. Moreover, let $Z_i = (Z_{i1}, \ldots, Z_{iN})$. Denote the distribution function of $Z_{ij}, j = 1, \ldots, N$ by F_i having mean μ_i. SAM and SAM-RS are used for multiple testing of

$$H_{0i} : \mu_i = 0 \quad \text{against} \quad H_{1i} : \mu_i \neq 0 \tag{1}$$

for $i = 1, \ldots, p$.

2.1 SAM algorithm for rank statistics

The SAM procedure is described in Tusher et al. (2001) and in a more general setting in Chu et al. (2001). Let us now discuss SAM-RS which is highly similar to SAM.

Let T be a linear signed-rank statistic:

$$T(Z_i) = \sum_{j=1}^{N} \text{sgn}(Z_{ij}) a(R(|Z_{ij}|)), \tag{2}$$

where $R(|Z_{ij}|)$ denotes the rank of $|Z_{ij}|$ in Z_i and $a(u)$ is a rank score function. Function $a(u) = u$ results in the popular Wilcoxon signed-rank statistic. First, compute $r_i = T(Z_i)$

for $i = 1, \ldots, p$. Let $r_{(i)}$ be the ith order statistic in $\vec{r} = (r_1, \ldots, r_p)$. Now, obtain B permutation versions of $\vec{r}$ by multiplying the scores within one column by -1 or +1 with equal probability. The correlation structure between genes is maintained by using the same multiplication factor within an experimental object. Denote the realization of $r_{(i)}$ in the bth sign permutation version by $r^b_{(i)}$ and calculate expected null scores $\bar{r}_{(i)} = \sum_{b=1}^{B} r^b_{(i)}/B$.

Then, estimate the number of rejections under H_{0i} for given threshold t:

$$\text{Null}(t) = \underset{b=1,\ldots,B}{\text{med}}\left(\#\{i : |r^b_{(i)} - \bar{r}_{(i)}| > t\}\right), \tag{3}$$

where $\text{med}_{b=1,\ldots,B}$ denotes the median over all permutations $b = 1, \ldots, B$. Similarly, we have the actual total number of rejections for given t:

$$\text{Total}(t) = \#\{i : |r_{(i)} - \bar{r}_{(i)}| > t\} \tag{4}$$

. Changing t allows for control of the proportion of false rejections, the False Discovery Rate,

$$\text{FDR}(t) = \frac{\text{False}(t)}{\text{Total}(t)} = \frac{\hat{\pi}_0 \text{Null}(t)}{\text{Total}(t)}, \tag{5}$$

where $\hat{\pi}_0$ is the estimated proportion of true null hypotheses. Hence, the higher $\hat{\pi}_0$, the larger the proportion of genes *falsely* called under the null hypotheses. FDR(t) is calculated for several values of t. When FDR(t') is at the aimed level for $t = t'$, those genes for which $|r_{(i)} - \bar{r}_{(i)}| > t'$ are called significant.

SAM-RS differs from SAM by using a linear signed-rank statistic instead of the t-type statistic

$$d_i = \frac{\bar{Z}_i}{s_i + s_0},$$

where $\bar{Z}_i = \sum_{j=1}^{N} Z_{ij}/N$, $s_i = \{\sum_{j=1}^{N}(Z_{ij} - \bar{Z}_i)/(N(N-1))\}^{1/2}$ and s_0 is a fudge factor (see Chu et al. (2001) for its definition). The fudge factor is a positive number; adding it to s_i prevents d_i from becoming very large when s_i is very small. Substitution of r by d in the procedure above results in the SAM procedure.

Originally, SAM used upper bound $\hat{\pi}_0 = 1$. The latest SAM software (versions 1.10 and higher) defines

$$\hat{\pi}_0 = \min\left(1, \frac{\#\{r_i \in (q_{25}, q_{75})\}}{0.5p}\right), \tag{6}$$

where q_{25} and q_{75} are the 25% and 75% quantiles of the joint permutation distribution of the r_i's. The choice for these quartiles and corresponding coefficient of $p : (75-25)/100 = 0.5$ is rather arbitrary. Other pairs of quantiles $(q_{\lambda*100}, q_{(1-\lambda)*100})$ can be used. Storey and Tibshirani (2001) deal with finding the optimal value of λ with respect to a mean square error criterion. In order to adapt to the SAM software, we apply estimation (6) in SAM-RS too.

In principle, estimation of the proportion rejections under the null hypotheses does not have to be based on the median as in (3), but instead may also be based on the average:

$$\text{Null}'(t) = \frac{1}{B}\sum_{b=1}^{B} \#\{i : |r^b_{(i)} - \bar{r}_{(i)}| > t\}.$$

In SAM, use of median based Null(t) instead of Null$'(t)$ makes the estimation more robust against outliers in the data. Since the rank transformation in SAM-RS already results in robustness against outliers, we could use Null$'(t)$ instead of Null(t) in SAM-RS with the benefit of a more precise estimation. However, to stay within the framework of the SAM software, we also use (3) for SAM-RS.

2.2 Bias of SAM and unbiasedness of SAM-RS

With an example Pan et al. (2001) showed that SAM may result in biased estimates FDR(t). This is due to bias in estimations of both Null(t) and Total(t). When Z_{ij} and $Z_{i'j}$ originate from two different distributions F_0 and F_1, then the expected (with respect to F_0 and F_1) permutational distributions of d_i and $d_{i'}$ are not the same. Hence, because the distribution of $d_{(i)}$ depends on the joint distribution of all d_i's, we have, under equally likely +1, -1 permutations

$$E\big[d_{(i)}|\text{all } Z_{ij} \text{ from } F_0\big] \neq E\big[d_{(i)}|Z_{ij} \text{ from } F_1 \neq F_0 \text{ for some } i\big], \tag{7}$$

where expectation is computed first with respect to the distributions of the Z_{ij}'s and then with respect to the permutational distribution of the signs. Therefore, $d_{(i)}$'s for which the corresponding Z's originate from F_0 are compared with wrong average scores $\bar{d}_{(i)} = \sum_{b=1}^{B} d^b_{(i)}/B$, since for some b's the Z's corresponding to $d^b_{(i)}$ may not originate from F_0. Then, the number of $d_{(i)}$'s for which $|d_{(i)} - \bar{d}_{(i)}| > t$ and the median of the number of $d^b_{(i)}$'s for which $|d^b_{(i)} - \bar{d}_{(i)}| > t$ are too large. Therefore, inequality (7) affects both Null(t) and Total(t).

Since the permutational distribution of rank score statistic r_i does not depend on the distribution functions of the Z_{ij}'s, we have, under equally likely +1, -1 permutations

$$E\big[r_{(i)}|\text{all } Z_{ij} \text{ from } F_0\big] = E\big[r_{(i)}|Z_{ij} \text{ from } F_1 \neq F_0 \text{ for some } i\big]. \tag{8}$$

Hence, estimation of Null(t) is unbiased in SAM-RS. Moreover, comparing $r_{(i)}$ with $\bar{r}_{(i)}$ to obtain Total(t) is fair, because $\bar{r}_{(i)}$ does not depend on the distribution functions of the Z's.

2.3 Choice of rank score function

Definition (2) gives us the freedom to choose a suitable rank score function $a(u)$. Wilcoxon rank scores, i.e. $a(u) = u$, are the most commonly used rank scores in nonparametric testing. However, the resulting statistic might be too discrete. Normal rank score statistics are less discrete and slightly more powerful against normal alternatives. These scores are based on the inverse standard normal cumulative distribution function Φ^{-1}. The normal signed-rank scores are defined as:

$$a(u) = \Phi^{-1}\big(\frac{1}{2} + \frac{1}{2(N+1)}\big).$$

It is also possible to use locally most powerful rank statistics for an arbitrary null density f_0, as developed in (Hájek et al., 1999, sec. 3.4). For example, one might consider to estimate f_0 with a normal mixture as in Pan et al. (2001), and base the rank scores on this estimate. A disadvantage of this approach is the need to model f_0. Moreover, although testing with those rank scores is still distribution-free, the whole procedure is not truly distribution-free anymore, because the scores are based on $\hat{f}_0$.

3 The k-rule: A more selective procedure

Instead of testing $\mu_i = 0$, one might consider a more selective procedure by testing simultaneously

$$\begin{aligned} &H^-_{0i} : \mu_i - ks_i = 0 \;\text{ against }\; H^-_{1i} : \mu_i - ks_i > 0 \;\text{ and} \\ &H^+_{0i} : \mu_i + ks_i = 0 \;\text{ against }\; H^+_{1i} : \mu_i + ks_i < 0. \end{aligned} \tag{9}$$

This procedure tends to select genes with means well separated from zero in terms of the number of standard errors with higher probability than those of which the means are close to zero. Similar to (4), the rejection region consists of those genes for which

$$r^-_{(i)} - \bar{r}^-_{(i)} > t \quad \text{or} \quad r^+_{(i)} - \bar{r}^+_{(i)} < -t,$$

where $r^-_{(i)}$ ($\bar{r}^-_{(i)}$) and $r^+_{(i)}$ ($\bar{r}^+_{(i)}$) are defined as $r_{(i)}$ ($\bar{r}_{(i)}$) in section 2.1, but here these (average) signed-rank score sums are computed from data $Z_{ij} - ks_i$ and $Z_{ij} + ks_i$, $j = 1, \ldots, N$, respectively. This is a not an exact, but instead conditional distribution-free procedure: given s_i, we have, under the null hypotheses, equally probable sign permutations. Hence, equality (8) is only true when expectations under F_0 and F_1 are computed conditionally on the observed s_i's. Also, using the sample standard error might be slightly unfair for some genes, because subtracting s_i does not have the same consequences for genes with symmetrical distributions as for those with very skewed distributions. However, it is good to have a procedure that selects genes with small standard errors relatively more frequently than those with large standard errors, without making (further) assumptions on the underlying distribution functions. We refer to this procedure as $k*$SE SAM-RS, where 'SE' stands for standard error.

Alternatively, one could standardize gene expressions first and then test the null hypotheses above with $s_i = 1$ for all i. In fact, standardization has no effect on the rank scores within one gene. Hence, one can immediately test simultaneously

$$\begin{aligned} &H^-_{0i} : \mu_i - k = 0 \quad \text{against} \quad H^-_{1i} : \mu_i - k > 0 \quad \text{and} \\ &H^+_{0i} : \mu_i + k = 0 \quad \text{against} \quad H^+_{1i} : \mu_i + k < 0. \end{aligned} \tag{10}$$

Then, selection of the genes is solely done on the basis of deviation larger than k from zero. This procedure is exact distribution-free. We refer to it as k SAM-RS. When applied on $\log_2$-ratios of gene expressions, k SAM-RS for $k = 1$ is the testing analogue of the popular 2-fold rule, which is available in the SAM software. In Sam version 1.10 and higher one may define a threshold fold value that is used as an extra selection criterion *outside* the statistical test.

Use of $k*$SE SAM-RS and k SAM-RS within SAM software is somewhat more complicated than for SAM-RS, because the current SAM software does not do one-sided testing. Also, control of the FDR(t) has to be done simultaneously on the set of genes for which H^-_{0i} is rejected and on the set of genes for which H^+_{0i} is rejected. Finally, the formula for $\hat{\pi}_0$ as defined in (6) is incorrect, because it is based on two-sided testing of $\mu_i = 0$. However, let us shortly explain how one can still apply $k*$SE SAM-RS, and hence also k SAM-RS, within the SAM software.

First, one needs to create two data sets: D^- with ks_i subtracted from Z_{ij} and D^+ with ks_i added to Z_{ij}. Then, apply signed-rank statistic (2) to both data sets. Let $\text{False}^-(t), \text{Null}^-(t)$ and $\text{Total}^-(t)$ be the one-sided equivalents of $\text{False}(t), \text{Null}(t)$ and $\text{Total}(t)$, which are defined in section 2.1. Rejection criteria $r^{b-}_{(i)} - \bar{r}^-_{(i)} > t$ and $r^-_{(i)} - \bar{r}^-_{(i)} > t$ determine the number of positively called genes for the bth sign permutation under H^-_{0i}, $\text{Null}^-(t)$, and the total number of positively called genes for the data, $\text{Total}^-(t)$, respectively. When one uses a symmetric rank score function, we have in expectation, under the sign permutation distribution $\text{Null}^-(t) = \text{Null}(t)/2$, because for every permutation of the signs, there is one equally likely sign permutation for which the order of the rank statistics inverses and hence also the number of negatively and positively called genes inverses. SAM software does not return $\text{Null}(t)$, but it does return $\text{False}(t)$ and (incorrect) $\hat{\pi}_0$. From (5) we observe that $\text{Null}(t)$ is found by dividing $\text{False}(t)$ by $\hat{\pi}_0$.

Next, we have to compute the number of false rejections in favor of H_{1i}^-, $\text{False}^-(t) = \hat{\pi}_0^- \,\text{Null}^-(t)$, where $\hat{\pi}_0^-$ is interpreted as one minus the proportion true H_{1i}^-. Analogous to (6) we may estimate

$$\hat{\pi}_0^- = \min\left(1, \frac{\#\{r_i^- \in D^- : r_i^- \leq q_{75}\}}{0.75p}\right).$$

Unfortunately, this number can not be computed from the SAM software output. However, one might use $\hat{\pi}_0^- = 1$ as an upper bound. In many cases this will not be so bad, because the proportion true H_{1i}^- is often small due to the fact that $H_{1i}^- : \mu_i - ks_i > 0$ is a rather strong statement when $k > 0$.

For finding $\text{Total}^-(t)$ we use the number of positively called genes in D^-, which is returned by the SAM software. Then,

$$\text{FDR}(t) = \frac{\text{False}^-(t) + \text{False}^+(-t)}{\text{Total}^-(t) + \text{Total}^+(-t)},$$

where $\text{False}^+(-t)$ and $\text{Total}^+(-t)$ are obtained from the analysis of D^+. Repeating this procedure for several values of t enables one to control the $\text{FDR}(t)$ for multiple simultaneous testing of (9).

4 Example data sets

We discuss two example data sets: firstly, a simulated data set and secondly, the example one-class response data set provided in the SAM software. Like in the output of SAM, we report $\text{FDR}(t)$ in percentages.

The simulated data set consists of 500 gene expression data for 8 experimental objects. The first 400*8 values of Z_{ij} are drawn from standard normal distribution $F_0 = N(0, 1)$, whereas the last 100*8 values are drawn from Gamma distribution $F_1 = \Gamma(4, 1/8)$, which has mean 1/2. Hence, a proper procedure should find approximately 100 significantly expressed genes.

Suppose that the target FDR is 1%. Running SAM on the simulated data set results in a $\text{FDR}(t) = 1.20\%$ and $\text{False}(t) = 0.76$ when selecting 69 genes. SAM-RS finds 100 significant genes at $\text{FDR}(t) = 0.77\%$ and $\text{False}(t) = 0.77$. SAM returns $\text{FDR}(t) = 5.19\%$ and $\text{False}(t) = 5.29$ when selecting 102 genes. Hence, in this case, the skewness of F_1 leads to dramatic overestimation of $\text{False}(t)$ and $\text{FDR}(t)$ when using SAM.

Next, we compare the results of SAM and SAM-RS for the example one-class response data set in the SAM software. This data set consists of 1000 gene expression data for 8 experimental objects. For both SAM and SAM-RS we used 500 sign permutations to obtain $\text{False}(t)$ as outlined in section 2.1. We show that SAM-RS selects less genes than SAM, because it does not underestimate $\text{FDR}(t)$, but it finds the important ones. Using a target FDR of 1%, SAM selects 401 genes for $t = 0.533$, with $\text{FDR}(t) = 1.227\%$ and $\text{False}(t) = 4.920$. SAM-RS selects 360 genes for $t = 1.085$, with $\text{FDR}(t) = 1.017\%$ and $\text{False}(t) = 3.660$. Figure 4 demonstrates that SAM-RS, which returns 172 negatively expressed genes, finds all top genes among the 207 negatively expressed ones returned by SAM. This figure shows that SAM-RS does a good job with respect to SAM, because the 35 negatively expressed genes not returned by SAM-RS are the ones that have the least negative scores in SAM. We see a similar picture for the positively expressed genes. In this example, the bias in the estimation $\text{FDR}(t)$ by SAM is not large. SAM would

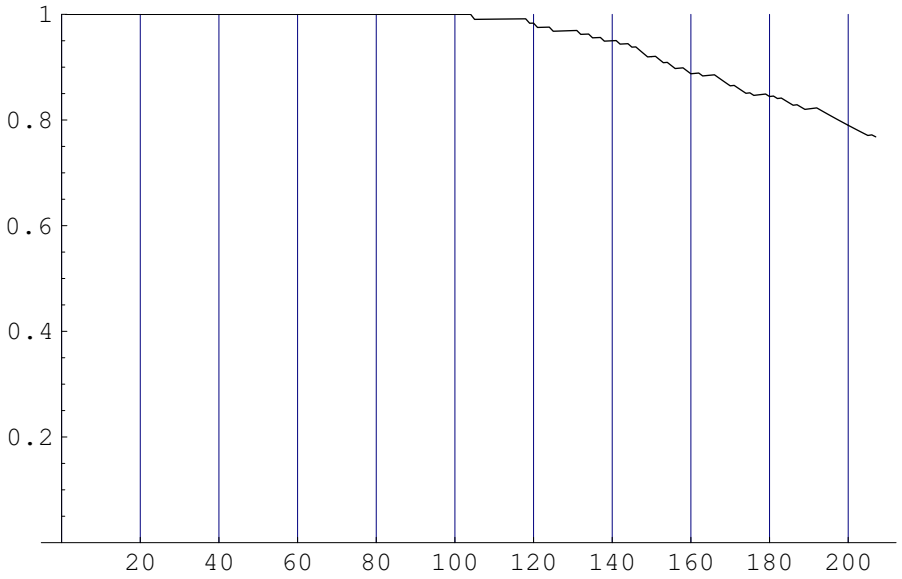

Figure 1: Number of most negatively significant genes called by SAM against the fraction of those genes also found by SAM-RS

select 363 genes, hence almost the same number as SAM-RS, for $\text{FDR}(t) = 0.806\%$ and $\text{False}(t) = 2.926$. Hence, the bias in the estimation of $\text{FDR}(t)$, with respect to the unbiased SAM-RS $\text{FDR}(t)$ is approximately 1.017-0.806=0.211, which corresponds to a relative underestimation of $(0.211/1.017) * 100\% = 21\%$.

We also applied k SAM-RS for $k = 1$ to SAM's example one-class data set. Using $t = 0.734$ in both D^- and D^+ and using upper bounds $\hat{\pi}_0^- = 1$ and $\hat{\pi}_0^+ = 1$, we found $\text{False}^-(t) \approx \text{Null}^-(t) = 1.824/(0.228 * 2) = 4$ and $\text{False}^+(t) \approx 1.984/(0.248 * 2) = 4$. This procedure finds 29 positive significant genes in D^- and 27 negative significant genes in D^+. Therefore, $\text{FDR}(0.734) \approx (4+4)/(27+29) = 14.3\%$. In this case, it is not possible to obtain a smaller value of $\text{FDR}(t)$ due to the discreteness of the test as discussed in section 2.3. Finally, we applied the threshold $k = 1$ after the testing procedure to those 360 genes selected by SAM-RS in the previous example. This procedure, which corresponds to the 2-fold rule when the data are $\log_2$ ratios, results in 160 genes called. The latter procedure is less selective than k SAM-RS, because $\log_2$ ratios of those genes called need not to be *significantly* larger than 1 or smaller than -1. Since the additional criterion is not included within the test, control of $\text{FDR}(t)$ is with respect to hypotheses (1). Hence, control of $\text{FDR}(t)$ with respect to the entire selection procedure is somewhat lost.

5 Other data structures

Let us shortly discuss some other data structures on which SAM-RS and the k-rules can be applied. All rank statistics are discussed and defined in Hollander and Wolfe (1999).

In case of unpaired two-class data one should use a two-sample linear rank statistic like the Wilcoxon rank-sum statistic. In a combined ordered sample, this is the sum of ranks corresponding to one of the two samples. Like for the one-class data it may be better to use a less discrete statistic like the Van der Waerden (normal scores) statistic which applies an inverse normal transformation to the ranks. Then, the joint null distribution

is obtained by permuting columns instead of signs. For class sample sizes n_1 and n_2 there are $\binom{n_1+n_2}{n_1}$ equiprobable permutations.

When the data are quantitative and one wants to test for correlation between response and the experimental conditions, Kendall's rank correlation coefficient or the less discrete Spearman's rank correlation coefficient can be used. In case of quantitative data, the experimental conditions correspond to a natural ranking, for instance ranking by time or temperature. Kendall's rank correlation coefficient is based on the number of inversions, which is the number of pairs $(i, j) : i < j < N$ for which response Y_i is smaller than response Y_j, whereas experimental condition X_i is larger than condition X_j. Spearman's rank correlation coefficient is based on the sum of squares of differences between the ranks of the experimental conditions and the corresponding ranks of the responses. Then for both rank correlation coefficients we have, under the null hypothesis: 'no rank correlation', that all $N!$ permutations of the columns are equally likely.

For randomly censored two-class survival data, the Log-rank (or Mantel) statistic is the most common option. This conditional test compares the number of failures of one of the two samples with the expected number of failures under the null hypothesis given the total number of failures within the combined sample at each failure time. In this case a slight problem may occur: rank scores will not be the same for every gene, because these are computed conditionally on the censoring structure within each gene. If the censoring distributions are the same for each gene, then equality 8 is still true and the FDR(t) computation stays unbiased. If not so, FDR(t) is (slightly) biased.

6 Discussion

We have observed that SAM-RS is a useful alternative to SAM. Due to its distribution-freeness it returns unbiased estimates of the FDR and it is robust against outliers in the data. Moreover, it fits nicely in the SAM software; one only has to transform the data to rank scores. When N is too small, SAM-RS might be too discrete and not powerful enough. Exact definition of 'too small' depends on the data structure, on the test statistic and on the proportion truly expressed genes in the data. Distributions of signed-rank and rank correlation statistics are often less discrete than those of two-sample rank statistics because of the larger numbers of permutations, 2^N and $N!$ versus $\binom{N}{n_1}$. We recommend to use normal scores based rank statistics instead of Wilcoxon-type statistics, because their distributions are less discrete. We observed that $N = 8$ was sufficiently large to detect top genes with SAM-RS in the example one-class data set provided by the SAM software.

The new procedures k SAM-RS and $k*$SE SAM-RS include more stringent selection criteria in the test. These procedures result in sets of genes with mean expressions significantly larger than k or $k*$SE. Using a threshold k or $k*$SE outside the test, which corresponds to k-fold rules for logarithmic data, results in sets of genes with mean expressions that are larger than this threshold and significantly larger than zero. Hence, application of k SAM-RS or $k*$SE SAM-RS instead of the additional threshold procedure gives the scientist the opportunity to make a stronger statement about the selected genes. Needless to say, the k and $k*$SE criteria can also be incorporated into 'normal' SAM.

Good news is that SAM-RS can easily be carried out by using the SAM software. All that is required is preprocessing of the data by transformation of the Z_{ij}'s to signed-rank scores. The author has written a Mathematica notebook that deals with this transformation and with exchanging the gene expressing data with Excel, which is the platform for the SAM software. This notebook is freely available from the author's web-site: `http://www.win.tue.nl/~markvdw`.

7 Acknowledgements

This study was carried out at the Norwegian Computing Center in Oslo under support of the EU-TMR programme on Spatial and Computational Statistics (grant number ERB-FMRX-CT960095). The author thanks Arnoldo Frigessi and Ingrid Glad for interesting discussions in the field of microarray analysis.

References

Benjamini, Y., and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc., B* **57**, 289–300.

Chu, G., B. Narasimhan, R. Tibshirani, and V. Tusher (2001). Sam, significance analysis of microarrays, users guide and technical document. Technical report Stanford University, http://www-stat.stanford.edu/~tibs/SAM.

Dudoit, S., Y. H Yang, T. P. Speed, and M. J. Callow (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578 University of California, Berkeley, http://www.stat.berkeley.edu/tech-reports/index.html.

Efron, B., R. Tibshirani, V. Goss, and G. Chu (2000). Microarrays and their use in a comparative experiment. Technical Report 2000-37B/213 Stanford University, http://www-stat.stanford.edu/~tibs/research.html.

Efron, B., R. Tibshirani, J.D. Storey, and V. Tusher (2001a). Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96**, 1151–1160.

Efron, B., J.D. Storey, and R. Tibshirani (2001b). Microarrays, empirical Bayes methods and false discovery rates. Technical Report 2001-23B/217 Stanford University, http://www-stat.stanford.edu/~tibs/research.html.

Hájek, J., Z. Šidák, and P.K. Sen (1999). *The Theory of Rank Tests, 2nd ed.* Academic Press, London.

Hollander, M., and D.A. Wolfe (1999). *Nonparametric statistical methods, 2nd ed.* Wiley, New York.

Klotz, J.H. (1963). Small sample power and efficiency for the one-sample Wilcoxon and normal scores test. *Ann. Math. Statist.* **34**, 624–632.

Pan, W., J. Lin, and C.T. Le (2001). A mixture model approach to detecting differentially expressed genes with microarray data. Technical Report 2001-11 Department of Biostatistics, University of Minnesota, http://www.biostat.umn.edu/cgi-bin/rrs?print+2001.

Sørlie, T., C.M. Perou, and R. et al. Tibshirani (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98**, 10869–10874.

Storey, J.D., and R. Tibshirani (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical Report 2001-12 Stanford University, http://www-stat.stanford.edu/~jstorey/.

Tusher, V.G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98**, 5116–5121.

Wolfinger, R.D., G. Gibson, and E.D. et al. Wolfinger (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comp. Biol.* **8**, 625–637.