Nonparametric Approaches to Detecting Differentially Expressed Genes in Replicated Microarray Experiments

Markus Neuhäuser and Fred C. Lam

Department of Mathematics and Statistics
University of Otago
PO Box 56, Dunedin, New Zealand

`mneuhaus@maths.otago.ac.nz, flam@maths.otago.ac.nz`

Abstract

Microarrays have quickly been established as an efficient tool for gene expression profiling. In this paper we consider the detection of differentially expressed genes with replicated measurements of expression levels of each gene under each condition. Several authors mentioned that data from microarrays are often not normally distributed, even when suitably preprocessed. Consequently, nonparametric tests, such as the Wilcoxon rank sum test and the Fisher-Pitman permutation test, were recommended. As a further powerful nonparametric test we propose the Baumgartner-Weiß-Schindler (BWS) test. However, when the population variances are unequal a significant result in these tests does not necessarily provide evidence for a difference in location. Note that, in data from microarray experiments, heterogeneous variances are common. Therefore, we suggest a two-stage procedure. If the BWS test applied in stage 1 is significant, a test for a difference in location only is carried out in stage 2. A bootstrap test based on the Welch t statistic can be used in stage 2. However, we demonstrate that a rank-based test recently proposed by Brunner and Munzel (2000) is more powerful.

Keywords: microarray, nonparametric tests, heteroscedasticity.

1 Introduction

DNA microarray technologies, such as cDNA arrays and oligonucleotide arrays, can be used to measure the expression of thousands of genes simultaneously. These technologies are rapidly becoming common laboratory tools and promise to revolutionize biological research. They are used in biomedical research, but also in other areas such as ecology and evolution (Gibson 2002). Often, the question is whether gene expression is different for two (or sometimes more) groups of organisms that differ with respect to a characteristic such as exposure to some environmental stimuli, genotype, or age (Gadbury et al. 2003). In this paper we consider the comparison of two groups in order to detect differentially expressed genes based on replicated measurements of expression levels of each gene.

From now on, the expression levels can refer to a summary measure of relative red to green channel intensity, a radioactive intensity, or a summary difference of the perfect match and mis-match scores; furthermore, the gene expression levels may have been preprocessed using dimension reduction, normalization and data transformation (Pan 2002).

Several authors pointed out that expression data from microarrys are often not distributed according to a normal distribution, even after some preprocessing (Hunter et al. 2001, Thomas et al. 2001, Pan 2002, Craig et al. 2003, Zhao and Pan 2003). According to Thomas et al. (2001) the normality assumption is certainly inappropriate for a subset of genes despite any given transformation. Therefore, nonparametric tests were recommended for the analysis of microarrays (Troyanskaya et al. 2002, Gadbury et al. 2003, Xu and Li 2003). The advantage of nonparametric methods is that no specific distribution has to be assumed. For example, the Wilcoxon rank sum test (equivalent to the Mann-Whitney U test) and the Fisher-Pitman permutation test were recommended (Troyanskaya et al. 2002). Regarding details about these two standard tests we refer to Manly (1997) and Hollander and Wolfe (1999).

The Fisher-Pitman permutation test is also called nonparametric t test because the t test statistic can be used, but its distribution is determined using permuted data sets. However, as the sample sizes, i.e. the numbers of replications, are usually very small in microarrays, the Wilcoxon test should also be carried out as a permutation test. The alternative is to rely on the asymptotic normality of the rank sum which is appropriate when the sample size exceeds eight in each group (Troyanskaya et al. 2003).

In a permutation test, all possible permutations under the null hypothesis are generated and the test statistic is calculated for each permutation. The null hypothesis can then be accepted or rejected using the permutation distribution of the test statistic, the p-value being the probability of the permutations giving a value of the test statistic as supportive or more supportive of the alternative than the observed value. Thus inference is based upon how extreme the observed test statistic is relative to other values that could have been obtained under the null hypothesis.

In this paper, we compare the Wilcoxon rank sum test and the Fisher-Pitman permutation test, hereafter WRS and FPP test, respectively, with the Baumgartner-Weiß-Schindler (BWS) test, a more novel test. In addition, we investigate some useful alternative tests for possibly unequal variances.

2 The BWS test

In 1998, Baumgartner et al. (1998) proposed a novel nonparametric test based on ranks. The proposed test statistic is $B = \frac{1}{2} \cdot (B_X + B_Y)$, where

$$B_X = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\left(R_i - \frac{N}{n_1} \cdot i\right)^2}{\frac{i}{n_1+1}\left(1 - \frac{i}{n_1+1}\right) \cdot \frac{n_2 N}{n_1}} \ ,$$

$$B_Y = \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{\left(H_j - \frac{N}{n_2} \cdot j\right)^2}{\frac{j}{n_2+1}\left(1 - \frac{j}{n_2+1}\right) \cdot \frac{n_1 N}{n_2}} \ ,$$

and $R_1 < ... < R_{n_1}$ ($H_1 < ... < H_{n_2}$) denote the combined-samples ranks of the n_1 (n_2) values from the first (second) group in increasing order of magnitude; $N = n_1 + n_2$ is the total sample size. Large values of B support the alternative that there are differences between the two groups. Baumgartner et al. (1998) determined the asymptotic distribution of B and demonstrated via simulation that the resulting asymptotic test is at least as powerful as commonly-used nonparametric tests, such as the WRS test.

The test based on B can also be carried out based on the exact permutation distribution of B. This exact test is recommended for very small sample sizes (Neuhäuser 2000). When comparing the two exact tests based on ranks (WRS and BWS), the BWS test is not only preferable in terms of power. A further advantage is that the exact permutation distribution of B is less discrete than that of the rank sum. For instance, there are $\binom{2 \cdot 10}{10}$

= 184,756 possible permutations in case of ten replications per group. Within these permutations the statistic B has 11,833 different values whereas the rank sum has only 101. The consequence is that (1) the BWS test is less conservative than the WRS test and (2) smaller p-values are possible. The latter point is particularly important in microarrays since many genes are considered simultaneously and, hence, the significance level may be adjusted for multiple testing.

In the presence of ties (observations with identical values) the usual way of dealing with these values is to assign average ranks. That is, we give tied observations the average of the ranks for which those observations are competing. Unfortunately, the asymptotic BWS test can

have an inflated type I error rate in this case (Neuhäuser 2002). However, all the exact permutation tests can be applied whether or not ties occur.

3 Heterogeneous variances

In gene expression heteroscedasticity is common, i.e. the variances within the two groups can be different (Thomas et al. 2001, Craig et al. 2003, Pepe et al. 2003). In a spotted microarray example Craig et al. (2003) found a larger variability for low intensity observations. The three above-mentioned tests can give a significant result for a test at the 5% level with much more than 0.05 probability when the population means are identical, but the population variances differ (see e.g. Hayes 2000, Neuhäuser 2000, Kasuya 2001).

In comparison to the tests FPP and WRS, the BWS test is more sensitive to different variances. Table 1 shows that the probability to get a significant result solely due to heteroscedasticity is larger for the BWS test. Note that the tests FPP and WRS have larger probabilities for significances when sample sizes are unequal, to be precise, when the smaller group has the larger variance.

However, irrespective of the sample sizes, a difference in variances can cause a significance. Therefore, a significant FPP, WRS, or BWS test does not necessarily provide evidence for a difference in means.

Standard deviations	Test		
σ_1, σ_2	FPP	WRS	BWS
1, 1	0.05	0.04	0.05
1, 2	0.06	0.05	0.09
1, 3	0.06	0.06	0.17
1, 4	0.07	0.07	0.26

Table 1: Proportions of significant results of permutation tests, based on simulated data from the normal distributions N(0, σ_1) and N(0, σ_2) [sample size per group: 10, α = 0.05, 10,000 simulation runs]

Therefore, when variances may be different, the tests, in particular the BWS test, are useful to test for any difference between the groups. Hence, the null hypothesis H_0 here is that the data in both groups follow an identical distribution. Let F and G denote the two distribution functions, we have H_0: $F = G$. Under the alternative H_1, there is a difference. Thus, H_1: $F(t) \neq G(t)$ for at least one t.

4 The nonparametric Behrens-Fisher problem

If H_0: $F = G$ were rejected, a researcher usually would like to know more than just that there is some difference between the two groups. Often, one is interested in whether values of one group tend to be larger than those of the other group.

For normally distributed data, to test the null hypothesis that the means, but not necessarily the variances, are equal is called the Behrens-Fisher problem. The null hypothesis tested in the generalized (nonparametric) Behrens-Fisher problem is $H_0^{BF}: p = \frac{1}{2}$ where the relative effect p is defined as

$$p = \Pr(X_i < Y_j) + \tfrac{1}{2}\Pr(X_i = Y_j)$$

with $X_i \sim F$ from group 1 and $Y_j \sim G$ from group 2 (Brunner and Munzel 2002, p. 53). When the distribution functions F and G are continuous, ties are not possible. In that case the relative effect can be calculated as $p = \Pr(X_i < Y_j)$. In the special case of two independent normal distributions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 we have

$$p = \Pr(X_i < Y_j) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$$

where Φ is the distribution function of the standard normal distribution (Reiser and Guttman, 1986). Since $\Phi(t) = \frac{1}{2}$ if and only if $t = 0$, $\mu_1 = \mu_2$ is equivalent to $p = \frac{1}{2}$, but the standard deviations variances σ_1 and σ_2 can differ. As a result, the generalized Behrens-Fisher problem is a special case of the classical one.

The observations in group 1 tend to be smaller in comparison to those of group 2 if $p > \frac{1}{2}$. In the case of $p < \frac{1}{2}$ the observations in group 1 tend to be larger than those of group 2.

Note that the hypothesis $F = G$ is a subset of H_0^{BF} (Brunner and Munzel, 2002, p. 234). Therefore, it is useful to test H_0^{BF} in a second step after rejection of the null hypothesis $F = G$ in a first stage. For the first step, we recommend the BWS test. But which test is appropriate for step 2?

A modification of the t test that can be used in the Behrens-Fisher problem was introduced by Welch (1947). This test has been applied for the statistical analysis of data from microarrays (Dudoit et al. 2002, Pan 2002). The test can be carried out as a nonparametric test, i.e. without the assumption of normally distributed data, in a bootstrap procedure. A permutation test is not possible since the distributions can be different under the null hypothesis H_0^{BF} (Efron and Tibshirani 1993, pp. 223-224, Brunner and Munzel 2002, p. 75).

The t statistic, modified for the Behrens-Fisher problem, is

$$T_{BF} = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

with $S_i^2 = \dfrac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2$, where the X_{ij}, $i = 1, 2$, $j = 1, ..., n_i$, are the observed values, and the $\overline{X}_i$'s are the means of the two groups. The bootstrap procedure is as follows (Efron and Tibshirani 1993, pp. 222-224):

(a) Calculate the statistic T_{BF} for the observed values.

(b) Transform the observed values so that both groups have the same mean: $\dot{X}_{ij} = X_{ij} - \overline{X}_i$, $i = 1, 2, j = 1, ..., n_i$.

(c) Draw samples of size n_1 and n_2 with replacement from the $\dot{X}_{ij}$ values and calculate the statistic T_{BF} for that bootstrap sample.

(d) Repeat step (c) M times.

The p-value is the proportion of bootstrap samples with a value of T_{BF} at least as large as T_{BF} with the original values. The number M should be large, Rózsa et al. (2000) recommend $M \geq 1{,}000$.

Brunner and Munzel (2000) recommended a rank-based test for the nonparametric Behrens-Fisher test. The test statistic is

$$W_{BF} = \sqrt{\frac{n_1 n_2}{N}} \cdot \frac{\overline{R}_{2.} - \overline{R}_{1.}}{\hat{\sigma}_{BF}},$$

where $\overline{R}_{i.}$ is the mean rank in group i and

$$\hat{\sigma}_{BF}^2 = \sum_{i=1}^{2}\frac{N\widetilde{S}_i^2}{N - n_i} \quad \text{with}$$

$$\widetilde{S}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(R_{ij} - R_{ij}^{(i)} - \overline{R}_{i.} + \frac{n_i + 1}{2}\right)^2.$$

Furthermore, $R_{ij}^{(i)}$ is the (within) rank of X_{ij}, i.e. the rank among the n_i obervations within group i. Note that

$$\hat{p} = \frac{1}{N}\left(\overline{R}_{2.} - \overline{R}_{1.}\right) + \frac{1}{2}$$

is an unbiased and consistent estimator of p.

The statistic W_{BF} is asymptotically standard normal (Brunner and Munzel, 2000). However, similar to the application of the Welch t test, the distribution should be approximated using a t distribution for small sample sizes. To be precise, one can use the t distribution with

$$df = \frac{\left(\sum\limits_{i=1}^{2}\dfrac{\widetilde{S}_i^2}{N - n_i}\right)^2}{\sum\limits_{i=1}^{2}\dfrac{\left(\widetilde{S}_i^2/(N - n_i)\right)^2}{n_i - 1}}.$$

5   Simulation study

We compared the bootstrap t test and the W_{BF} test in a simulation study performed using SAS version 8.2; 10,000 simulation runs were generated for each particular configuration. Both tests control the type I error rate, although sometimes an acual size marginally larger than 0.05 was found. For the W_{BF} test this is consistent with Brunner and Munzel's (2000) results: they found type I error rates between 0.046 and 0.057 in the case of a nominal significance level $\alpha = 0.05$.

The comparsion in terms of power is shown in Table 2. After simulating the data and transforming the distributions to have a median of 0, the values in one group were multiplied with MF and shifted by the amount θ. Consequently, in case of MF = 1 there is a difference in location only. For MF = 2 there is an additional difference in variability. The contaminated normal distribution is a mixture of normals defined as follows: The data are standard normal with probability 0.7, and with probability 0.3 they are normally distributed with mean 5 and standard deviation 4. Note that Xu and Li (2003) also investigated a contaminated normal distribution. Furthermore, the figure 1 in Pepe et al. (2003, p. 134) indicates that such a distribution is appropriate for gene expression data from microarray experiments.

Distribution	θ	MF = 1		MF = 2	
		W_{BF}	T_{BF}	W_{BF}	T_{BF}
Uniform on (0,1)	0.4	0.80	0.83	0.31	0.39
Standard normal	1.5	0.88	0.88	0.48	0.48
t with df = 3	2	0.85	0.72	0.53	0.40
χ^2 with df = 3	4	0.93	0.88	0.85	0.66
Exponential ($\lambda = 1$)	1.5	0.90	0.83	0.87	0.61
Contaminat. normal	4	0.78	0.68	0.73	0.40

Table 2: Simulated power of the W_{BF} test and the bootstrap t test based on T_{BF} for different distributions [sample size per group: 10, α = 0.05, 10,000 simulation runs, MF: multiplication factor, see text]

When the data come from normal distributions there is no difference in power. For most other distributions the W_{BF} test is the more powerful one. The bootstrap t test outperformed the W_{BF} test for the uniform distribution only. However, in that case the difference in power between the tests is much smaller than that for the other distributions investigated in our simulation study. Consequently, the results indicate that the W_{BF} test is a better choice.

6   Conclusion

Our approach for the identification of differentially expressed genes is to consider a univariate testing problem for each gene. A correction for the multiplicity of genes is a following step that is, as the previous step of normalizing the data, outside the scope of this paper. A common approach to this problem is to consider a procedure for testing the genes simultaneously for differential expression with the test on an individual gene being implied in the simultaneous test. Such a procedure was recommended, for example, by Zaykin et al. (2002) and Storey and Tibshirani (2003).

The Baumgartner-Weiß-Schindler test is useful for testing whether there is a difference between the distributions of the two groups. We recommend to apply this test. If a

significant difference is found, a test for the nonparametric Behrens-Fisher problem can then be applied in a following step. A powerful test for this second stage is the W_{BF} test. These two tests can be applied for any gene. Note that no multiplicity adjustment is necessary within genes since the W_{BF} test is carried out only if the first stage's test is significant (see e.g. Bauer 1991).

The advantage of the two-step procedure is that the first step can identify differences other than simple location shifts. The second step answers the question whether there is evidence that a difference found in stage 1 is, at least partly, due to location differences.

The tests we recommend are based on ranks. These tests have the advantage that they are not sensitive to outliers. Note that outliers frequently occur in microarray experiments (Lönnstedt and Speed, 2000).

Only two-sided alternative hypotheses are considered here; one-sided alternatives can be handled in a similar manner. Owing to the squares in the numerators of B_X and B_Y, the statistic B is not suitable for a one-sided test, but a modification with absolute values instead of squares was proposed for one-sided test problems (Neuhäuser 2001). Note that the statistic B considered here is the one presented by Baumgartner et al. (1998). A novel statistic recently introduced for replicated microarray data by Lönnstedt and Speed (2002) is also called B, but it is a different statistic.

7   References

Bauer, P. (1991): Multiple testing in clinical trials. *Statistics in Medicine* **10**: 871-890.

Baumgartner, W., Weiß, P. and Schindler, H. (1998): A nonparametric test for the general two-sample problem. *Biometrics* **54**: 1129-1135.

Brunner, E. and Munzel, U. (2000): The nonparametric Behrens-Fisher problem: asymptotic theory and a small sample approximation. *Biometrical Journal* **42**: 17-25.

Brunner, E. and Munzel, U. (2002): *Nichtparametrische Datenanalyse*. Berlin, Springer.

Craig, B.A., Black, M.A. and Doerge, R.W. (2003): Gene expression data: the technology and statistical analysis. *Journal of Agricultural, Biological, and Environmental Statistics* **8**: 1-28.

Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002): Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**: 111-139.

Efron, B. and Tibshirani, R.J. (1993): *An introduction to the bootstrap*. New York, Chapman and Hall.

Gadbury, G.L., Page, G.P., Heo, M., Mountz, J.D. and Allison, D.B. (2003): Randomization tests for small samples: an application for genetic expression data. *Applied Statistics* **52**: 365-376.

Gibson, G. (2002): Microarrays in ecology and evolution: a preview. *Molecular Ecology* **11**: 17-24.

Hayes, A.F. (2000): Randomization tests and the equality of variance assumption when comparing group means. *Animal Behaviour* **59**: 653-656.

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

Hunter, L., Taylor, R.C., Leach, S.M. and Simon, R. (2001): GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics* **17** (Suppl. 1): S115-S122.

Kasuya, E. (2001): Mann-Whitney *U* test when variances are unequal. *Animal Behaviour* **61**: 1247-1249.

Lönnstedt, I. and Speed, T. (2002): Replicated microarray data. *Statistica Sinica* **12**: 31-46.

Manly, B.F.J. (1997): *Randomization, bootstrap and Monte Carlo methods in biology*. London, Chapman and Hall (2nd edition).

Neuhäuser, M. (2000): An exact two-sample test based on the Baumgartner-Weiß-Schindler statistic and a modification of Lepage's test. *Communications in Statistics - Theory and Methods* **29**: 67-78.

Neuhäuser, M. (2001): One-sided two-sample and trend tests based on a modified Baumgartner-Weiß-Schindler statistic. *Journal of Nonparametric Statistics* **13**: 729-739.

Neuhäuser, M. (2002): The Baumgartner-Weiß-Schindler test in the presence of ties (letter to the editor). *Biometrics* **58**: 250.

Pan, W. (2002): A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**: 546-554.

Pepe, M.S., Longton, G., Anderson, G.L. and Schummer, M. (2003): Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**: 133-142.

Reiser, B. and Guttman, I. (1986): Statistical inference for Pr($Y < X$): the normal case. *Technometrics* **28**: 253-257.

Rózsa, L., Reiczigel, J. and Majoros, G. (2000): Quantifying parasites in samples of hosts. *Journal of Parasitology* **86**: 228-232.

Storey, J.D. and Tibshirani, R. (2003): Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA* **100**: 9440-9445.

Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001): An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* **11**: 1227-1236.

Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D. and Altman, R.B. (2002): Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18**: 1454-1461.

Welch, B.L. (1937): The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**: 350-362.

Xu, R. and Li, X. (2003): A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics* **19**: 1284-1289.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H. and Weir, B.S. (2002): Truncated product method for combining *P*-values. *Genetic Epidemiology* **22**: 170-185.

Zhao, Y. and Pan, W. (2003): Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics* **19**: 1046-1054.