International Journal of Neural Systems, Vol. 16, No. 5 (2006) 353–362
© World Scientific Publishing Company

MIXTURE MODELS FOR DETECTING DIFFERENTIALLY EXPRESSED GENES IN MICROARRAYS

LIAT BEN-TOVIM JONES[*] and RICHARD BEAN[†]
ARC Centre in Bioinformatics, Institute for Molecular Bioscience (IMB),
University of Queensland, St. Lucia
Brisbane, 4072, Australia
[*]*liatj@maths.uq.edu.au*
[†]*rbean@maths.uq.edu.au*

GEOFFREY J. MCLACHLAN
Department of Mathematics and ARC Centre in Bioinformatics
and ARC Special Research Centre for Functional and Applied Genomics, IMB,
University of Queensland, St. Lucia
Brisbane, 4072, Australia
gjm@maths.uq.edu.au

JUSTIN XI ZHU
ARC Centre in Bioinformatics, IMB,
University of Queensland, St. Lucia
Brisbane, 4072, Australia
j.zhu@imb.uq.edu.au

An important and common problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes. As this problem concerns the selection of significant genes from a large pool of candidate genes, it needs to be carried out within the framework of multiple hypothesis testing. In this paper, we focus on the use of mixture models to handle the multiplicity issue. With this approach, a measure of the local FDR (false discovery rate) is provided for each gene. An attractive feature of the mixture model approach is that it provides a framework for the estimation of the prior probability that a gene is not differentially expressed, and this probability can subsequently be used in forming a decision rule. The rule can also be formed to take the false negative rate into account. We apply this approach to a well-known publicly available data set on breast cancer, and discuss our findings with reference to other approaches.

1. Introduction

DNA microarrays allow the simultaneous measurement of the expression levels of tens of thousands of genes for a single biological sample; see, for example, McLachlan *et al.*[1] Here the term expression level of a gene refers to the concentration of its corresponding bound mRNA as measured by the fluorescence intensity in the microarray experiment. A major objective in these experiments is to find genes that are differentially expressed in a given number of classes. In cancer studies, the classes may correspond to normal versus tumor tissues, or to different subtypes of a particular cancer. Comparing gene expression profiles across these classes gives insight into the roles of these genes, and is important in making new biological discoveries. Yet now a real goal for microarrays is to establish their use as tools in medicine. This requires the identification of subsets of genes (marker genes) potentially useful in cancer diagnosis and prognosis.

In the early days of microarray technology, a simple fold change test with an arbitrary cut-off value was used to determine differentially expressed genes. This method is now known to be unreliable as it does not take into account the statistical variability. In order to determine statistical significance, a test such as the t-test, can be performed for each gene. However, when many hypotheses are tested the probability of a type I error (false positive) occurring increases sharply with the number of hypotheses. This multiplicity poses a considerable problem in microarray data, where there are many thousands of gene expression values.

It is also clear that single genes do not act independently, rather groups of genes involved in a particular biological pathway may be under similar control (co-regulated genes). In addition there is often dependency in the measurement errors in microarray experiments. Both these factors contribute to correlation between the test statistics.

Recently, a number of sophisticated statistical methods have been proposed, including several nonparametric methods. Tusher *et al.*[2] in their significance analysis method (SAM), proposed a refinement on the standard Student's t-statistic. Because of the large number of genes in microarray experiments, there will always be some genes with a very small sum of squares across replicates, so that their (absolute) t-values will be very large whether or not their averages are large. The modified t-statistic of Tusher *et al.*[2] avoids this problem. Pan *et al.*[3] also considered a nonparametric approach in their mixture model method (MMM). These methods are reviewed in Pan.[4]

In this paper, we initially present the statistical problem and show how a prediction rule based on a two-component mixture model can be applied. In particular, we show how the mixture model approach can handle the multiplicity issue. It provides a measure of the local FDR (false discovery rate), but can be used in the spirit of the q-value. In the latter case, an upper bound, c_o, can be obtained on the posterior probability of nondifferential expression, to ensure that the FDR is bounded at some desired level α.

We apply this method to real data, in the well-known breast cancer study of Hedenfalk *et al.*[5] with the aim of identifying new genes which are differentially expressed between BRCA1 and BRCA2 tumors. We compare our findings with those of Storey and Tibshirani,[6] and of Broët *et al.*[7] who also analysed this data set using different approaches.

To address the issue of co-regulated genes and dependency between the test statistics we consider the simulation experiment as in Allison *et al.*[8]

2. Two-Component Mixture Model Frame-Work

2.1. *Definition of model*

We focus on a decision-theoretic approach to the problem of finding genes that are differentially expressed. We use a prediction rule approach based on a two-component mixture model as formulated in Lee *et al.*[9] and Efron *et al.*[10] We let G denote the population of genes under consideration. It can be decomposed into G_0 and G_1, where G_0 is the population of genes that are not differentially expressed, and G_1 is the complement of G_0; that is, G_1 contains the genes that are differentially expressed.

We let the random variable Z_{ij} be defined to be one or zero according as the jth gene belongs to G_i or not ($i = 0, 1$; $j = 1, \ldots, N$). We define H_j to be zero or one according as to whether the null hypothesis of no differential expression does or does not hold for the jth gene. Thus Z_{1j} is zero or one according as to whether H_j is zero or one.

The prior probability that the jth gene belongs to G_0 is assumed to be π_0 for all j. That is, $\pi_0 = \mathrm{pr}\{H_j = 0\}$ and $\pi_1 = \mathrm{pr}\{H_j = 1\}$. Assuming that the test statistics W_j all have the same distribution in G_i, we let $f_i(w_j)$ denote the density of W_j in G_i ($i = 0, 1$). The unconditional density $f(w_j)$ of W_j is given by the two-component mixture model

$$f(w_j) = \pi_0\, f_0(w_j) + \pi_1\, f_1(w_j). \qquad (1)$$

Using Bayes Theorem, the posterior probability that the jth gene is not differentially expressed (that is, belongs to G_0) is given by

$$\tau_0(w_j) = \pi_0 f_0(w_j)/f(w_j) \quad (j = 1, \ldots, N). \qquad (2)$$

In this framework, the gene-specific posterior probabilities $\tau_0(w_j)$ provide the basis for optimal statistical inference about differential expression.

2.2. *Bayes decision rule*

Let e_{01} and e_{10} denote the two errors when a rule is used to assign a gene to either G_0 or G_1, where

e_{ij} is the probability that a gene from G_i is assigned to G_j $(i, j = 0, 1)$. That is, e_{01} is the probability of a false positive and e_{10} is the probability of a false negative. Then the risk is given by

$$\text{Risk} = (1 - c)\pi_0 e_{01} + c\pi_1 e_{10}, \tag{3}$$

where $(1 - c)$ is the cost of a false positive. As the risk depends only on the ratio of the costs of misallocation, they have been scaled to add to one without loss of generality.

The Bayes rule, which is the rule that minimizes the risk (3), assigns a gene to G_1 if

$$\tau_0(w_j) \leq c; \tag{4}$$

otherwise, the jth gene is assigned to G_0. In the case of equal costs of misallocation $(c = 0.5)$, the cutoff point for the posterior probability $\tau_0(w_j)$ in (4) reduces to 0.5.

2.3. *The FDR and FNR*

When many hypotheses are tested, the probability that a type I error (false positive) is made increases rapidly with the number of hypotheses. The Bonferroni method is perhaps the best known method for dealing with this problem. It controls the family-wise error rate (FWER), which is the probability that at least one false positive error will be made. Control of the FWER is useful for situations where the aim is to identify a small number of genes that are truly differentially expressed. However, in the case of exploratory type microarray analyses, approaches to control the FWER are too strict and will lead to missed findings. Here it is more appropriate to emphasize the proportion of false positives among the identified differentially expressed genes. The false discovery rate (FDR), introduced by Benjamini and Hochberg,[11] is essentially the expectation of this proportion and is widely used for microarray analyses. Similarly, the false nondiscovery rate (FNR) can be defined as the expected proportion of false negatives among the genes identified as not differentially expressed (Genovese and Wasserman[12]). We are unable to estimate the various error rates using cross-validation, as the class of origin of each observation (gene) is unknown; that is, we do not know whether a gene is differentially expressed or not. Thus we have to estimate these error rates using methods developed for unclassified data in terms of their posterior probabilities of class membership, as discussed in McLachlan[13] (Sec. 10.5.2).

2.4. *Estimated FDR*

In practice, we do not know π_0 nor the density $f(w_j)$, and perhaps not $f_0(w_j)$. In some instances, the latter may be known as we may have chosen our test statistic so that its null distribution is known (or known to a good approximation). For example, we shall work with the oneway analysis of variance F-statistic, which can be so transformed that its null distribution is approximately the standard normal.

Alternatively, null replications of the test statistic might be created, for example, by the bootstrap or permutation methods. We shall estimate the population density $f(w)$ by maximum likelihood after its formulation using a mixture model. But it can be estimated also nonparametrically by its empirical distribution based on the observed test statistics w_j.

If $\hat{\pi}_0, \hat{f}_0(w_j)$, and $\hat{f}(w_j)$ denote estimates of $\pi_0, f_0(w_j)$, and $f(w_j)$, respectively, the gene-specific summaries of differential expression can be expressed in terms of the estimated posterior probabilities $\hat{\tau}_0(w_j)$, where

$$\hat{\tau}_0(w_j) = \hat{\pi}_0 \hat{f}_0(w_j)/\hat{f}(w_j) \quad (j = 1, \ldots, N) \tag{5}$$

is the estimated posterior probability that the jth gene is not differentially expressed. An optimal ranking of the genes can therefore be obtained by ranking the genes according to the $\hat{\tau}_0(w_j)$ ranked from smallest to largest. A short list of genes can be obtained by including all genes with $\hat{\tau}_0(w_j)$ less than some threshold c_o or by taking the top N_o genes in the ranked list.

Suppose that we select all genes with

$$\hat{\tau}_0(w_j) \leq c_o. \tag{6}$$

Then an estimate of the FDR is given by

$$\widehat{\text{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\, I_{[0,c_o]}(\hat{\tau}_0(w_j))/N_r, \tag{7}$$

where

$$N_r = \sum_{j=1}^{N} I_{[0,c_o]}(\hat{\tau}_0(w_j)) \tag{8}$$

is the number of the selected genes in the list. Here $I_A(w)$ is the indicator function that is one if w belongs to the interval A and is zero otherwise.

Thus we can find a data-dependent $c_o \leq 1$ as large as possible such that $\widehat{\text{FDR}} \leq \alpha$. This assumes that there will be some genes with $\hat{\tau}_0(w_j) \leq \alpha$, which will be true in the typical situation in practice. This bound is approximate due to the use of estimates in forming the posterior probabilities of nondifferential expression and so it depends on the fit of the densities $f_0(w_j)$ and $f(w_j)$.

2.5. *Bayes risk in terms of estimated FDR and FNR*

The Bayes prediction rule minimizes the risk of an allocation defined by Eq. (3). We can estimate the error of a false positive e_{01} and the error of a false negative e_{10} by

$$\hat{e}_{01} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\hat{z}_{1j} \Big/ \sum_{j=1}^{N} \hat{\tau}_0(w_j) \tag{9}$$

and

$$\hat{e}_{10} = \sum_{j=1}^{N} \hat{\tau}_1(w_j)\hat{z}_{0j} \Big/ \sum_{j=1}^{N} \hat{\tau}_1(w_j) \tag{10}$$

respectively, where $\hat{z}_{0j}$ is taken to be zero or one according as to whether $\hat{\tau}_0(w_j)$ is less than or greater than c in (4), and $\hat{z}_{1j} = 1 - \hat{z}_{0j}$. Also, we can estimate the prior probability π_0 as

$$\hat{\pi}_0 = \sum_{j=1}^{N} \hat{\tau}_0(w_j)/N. \tag{11}$$

On substituting these estimates (9) to (11) into the right-hand side of (3), the estimated risk can be written as

$$\widehat{\text{Risk}} = (1-c)\hat{\omega}\widehat{\text{FDR}} + c(1-\hat{\omega})\widehat{\text{FNR}}, \tag{12}$$

where

$$\widehat{\text{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\hat{z}_{1j} / \sum_{j=1}^{N} \hat{z}_{1j} \tag{13}$$

and

$$\widehat{\text{FNR}} = \sum_{j=1}^{N} \hat{\tau}_1(w_j)\hat{z}_{0j} / \sum_{j=1}^{N} \hat{z}_{0j} \tag{14}$$

are estimates of the FDR and FNR respectively, and where

$$\begin{aligned}\hat{\omega} &= \sum_{j=1}^{N} \hat{z}_{1j}/N \\ &= N_r/N \end{aligned} \tag{15}$$

is an estimate of the probability that a gene is selected. (Note that (13) is a restatement of (7).)

Thus unlike the tests or rules that are designed to control just the FDR, the Bayes rule approach in its selection of the genes can be viewed as controlling a linear combination of the FDR and FNR. The balance between the FDR and the FNR is controlled by the threshold c.

3. Estimation of Posterior Probabilities

3.1. *Mixture model approach*

We choose our test statistic W_j so that it has a normal distribution under the null hypothesis H_j that the jth gene is not differentially expressed. For example, if F_j denotes the usual test statistic (see Cochran and Cox,[14]) in a one-way analysis of variance of M observations from g classes, then we follow Broët *et al.*[15] and transform the F_j statistic as

$$W_j = \frac{\left(1 - \frac{2}{9(M-g)}\right) F_j^{\frac{1}{3}} - \left(1 - \frac{2}{9(g-1)}\right)}{\sqrt{\frac{2}{9(M-g)} F_j^{\frac{2}{3}} + \frac{2}{9(g-1)}}} \tag{16}$$

The distribution of the transformed statistic W_j is approximately a standard normal under the null hypothesis that the jth gene is not differentially expressed (that is, given its membership of population G_0). As noted in Broët *et al.*[15] it is remarkably accurate for $(M-g) \geq 10$. With this transformation, we can take the null density $f_0(w_j)$ to be the standard normal density (which has mean zero and unit variance). In order to estimate the mixing proportion π_0 and the mixture density $f(w_j)$, we postulate it to have the h-component normal mixture form

$$f(w_j) = \sum_{i=0}^{h-1} \pi_i \phi(w_j;\, \mu_i,\, \sigma_i^2), \tag{17}$$

where we specify $\mu_0 = 0$ and $\sigma_0^2 = 1$. In (17), $\phi(w_j;\, \mu_i,\, \sigma_i^2)$ denotes the normal density with mean μ_i and variance σ_i^2. We suggest starting with $h = 2$, adding more components if considered necessary as judged using the Bayesian Information Criterion (BIC).

3.2. *Use of P-values*

An an alternative to working with the test statistic W_j, we could follow the approach of Allison *et al.*[8] and use the associated P-value p_j. We can find these P-values using permutation methods whereby we permute the class labels. Using just the B permutations of the class labels for the gene-specific statistic W_j, the P-value for $W_j = w_j$ is assessed as

$$p_j = \frac{\#\{b : w_{0j}^{(b)} \geq w_j\}}{B}, \tag{18}$$

where $w_{0j}^{(b)}$ is the null version of w_j after the bth permutation of the class labels.

3.3. *Link with FDR*

Suppose that $\tau_0(w)$ is monotonic (decreasing in w). Then the rule (6) for declaring the jth gene to be differentially expressed is equivalent to

$$w \geq w_o, \tag{19}$$

where w_o is the value of w such that $\tau_0(w_o) = c_o$. The associated FDR, actually the positive FDR (Storey[16]), is given by

$$\pi_0 \frac{1 - F_0(w_o)}{1 - F(w_o)}. \tag{20}$$

Using (17), the positive FDR can be approximated using the fully parametric estimate for $F(w_o)$,

$$\hat{F}(w_o) = \pi_0 \Phi(w_o) + \sum_{i=1}^{h-1} \hat{\pi}_i \Phi(\frac{w_o - \hat{\mu}_i}{\hat{\sigma}_i}) \tag{21}$$

in the right-hand side of (20).

In the case where $\tau_0(w_j)$ is monotonic (decreasing in w_j), the inequality

$$\tau_0(w_j) < c_0 \tag{22}$$

is equivalent to

$$w_j > w_0 \tag{23}$$

for some threshold value w_0 of w_j. From (23), the (positive) FDR can be expressed as

$$\pi_0 \frac{1 - F_0(w_o)}{1 - F(w_o)}. \tag{24}$$

Alternatively, we could choose w_o, and hence c_o, so that (20) is equal to α. It thus also has an interpretation in terms of the q-value of Storey.[16] For if all genes with $\tau_0(w) \leq c_o$ are declared to be differentially expressed, then the FDR will be bounded above by α; see Efron *et al.*[10]

Concerning the link of this approach with the tail-area methodology of Benjamini and Hochberg,[11] suppose that the right-hand side of (20) is monotonic (decreasing) in w_0. Then as shown explicitly in Wit and McClure,[17] if we set π_0 equal to one and estimate $F(w_0)$ by its empirical distribution in the right-hand side of (20), the consequent rule is equivalent to the Benjamini-Hochberg procedure.

4. Application to Hedenfalk Breast Cancer Data

We analyze the publicly available cDNA microarray data set of Hedenfalk *et al.*[5] They studied the gene expression profiles of tumors from women with hereditary BRCA1- ($n_1 = 7$) and BRCA2-mutation positive cancer ($n_2 = 8$), here referred to as BRCA1 and BRCA2, as well as sporadic cases of breast cancer.

Hedenfalk *et al.* initially considered genes which could differentiate between the three types of breast cancer (BRCA1, BRCA2 and sporadic). They computed a modified F-statistic and used it to assign a P-value to each gene. A threshold of $\alpha = 0.001$ was selected to find 51 genes from a total of $N = 3,226$ that show differential gene expression. One of the main goals of the study was to identify the genes differentially expressed between the BRCA1 and BRCA2 cancers. They used a combination of three methods (modified t-test, weighted gene analysis and mutual-information scoring), and identified 176 significant genes.

Here we consider the gene expression data from the BRCA1 and BRCA2 tumors only. We use a subset of 3,170 genes, having eliminated genes with one or more measurements greater than 20, which was several interquartile ranges away from the interquartile range of all the data (as in Ref. 6). We then logged the data and standardised each patient's data to have mean 0 and variance 1. We applied our decision-theoretic approach to this data set. In Table 1, we report the estimated values of the FDR, calculated using (13), for various levels of the threshold c_o.

Table 1. Estimated FDR for various levels of c_o.

c_o	N_r	$\widehat{\text{FDR}}$
0.5	1086	0.27
0.4	817	0.21
0.3	584	0.16
0.2	378	0.11
0.1	158	0.06

It can be seen that if we were to declare the jth gene to be differentially expressed if $\tau_0(w_j) \leq 0.1$, then 158 genes would be selected as being significant, with an estimated FDR equal to 0.06. The prior probability of a gene not being differentially expressed (π_0) was estimated to be 0.465. We found that the above estimates, based on the semi-parametric version (13), were the same (to the second decimal place) as those calculated using the fully parametric estimate given in (20).

Of these 158 significant genes, 92 are over-expressed in BRCA1 tumors relative to BRCA2. Hedenfalk *et al.*[5] and also Storey and Tibshirani[6] in their further analysis of this data set, found too that a large block of genes are over-expressed in BRCA1. In particular, these included genes involved in DNA repair and cell death, such as MSH2 (DNA repair) and PDCD5 (induction of apoptosis), also identified by us. In their paper, Hedenfalk *et al.* noted that the finding of these over-expressed genes suggests that the BRCA1 mutation leads to a constitutive stress-type state.

Storey and Tibshirani identified 160 genes to be significant for differential expression between BRCA1 and BRCA2 by thresholding genes with q-values less than or equal to $\alpha = 0.05$ (an arbitrary cut-off value). Here the q-value of a particular gene is the expected proportion of false positives incurred when calling that gene significant, so that 8 of their 160 genes were expected to be false positives.

On comparing our 158 genes with the 160 identified by Storey and Tibshirani, we found that there were 122 genes in common. Of the 36 excluded genes, 10 were included in the Hedenfalk set of 176. The functional classes (where known) of the remaining 26 genes are shown in Table 2.

Of the 38 genes found by Storey and Tibshirani but not by the present approach, 28 were included in the Hedenfalk set.

We also applied the SAM (v2.0) method of Tusher *et al.*[2] to the data set. Using an FDR cut-off of 5%, 210 genes were selected as significant. Of these, 109 were in common with the 158 genes chosen by us, and 132 in common with the 160 genes as picked by Storey and Tibshirani.

Broët *et al.*[7] recently also applied a mixture model appproach to identify differentially expressed genes in this data set. However, they implemented a Bayesian approach, in contrast to the frequentist approach as applied here. They obtained a slightly different estimate for π_0 of 0.52, hence rejecting 52% of the genes as not differentially expressed, as opposed to our value of 46.5%. In their approach, they did not constrain the variance of the first component to be one because it presents computational problems implementing the Bayesian solution via MCMC methods. However, using the frequentist approach, we were able to fix the variance to be one. As Broët *et al.*'s list of genes was not made available, we were unable to compare our gene list to theirs.

Table 2. Functional classes for uniquely identified genes.

Functional class	Gene identifier
Kinase Activity (plus protein or nucleotide binding)	MAST4, ITPK1, PRKCBP1, MADD
Nucelotide Binding	RMB17, HARS
Protein Binding	CLTC, TNFAIP1
Receptor activity/ Protein Binding	ITGB5, ITGA3
Signal transduction/ nucleotide binding	RHOC
Hydrolase activity	RNPEP, HDAC3, GNS
Protease inhibitor	A2M
Oxidoreductase/Dehydrogenase activity	HSD17B4, ACOX1
Transcription factor activity	GATA3, ZNF500
Unknown	LRBA, PPP1R15A

5. Simulation Study

Allison *et al.*[8] were interested in looking at the effect of the assumption of independently distributed expression levels of the genes. To this end, they generated gene expression levels for M experiments (with $M/2$ "mice" per experimental group) and for $N = 3000$ genes. The M vectors y_j of dimension N were generated randomly from a multivariate normal distribution with covariance matrix specified to be

$$\Sigma = \sigma^2 B \otimes I_6 \tag{25}$$

and

$$B = 1_{500}\, 1_{500}^{\mathrm{T}} \rho + (1-\rho) 1_{500}.$$

Here 1_{500} denotes the unit vector of length 500 and I_m is the $m \times m$ identity matrix.

For the simulations the common variance was $\sigma^2 = 4$, while the correlation ρ varied over three values of 0 (independence), 0.4 (moderate dependence), and 0.8 (strong dependence). They noted that this covariance structure seems plausible since groups of genes are likely to be coexpressed, but it is unlikely that a particular gene is correlated with all other genes. For 20% of the genes (600 randomly selected), a true mean difference in expression between the two classes of mice was incorporated by adding d to the gene measurements y_j from $j = \frac{1}{2}M + 1$ through to M.

We applied our mixture model approach, using $d = 0, 4, 8$ and $M = 10$. As before, we transformed the pooled t-statistic according to (16), with $F_j = t_j^2$. The two cases of $d = 4, 8$ (where there is a true mean difference between the groups), with varying levels of dependence, are shown in Figs. 1–6. We fitted normal components with the restriction that one component must be $N(0, 1)$, that is, a theoretical null component. In Figs. 1–6, these components are superimposed on the same histograms. Figures 7–12 are similar to Figs. 1–6 except that we do not apply restrictions to the null component of the

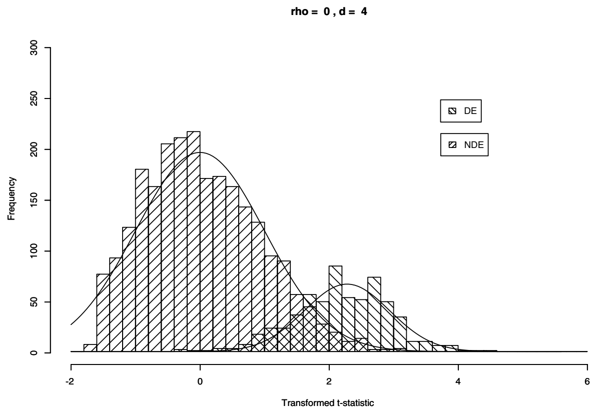

Fig. 1. Independence and mean difference of 4 with theoretical null.

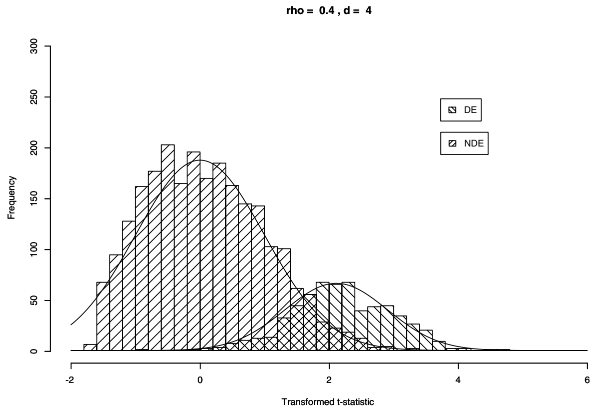

Fig. 2. Moderate dependence and mean difference of 4 with theoretical null.

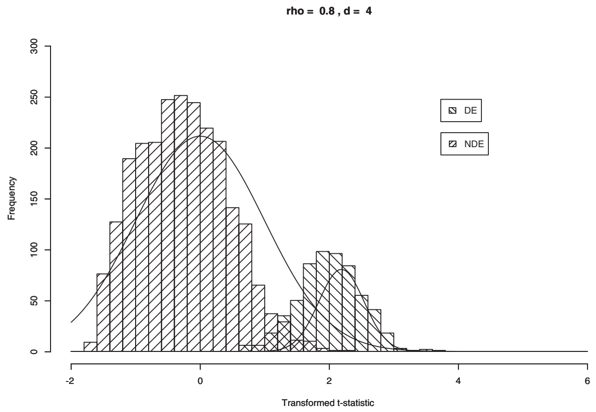

Fig. 3. Strong dependence and mean difference of 4 with theoretical null.

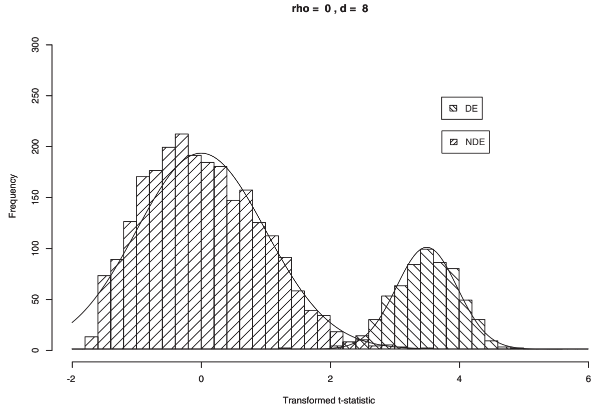

Fig. 4. Independence and mean difference of 8 with theoretical null.

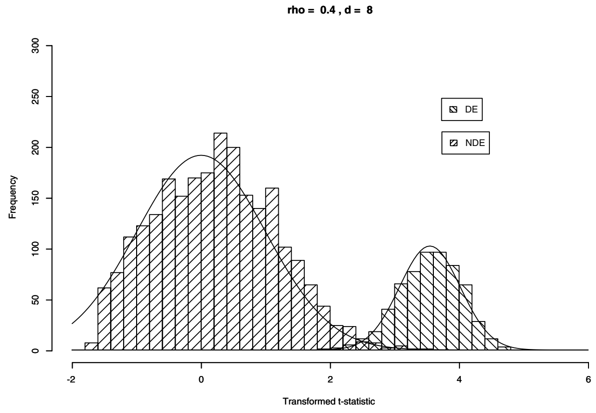

Fig. 5. Moderate dependence and mean difference of 8 with theoretical null.

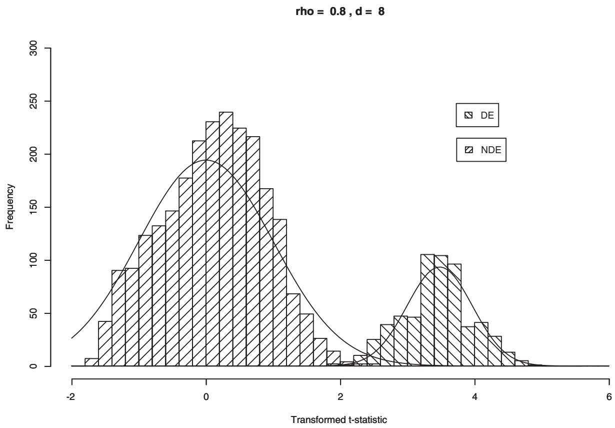

Fig. 6. Strong dependence and mean difference of 8 with theoretical null.

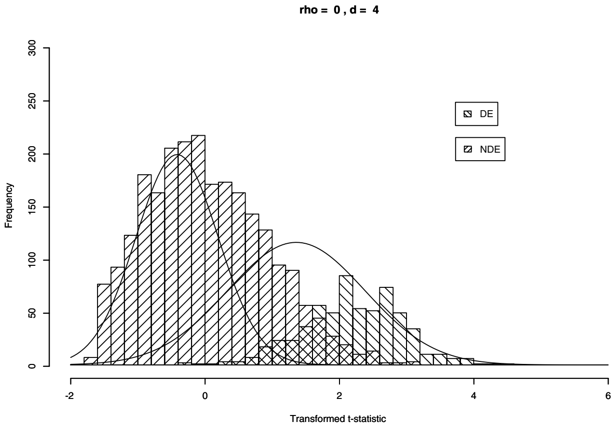

Fig. 7. Independence and mean difference of 4 with empirical null.

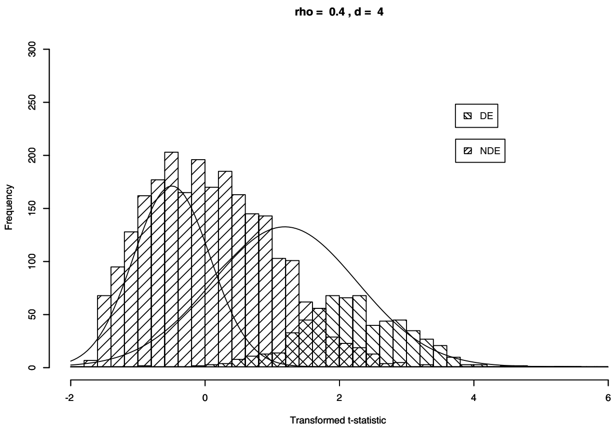

Fig. 8. Moderate dependence and mean difference of 4 with empirical null.

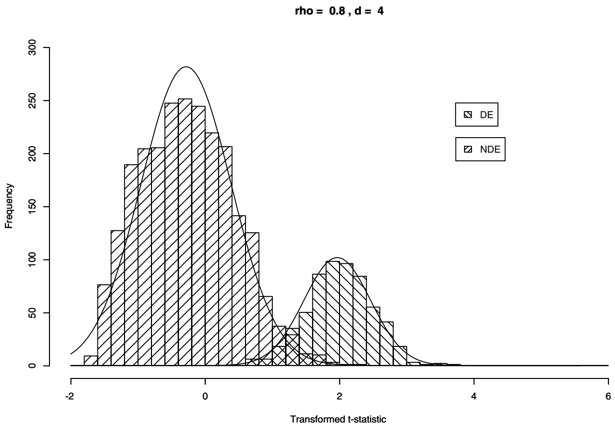

Fig. 9. Strong dependence and mean difference of 4 with empirical null

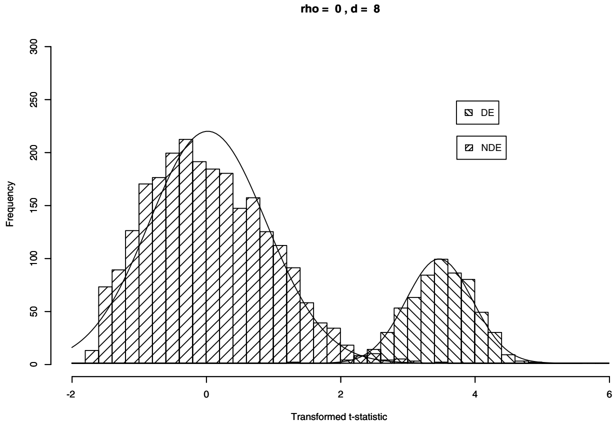

Fig. 10. Independence and mean difference of 8 with empirical null.

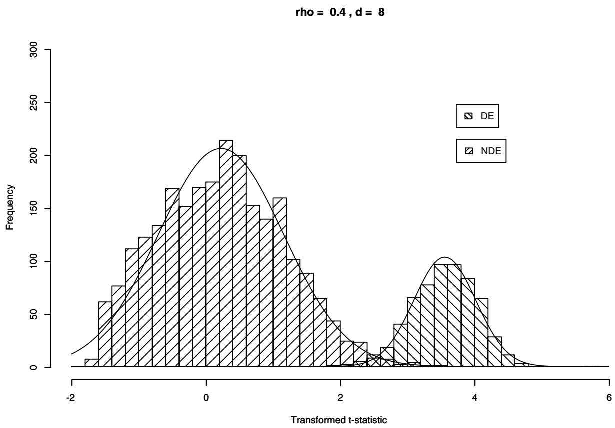

Fig. 11. Moderate dependence and mean difference of 8 with empirical null.

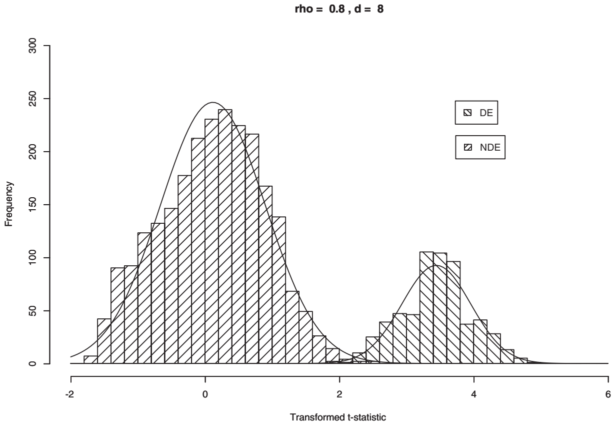

Fig. 12. Strong dependence and mean difference of 8 with empirical null.

two-component normal mixture model fittted. Following Efron,[18] we call the component with the smaller mean the *empirical* null component.

In each plot, the first component (the null component) corresponds to the nondifferentially expressed genes (NDE) and the second component to the differentially expressed genes (DE).

It can be seen that as the correlation increases, the fit of the theoretical null component becomes poorer. In the case of $d = 8$ for which the Mahalanobis distance (Δ) between the means of the DE and NDE genes is large ($\Delta = 8/2 = 4$), the empirical null provides an improved fit to the NDE genes. But fitting either a theoretical or empirical null component gives a π_0 value of almost exactly 0.8, that is, the true π_0 value.

When the mean difference between the DE and NDE genes is only $d = 4$ (that is, the Mahalanobis distance is only moderate with Δ is 2), it can be seen that the fit of the theoretical null component is very poor in the case of strong correlation ($\rho = 0.8$). In this case, it can be seen that the empirical null provides an improved fit to the NDE genes. For moderate correlation ($\rho = 0.4$) the fit of the empirical null is quite poor, but it is not needed, as the fit of the theoretical null is adequate.

The t-statistic, t_j, transformed according to (16) has a minimum value of $-7/3\sqrt{2}$ when $F_j = 0$. Thus some of the histograms appear to taper off sharply at the left hand end of the plots. This has led McLachlan *et al.*[19] to work with a normal score-based statistic, which is similar to that used in Efron.[18]

6. Conclusions

We use a mixture model-based approach to finding differentially expressed genes in microarray data, and show that for the Hedenfalk data set this approach can provide useful information beyond that of other methods.

We consider also a simulation study, with varying levels of correlation between groups of genes and in the mean difference, d, in their expression levels between the two classes. Not surprisingly, it is demonstrated that for high values of d, the correlation has little impact on the detection of differentially expressed genes. However, for moderate values of d, the correlation can affect this detection as the theoretical null distribution would not appear to fit the observed distribution of the null genes. In situations where this is the case, an improved fit is given by the so-called empirical null distribution obtained by relaxing the imposition of a zero mean and unit variance on the null component in the two-component mixture model fitted to the data.

Finally, it is worth noting that genes which score as most significant using standard methods for multiple hypothesis testing may not necessarily be of most biological relevance (see Ref. 7). Genes with more subtle changes in their expression levels, indicating that they are more tightly regulated, may be of more importance in the biology of tumor formation.

References

1. G. J. McLachlan, K. A. Do and C. Ambroise, *Analyzing Microarray Gene Expression Data* (New York, Wiley, 2004).
2. V. G. Tusher, R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, in *Proceedings of the National Academy of Sciences USA* **98** (2001), pp. 5116–5121.
3. W. Pan, J. Lin, and C. T. Le, A mixture model approach to detecting differentially expressed genes with microarray data, *Functional and Integrative Genomics* **3** (2003) 117–124.
4. W. Pan, A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, *Bioinformatics* **18** (2002) 546–554.
5. I. Hedenfalk, D. Duggan, Y. D. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi *et al.*, Gene-expression profiles in hereditary breast cancer, *The New England Journal of Medicine* **344** (2001) 539–548.

6. J. D. Storey and R. Tibshirani, Statistical significance for genome-wide studies, in *Proceedings of the National Academy of Sciences USA* **100** (2003) 9440–9445.
7. P. Broët, A. Lewin, S. Richardson, C. Dalmasso and H. Magdelenat, A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments, *Bioinformatics* **20** (2004) 2562–2571.
8. D. B. Allison, G. L. Gadbury, M. Heo, J. R. Fernandez, C.-K. Lee, T. A. Prolla and R. Weindruch, A mixture model approach for the analysis of microarray gene expression data, *Computational Statistics and Data Analysis* **39** (2002) 1–20.
9. M.-L. T. Lee, F. C. Kuo, G. A. Whitmore and J. Sklar, Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations, in *Proceedings of the National Academy of Sciences USA* **97** (2000) 9834–9838.
10. B. Efron, R. Tibshirani, J. D. Storey and V. Tusher, Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* **96** (2001) 1151–1160.
11. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society* **B57** (1995) 289–300.
12. C. R. Genovese and L. Wasserman, Operating characteristics and extensions of the false discovery rate procedure, *Journal of the Royal Statistical Society B* **64** (2002) 499–517.
13. G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition* (New York, Wiley, 1992).
14. W. G. Cochran and G. Cox, *Experimental Designs* (New York, Wiley, 1957).
15. P. Broët, S. Richardson and F. Radvanyi, Bayesian hierarchical model for identifying changes in gene expression from microarray experiments, *Journal of Computational Biology* **9** (2002) 671–683.
16. J. Storey, The positive false discovery rate: A Bayesian interpretation and the q-value, *Annals of Statistics* **31** (2004) 2013–2035.
17. E. Wit and J. McClure, *Statistics for Microarrays: Design, Analysis and Inference* (Chichester, Wiley, 2004).
18. B. Efron, Large-scale simultaneous hypothesis Testing: The choice of a null hypothesis, *Journal of the American Statistical Association* **99** (2004) 96–104.
19. G. J. McLachlan, R. W. Bean and L. Ben-Tovim Jones, A simple implementation of a normal mixture approach to the detection of differential expression, unpublished manuscript (2005).