BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1608–1615
doi:10.1093/bioinformatics/btl148

Gene expression

A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays

G.J. McLachlan[1,2,*], R.W. Bean[2] and L. Ben-Tovim Jones[2]

[1]Department of Mathematics, University of Queensland and [2]ARC Centre in Bioinformatics, Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane 4072, Australia

Received on January 25, 2006; revised on April 10, 2006; accepted on April 12, 2006
Advance Access publication April 21, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: An important problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes. We provide a straightforward and easily implemented method for estimating the posterior probability that an individual gene is null. The problem can be expressed in a two-component mixture framework, using an empirical Bayes approach. Current methods of implementing this approach either have some limitations due to the minimal assumptions made or with more specific assumptions are computationally intensive.

Results: By converting to a *z*-score the value of the test statistic used to test the significance of each gene, we propose a simple two-component normal mixture that models adequately the distribution of this score. The usefulness of our approach is demonstrated on three real datasets.

Availability: An R-program for implementing the approach is freely available at http://www.maths.uq.edu.au/~gjm/

Contact: gjm@maths.uq.edu.au

Supplementary information: http://www.maths.uq.edu.au/~gjm/bioinf061supp_data.pdf

1 INTRODUCTION

Often the first step, and indeed the major goal for many microarray studies, is the detection of genes that are differentially expressed in a known number of classes, $C_1, \ldots, C_g$. Statistical significance of differential expression can be tested by performing a test for each gene. When many hypotheses are tested, the probability that a type I error (a false positive error) is committed increases sharply with the number of hypotheses. In this paper, we focus on the use of a two-component mixture model to handle the multiplicity issue. This model is becoming more widely adopted in the context of microarrays, where one component density corresponds to that of the test statistic for genes that are not differentially expressed, and the other component density to that of the test statistic for genes that are differentially expressed. For the adopted test statistic, we propose that its values be transformed to *z*-scores, whose null and non-null distributions can be represented by a single normal each. We show how this two-component normal mixture model can be fitted very quickly via the EM algorithm started from a point that is completely determined by an initial specification of the proportion π_0 of genes that are not differentially expressed. A procedure for determining suitable initial values for π_0 is suggested in the case where the null density is taken to be standard normal (the theoretical null distribution). We also consider the provision of an initial partition of the genes into two groups for the application of the EM algorithm in the case where the adoption of the theoretical null distribution would appear not to be appropriate and an empirical null distribution needs to be used. We demonstrate our approach on three datasets that have been analyzed previously in the bioinformatics literature.

2 BACKGROUND

2.1 Notation

Although biological experiments vary considerably in their design, the data generated by microarrays can be viewed as a matrix of expression levels. For M microarray experiments (corresponding to M tissue samples), where we measure the expression levels of N genes in each experiment, the results can be represented by the $N \times M$ matrix. Typically, M is no more than 100 (usually much less in the present context), while the number of genes N is of the order of 10^4. The M tissue samples on the N available genes are classified with respect to g different classes, and it is assumed that the (logged) expression levels have been preprocessed with adjustment for array effects.

2.2 Detection of differential expression

Differential expression of a gene means that the (class-conditional) distribution of its expression levels is not the same for all g classes. These distributions can differ in any possible way, but the statistics usually adopted are designed to be sensitive to primarily a difference in the means; for example, the one-way analysis of variance (ANOVA) F-statistic. Even so, the gene hypotheses being tested are of equality of distributions across the g classes, which allows the use of permutation methods to estimate P-values if necessary.

In the special case of $g = 2$ classes, the one-way ANOVA F-statistic reduces to the square of the classical (pooled) t-statistic. Various refinements of the t-statistic have been suggested; see, for example, the procedure of Tusher *et al*. (2001).

3 TWO-COMPONENT MIXTURE MODEL

3.1 Posterior probability of non-differential expression

In this paper, we focus on a decision–theoretic approach to the problem of finding genes that are differentially expressed. We

[*]To whom correspondence should be addressed.

use a prediction rule approach based on a two-component mixture model as formulated in Lee *et al*. (2000) and Efron *et al*. (2001). We let G denote the population of genes under consideration. It can be decomposed into two groups G_0 and G_1, where G_0 is the group 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 π_i denote the prior probability of a gene belonging to G_i ($i = 0, 1$), and assume that the common density of the test statistic W_j for a gene j in G_i is $f_i(w_j)$. The unconditional density of W_j is then given by the two-component mixture model,

$$f(w_j) = \pi_0 f_0(w_j) + \pi_1 f_1(w_j). \tag{1}$$

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

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

In this framework, the gene-specific posterior probabilities provide the basis for optimal statistical inference about differential expression. The posterior probability $\tau_0(w_j)$ has been termed the local false discovery rate (local FDR) by Efron and Tibshirani (2002). It quantifies the gene-specific evidence for each gene. As noted by Efron (2004), it can be viewed as an empirical Bayes version of the Benjamini–Hochberg (1995) methodology, using densities rather than tail areas.

It can be seen from (2) that in order to use this posterior probability of non-differential expression in practice, we need to be able to estimate π_0, the mixture density $f(w_j)$ and the null density $f_0(w_j)$, or equivalently, the ratio of densities $f_0(w_j)/f(w_j)$. Efron *et al*. (2001) have developed a simple empirical Bayes approach to this problem with minimal assumptions. This problem has been studied since under more specific assumptions, including works by Newton *et al*. (2001, 2004), Lönnstedt and Speed (2002), Pan *et al*. (2003), Zhao and Pan (2003), Broët *et al*. (2004), Newton *et al*. (2004), Smyth (2004), Do *et al*. (2005) and Gottardo *et al*. (2006), among many others. The fully parametric methods that have been proposed are computationally intensive.

3.2 Bayes decision rule

Let e_{01} and e_{10} denote the two errors when a rule is used to assign a gene as being differentially expressed or not, where e_{01} is the probability of a false positive and e_{10} is the probability of a false negative. That is, the sensitivity is $1 - e_{10}$ and the specificity is $1 - e_{01}$. The so-called risk of allocation 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 $\tau_0(w_j) \leq c$; otherwise, the j-th gene is assigned to G_0.

4 SELECTION OF GENES

In practice, we do not know the prior probability π_0 nor the densities $f_0(w_j)$ and $f(w_j)$, which will have to be estimated. We shall shortly discuss a simple and quick approach to this estimation problem. If $\hat{\pi}_0$, $\hat{f}_0(w_j)$ and $\hat{f}_1(w_j)$ denote estimates of π_0, $f_0(w_j)$ and $f_1(w_j)$, respectively, the gene-specific summaries of differential expression can be expressed in terms of the estimated posterior probabilities $\hat{\tau}_0(w_j)$, where

$$\hat{\tau}_0(w_j) = \hat{\pi}_0 \hat{f}_0(w_j)/\hat{f}(w_j) \quad (j = 1, \ldots, N) \tag{4}$$

is the estimated posterior probability that the j-th 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.

4.1 FDR

Suppose that we select all genes with

$$\hat{\tau}_0(w_j) \leq c_o. \tag{5}$$

Then McLachlan *et al*. (2004) have proposed that the false discovery rate (FDR) of Benjamini–Hochberg (1995) can be estimated as

$$\widehat{\text{FDR}} = \frac{\sum_{j=1}^{N} \hat{\tau}_0(w_j) I_{[0,\, c_o]}(\hat{\tau}_0(w_j))}{N_r}, \tag{6}$$

where N_r is the number of selected genes and $I_A(x)$ is the indicator function, which is one if $x \in A$ and is zero otherwise.

Similarly, the false non-discovery rate (FNDR) can be estimated by

$$\widehat{\text{FNDR}} = \frac{\sum_{j=1}^{N} \hat{\tau}_1(w_j) I_{(c_o,\, \infty)}(\hat{\tau}_0(w_j))}{(N - N_r)}. \tag{7}$$

We can also estimate the false positive rate (FPR), e_{01}, and the false negative rate (FNR), e_{10}, in a similar manner to give

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

and

$$\widehat{\text{FNR}} = \frac{\sum_{j=1}^{N} \hat{\tau}_1(w_j)\; I_{(c_o,\, \infty)}(\hat{\tau}_0(w_j))}{\sum_{j=1}^{N} \hat{\tau}_1(w_j)}, \tag{9}$$

respectively.

When controlling the FDR, it is important to have a guide to the value of the associated FNR in particular, as setting the FDR too low may result in too many false negatives in situations where the genes of interest (related to biological pathway or target drug) are not necessarily the top ranked genes; see, for example, Pawitan *et al*. (2005). The local FDR in the form of the posterior probability of non-differential expression of a gene has an advantage over the global measure of FDR in interpreting the data for an individual gene; see more details in Efron (2005b) and also Supplementary information.

5 USE OF Z-SCORE

5.1 Normal transformation

We let W_j denote the test statistic for the test of the null hypothesis

$$H_j: j\text{-th gene is not differentially expressed.} \tag{10}$$

For example, as discussed above, W_j might be the t- or F-statistic, depending on whether there are two or multiple classes. Whatever the test statistic, we proceed in a similar manner as in Efron (2004) and transform the observed value of the test statistic to a z-score given by

$$z_j = \Phi^{-1}(1 - P_j), \tag{11}$$

where P_j is the P-value for the value w_j of the original test statistic W_j and Φ is the $N(0, 1)$ distribution function. Thus

$$P_j = 1 - F_0(w_j) + F_0(-w_j), \tag{12}$$

where F_0 is the null distribution of W_j. If F_0 is the true null distribution, then the null distribution of the test statistic Z_j corresponding to z_j is exactly standard normal. With this definition of z_j, departures from the null are indicated by large positive values of z_j. Our transformation (11) is slightly different to that in Efron (2004), as we wish that only large positive values of the z-score be consistent with the alternative hypothesis; that is, we want the latter to be (upper) one-sided so that the non-null distribution of the z-score can be represented by a single normal distribution rather than a mixture in equal proportions of two normal components with means of opposite sign. Previously, Allison *et al*. (2002) had considered mixture modelling of the P-values directly in terms of a mixture of beta distributions with the uniform (0, 1) distribution (a special form of a beta distribution) as the null component. Pounds and Morris (2003) considered a less flexible beta mixture model for the P-values, being a mixture of a uniform (0, 1) distribution for the null and a single beta distribution for the non-null component. In the work of Broët *et al*. (2004), they used a transformation similar to the approximation of Wilson and Hilferty (1931) for the χ^2 distribution to transform the value F_j for the F-statistic for the j-th gene to an approximate z-score.

5.2 Permutation assessment of *P*-value

In cases where we are unwilling to assume the null distribution F_0 of the original test statistic W_j for use in our normal transformation (11), we can obtain an assessment of the P-value P_j via permutation methods. We can use just permutations of the class labels for the gene-specific statistic W_j. This suffers from a granularity problem, since it estimates the P-value with a resolution of only $1/B$, where B is the number of the permutations. Hence it is common to pool over all N genes; see Supplementary information for details. The drawback of pooling the null statistics across the genes to assess the null distribution of W_j is that one is using different distributions unless all the null hypotheses H_j are true. The distribution of the null values of the differentially expressed genes is different from that of the truly null genes, and so the tails of the true null distribution of the test statistic is overestimated, leading to conservative inferences; see, for example, Pan (2003), Guo and Pan (2005) and Xie *et al*. (2005).

6 TWO-COMPONENT NORMAL MIXTURE

We now proceed to show that by working in terms of the z_j-scores as defined by (11), we can provide a parametric version of the two-component mixture model (1) that is easy to fit. The density of the test statistic Z_j corresponding to the use of the z-score (11) for the j-th gene is to be represented by the two-component normal mixture model

$$f(z_j) = \pi_0 f_0(z_j) + \pi_1 f_1(z_j), \tag{13}$$

where $\pi_1 = 1 - \pi_0$. In (13), $f_0(z_j) = \phi(z_j; 0, 1)$ is the (theoretical) null density of Z_j, where $\phi(z; \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2, and $f_1(z_j)$ is the non-null density of Z_j. It can be approximated with arbitrary accuracy by taking q sufficiently large in the normal mixture representation

$$f_1(z_j) = \sum_{h=1}^{q} \pi_{1h}\phi(z_j; \mu_{1h}, \sigma_{1h}^2). \tag{14}$$

For the datasets that we have analyzed, it has been sufficient to use just a single normal component ($q = 1$) in (14). In such cases, we can write (13) as

$$f(z_j) = \pi_0\phi(z_j; 0, 1) + \pi_1\phi(z_j; \mu_1, \sigma_1^2). \tag{15}$$

As pointed out in a series of papers by Efron (2004, 2005a, b), for some microarray datasets the normal scores do not appear to have the theoretical null distribution, which is the standard normal. In this case, Efron has considered the estimation of the actual null distribution called the empirical null as distinct from the theoretical null. As explained in Efron (2005b), the two-component mixture model (1) assumes two classes, null and non-null, whereas in reality the differences between the genes range smoothly from zero or near zero to very large.

In the case where the theoretical null distribution does not appear to be valid and the use of an empirical null distribution would seem appropriate, we shall adopt the two-component mixture model obtained by replacing the standard normal density by a normal with mean μ_0 and variance σ_0^2 to be inferred from the data. That is, the density of the z_j-score is modelled as

$$f(z_j) = \pi_0\phi(z_j; \mu_0, \sigma_0^2) + \pi_1\phi(z_j; \mu_1, \sigma_1^2). \tag{16}$$

In the sequel, we shall model the density of the z_j-score by (16). In the case of the theoretical $N(0, 1)$ null being adopted, we shall set $\mu_0 = 0$ and $\sigma_0^2 = 1$ in (16).

7 FITTING OF NORMAL MIXTURE MODEL

7.1 Theoretical null

We now consider the fitting of the two-component mixture model (15) to the z_j, firstly with the theoretical $N(0, 1)$ null adopted. In order to fit the two-component normal mixture (15), we need to be able to estimate π_0, μ_1 and σ_1^2. This is effected by maximum likelihood (ML) via the EM algorithm of Dempster *et al*. (1977), using the EMMIX program as described in McLachlan and Peel (2000); see also McLachlan and Krishnan (1997). To provide a suitable starting value for the EM algorithm in this task, we note that the ML estimate of the parameters in a two-component mixture model satisfies the moment equations obtained by equating the sample mean and variance of the mixture to their population counterparts, which gives

$$\bar{z} = \hat{\pi}_0\hat{\mu}_0 + \hat{\pi}_1\hat{\mu}_1 \tag{17}$$

and

$$s_z^2 = \hat{\pi}_0\hat{\sigma}_0^2 + \hat{\pi}_1\hat{\sigma}_1^2 + \hat{\pi}_0\hat{\pi}_1(\hat{\mu}_0 - \hat{\mu}_1)^2, \tag{18}$$

where $\hat{\pi}_1 = 1 - \hat{\pi}_0$. For the theoretical null, $\hat{\mu}_0 = 0$ and $\hat{\sigma}_0^2 = 1$ and on substituting for them in (17) and (18), we obtain

$$\hat{\mu}_1 = \bar{z}/(1 - \hat{\pi}_0) \tag{19}$$

and

$$\hat{\sigma}_1^2 = \{s_z^2 - \hat{\pi}_0 - \hat{\pi}_0(1 - \hat{\pi}_0)\hat{\mu}_1^2\}/(1 - \hat{\pi}_0). \tag{20}$$

Hence with the specification of an initial value $\pi_0^{(0)}$ for $\hat{\pi}_0$, initial values for the other parameters to be estimated, μ_1 and σ_1^2, are automatically obtained from (19) and (20). If there is a problem in so finding a suitable solution for $\mu_1^{(0)}$ and $\sigma_1^{(0)^2}$, it gives a clue that perhaps the theoretical null is inappropriate and that consideration should be given to the use of an empirical null, as to be discussed shortly.

Following the approach of Storey and Tibshirani (2003) to the estimation of π_0, we can obtain an initial estimate $\pi_0^{(0)}$ for use in (19) and (20) by taking $\pi_0^{(0)}$ to be

$$\pi_0^{(0)}(\xi) = \#\{z_j : z_j < \xi\}/\{N\Phi(\xi)\}, \tag{21}$$

for an appropriate value of ξ. There is an inherent bias-variance trade-off in the choice of ξ. In most cases, as ξ grows smaller, the bias of $\pi_0^{(0)}(\xi)$ grows larger, but the variance becomes smaller.

7.2 Empirical null

In this case, we do not assume that the mean μ_0 and variance σ_0^2 of the null distribution are zero and one, respectively, but rather they are estimated in addition to the other parameters π_0, μ_1 and σ_1^2. For an initial value $\pi_0^{(0)}$ for π_0, we let n_0 be the greatest integer less than or equal to $N\pi_0^{(0)}$, and assign the n_0 smallest values of the z_j to one class corresponding to the null component and the remaining $N - n_0$ to the other class corresponding to the alternative component. We then obtain initial values for the mean and variances of the null and alternative components by taking them equal to the means and variances of the corresponding classes so formed. The two-component mixture model is then run from these starting values for the parameters. For the real datasets to be considered next, the fitting of the normal mixture model takes <0.1 s. The calculation of the t-statistics in the first instance for the individual genes takes ~4 s for the Hedenfalk dataset and 2 s for the other two sets on a Pentium IV 1.8 GHz computer.

8 EXAMPLE 1: BREAST CANCER DATA

For our first example, we consider some data from the study of Hedenfalk *et al*. (2001), which examined gene expressions in breast cancer tissues from women who were carriers of the hereditary BRCA1 or BRCA2 gene mutations, predisposing to breast cancer. The dataset comprised the measurement of $N = 3226$ genes using cDNA arrays, for $n_1 = 7$ BRCA1 tumours and $n_2 = 8$ BRCA2 tumours. We column normalized the logged expression values, and ran our analysis with the aim of finding differentially expressed genes between the tumours associated with the different mutations. As in Efron (2004), we adopted the classical pooled t-statistic as our test statistic W_j for each gene j and we used the t-distribution function with 13 degrees of freedom, F_{13}, as the null distribution of W_j in the computation of the P-value P_j from (12).

We fitted the two-component normal mixture model (15) with the standard normal $N(0, 1)$ as the theoretical null, using various values of $\pi_0^{(0)}$, as obtained from (21). For example, using (21) for $\xi = 0$ and -0.675 led to the initial values of 0.70 and 0.66 for $\pi_0^{(0)}$. The fit we obtained (corresponding to the largest local maximum) is given by $\hat{\pi}_0 = 0.65$, $\hat{\mu}_1 = 1.49$, and $\hat{\sigma}_1^2 = 0.94$. In Figure 1, we display the fitted mixture density superimposed on the histogram of z_j-scores, along with its two components, the theoretical $N(0, 1)$ null density and the $N(1.49, 0.94)$ non-null density weighted by their prior probabilities of $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$. It can be seen that this two-component normal mixture model gives a good fit to the empirical distribution of the z_j-scores.

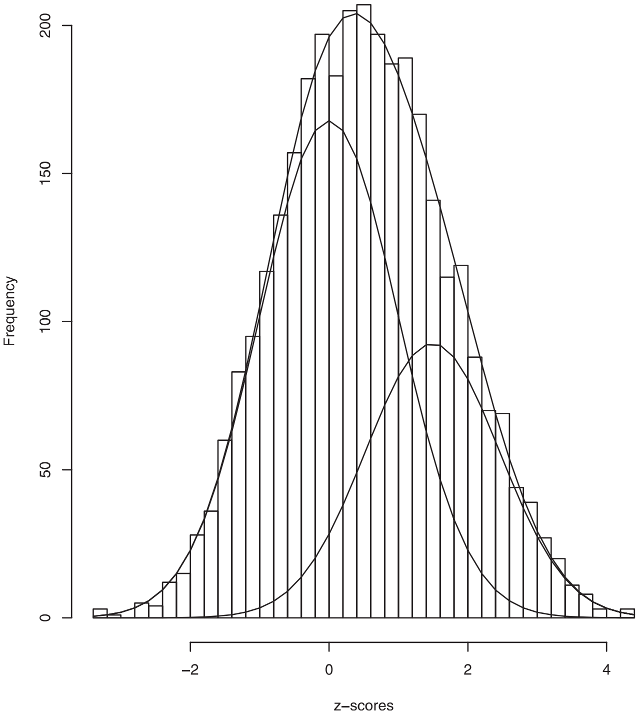

Fig. 1. Breast cancer data: plot of fitted two-component normal mixture model with theoretical $N(0, 1)$ null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$) imposed on histogram of z-scores.

In Table 1, we have listed the FDR estimated from (6) for various levels of the threshold c_o in (5). It can be seen, for example, that if c_o is set equal to 0.1, then the estimated FDR is 0.06 and $N_r = 143$ genes would be declared to be differentially expressed. It is not suggested that the FDR should be controlled to be around 0.05. It is just that in this example, its control at this approximate level yields a number (143) of differentially expressed genes that is not too unwieldy for a biologist to handle in subsequent confirmatory experiments; the choice of c_o is discussed in Efron (2005b).

In the original paper, Hedenfalk *et al*. (2001) selected 176 genes based on a modified F-test, with a p-value cut off of 0.001. Comparing genes which were selected in our set of 143, we found 107 in common, including genes involved in DNA repair and cell death, which are over-expressed in BRCA1-mutation-positive tumours, such as MSH2 (DNA repair) and PDCD5 (induction of apoptosis). Storey and Tibshirani (2003) in their analysis of this dataset, selected 160 genes by thresholding genes with q-values less than or equal to $\alpha = 0.05$ (an arbitrary cut-off value), of which there are 113 in common with our set of 143. Overall, 101 genes were selected in common to all three studies, with 24 genes unique to

Table 1. Estimated FDR and other error rates for various levels of the threshold $c_{\rm o}$ applied to the posterior probability of non-differential expression for the breast cancer data, where N_r is the number of selected genes

$c_{\rm o}$	N_r	$\widehat{\rm FDR}$	$\widehat{\rm FNDR}$	$\widehat{\rm FNR}$	$\widehat{\rm FPR}$
0.1	143	0.06	0.32	0.88	0.004
0.2	338	0.11	0.28	0.73	0.02
0.3	539	0.16	0.25	0.60	0.04
0.4	742	0.21	0.22	0.48	0.08
0.5	971	0.27	0.18	0.37	0.12

our set. We searched publicly available databases for the biological functions of these genes, and found these included DNA repair, cell cycle control and cell death, suggesting good evidence for inclusion of these genes (see the Supplementary information for a fuller description).

Concerning the other type of allocation rates for the choice of $c_{\rm o} = 0.1$ in (5), the estimates of the FNDR, FNR and FPR are equal to 0.32, 0.88 and 0.004, respectively. The FNR of 0.88 means that there would be quite a few false negatives among the genes declared to be null (not differentially expressed).

Among other analyses of this dataset, π_0 was estimated to be 0.52 by Broët *et al*. (2004), 0.64 by Gottardo *et al*. (2006), 0.61 by Ploner *et al*. (2006) and 0.47 by Storey (2002). In the fully parametric Bayesian approach of Broët *et al*. (2004), the mean of the null component was fixed at zero, but the variance was allowed to be free during the estimation process for computational convenience. In Ploner *et al*. (2006), 56 genes with highly extreme expression values were first removed as in Storey and Tibshirani (2003).

We also considered the fitting of the two-component normal mixture model (16) with the null component mean and variance, μ_0 and σ_0^2, now estimated in addition to π_0 and the non-null mean and variance, μ_1 and σ_1^2. We found that this fit from using the empirical null in place of the $N(0, 1)$ theoretical null is similar to the fit in Figure 1; see Figure 1 in Supplementary information. Efron (2003) writes that 'there is ample reason to distrust the theoretical null' in the case of the Hedenfalk data. The difference in our findings may be due to the fact that our gene expression data seems to differ when compared with the expression data presented in Efron and Tibshirani (2002). In other analyses of this dataset, Newton *et al*. (2001), Tusher *et al*. (2001) and Gottardo *et al*. (2006) concluded that there were 375, 374 and 291 genes, respectively, differentially expressed when the FDR is controlled at the 10% level. It can be seen from Table 1 that our approach gives 338 genes as being differentially expressed when the threshold of $c_{\rm o} = 0.2$ is imposed on the posterior probability of non-differential expression for which the implied FDR is 11% and the FNR is 73%. The corresponding values with the use of the empirical null are 12 and 76% for the FDR and FNR, respectively, with 235 genes declared to be differentially expressed. According to the Bayesian Information criterion (BIC), the empirical null would not be selected in favor of the theoretical $N(0, 1)$ null (see Supplementary information). The (main) reason for fewer genes being declared differentially expressed with the use of the empirical than with the theoretical null is that the estimate of π_0 is greater ($\hat{\pi}_0 = 0.73$).

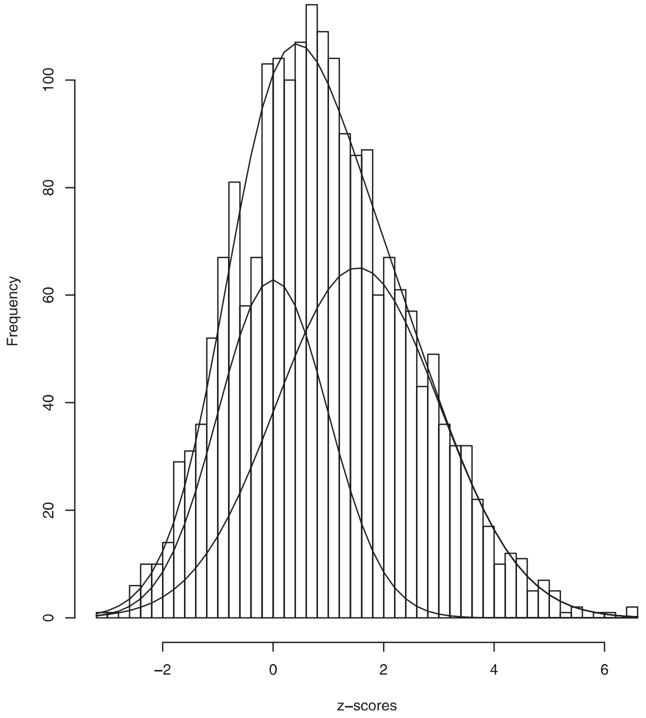

Fig. 2. Colon cancer data: plot of fitted two-component normal mixture model with theoretical $N(0, 1)$ null and non-null components (weighted respectively by $\hat{\pi}_0$ and $((1 - \hat{\pi}_0))$ imposed on histogram of z-scores.

9 EXAMPLE 2: COLON CANCER DATA

We apply our method next to the well-known study of Alon *et al*. (1999), where Affymetrix arrays were used to compare the gene expressions from colon cancer tissues with normal colon tissues. They measured expressions of over 6500 genes in $n_1 = 40$ tumor and $n_2 = 22$ normal colon tissue samples. The samples were taken from 40 different patients, so that 22 patients supplied both a normal and a tumor tissue sample. We consider the set of $N = 2000$ genes with highest minimal intensity across the samples as in Alon *et al*. (1999).

As in the last example, we adopted the (pooled) t-statistic as our test statistic W_j for each gene j, and so we used the t-distribution function with 60 degrees of freedom, F_{60}, as the null distribution of W_j in the computation of the P-value P_j from (12). We fitted the two-component normal mixture model (15) with the theoretical $N(0, 1)$ null to the z_j-scores, which gave the fit, $\hat{\pi}_0 = 0.39$, $\hat{\mu}_1 = 1.53$ and $\hat{\sigma}_1^2 = 2.21$. A plot of the normal mixture density and its two components is displayed in Figure 2. Our estimate of 0.39 for π_0 coincides with the empirical Bayes estimate reported in Do *et al*. (2005).

If we again consider a threshold of $c_o = 0.1$ in (5), then 433 genes would be declared to be differentially expressed with an implied FDR and FNR of 3 and 65%, respectively. The smooth muscle genes provide a positive control for differential gene expression, as the normal tissues were later found to have included smooth muscle tissue in the biopsy samples. We find that for the smooth muscle genes (J02854, T60155, M63391, D31885, X74295, X12369) each has an estimated posterior probability of non-differential expression that is <0.0015, and these genes are found within our top 66 ranked genes. For this dataset, Do *et al*. (2005) concluded that each gene in this cluster has an estimated posterior probability (of non-differential expression) of 0.002 or less.

In their original analysis, Alon *et al*. (1999) identified a ribosomal gene cluster, associated with over-expression in tumor tissues relative to normal tissues. Of these 29 genes (as listed in Table 1 of Alon *et al*. 1999), only 10 are declared differentially expressed at a threshold of c_o of 0.05 (within a subset of 296 genes). Do *et al*. (2005) similarly identified only 13 genes with posterior probabilities of non-differential expression of <0.05. Their two genes with weakest discriminatory ability were H77302 and T63484, with posterior probabilities of non-differential expression of 0.44 and 0.49; ours are H77302 and R85464, with posterior probabilities of 0.59 and 0.65. Our results, supported by the findings in Do *et al*. (2005), suggest that only a minority of the ribosomal genes are actually differentially expressed between the tumor and normal classes. We also considered the use of the empirical null instead of the theoretical $N(0, 1)$ null for this dataset. The plot of the normal mixture density is given in Figure 2 in the Supplementary information. For a threshold of $c_o = 0.1$ in (5), 253 genes would be declared to be differentially expressed with an implied FDR and FNR of 4 and 74%, respectively. The reason for fewer genes being declared differentially expressed with the use of the empirical than with the theoretical null is that the estimate of π_0 is greater ($\hat{\pi}_0 = 0.53$). According to the BIC, the empirical null would not be selected in favor of the theoretical $N(0, 1)$ null (see Supplementary Information).

10 EXAMPLE 3: HIV DATA

We consider finally the HIV dataset of van't Wout *et al*. (2003), and as discussed in Gottardo *et al*. (2005). van't Wout *et al*. used cDNA arrays to measure expression levels of $N = 7680$ genes in CD4-T-cell lines, at time $t = 24$ h after infection with the HIV-1 virus. There were four control slides (pooled mRNA from uninfected cells) and four slides run using pooled mRNA from the infected cells, so that $n_1 = 4$ and $n_2 = 4$. We column normalized the logged expression values, using the dataset as downloaded from http://hajek.stat.ubc.ca/∼raph. The dataset contains 12 HIV-1 genes which were used as positive controls, as they are known to be differentially expressed in infected versus uninfected cells. As in Efron (2004), we adopted the (pooled) *t*-statistic and used the *t* distribution with 6 degrees of freedom for F_0 in the computation of the *P*-value P_j from (12). As noted by Efron (2004, 2005b), this dataset is an example of one where the theoretical $N(0, 1)$ is inapplicable due to the underdispersion of the z_j-scores about the origin (Fig. 3). A fit of a single normal to all the z_j-scores gives an $N(-0.16, 1.06)$ distribution, clearly showing that it is not appropriate to adopt the theoretical $N(0, 1)$ null component in our two-component mixture model (15). As there are only 6 degrees available for the use of the parametric *t* assumption for the null distribution F_0 of the test statistic, we also computed the *P*-values using the permutation estimate with $B = 35$ permutations (pooled over the genes). It gave a similar fit, $N(-0.146, 0.997)$, for a single normal fitted to the z_j-scores. As explained by Efron (2005b), permutation methods in the context of this problem should be considered as a means for providing improved versions of the theoretical null rather than empirical nulls.

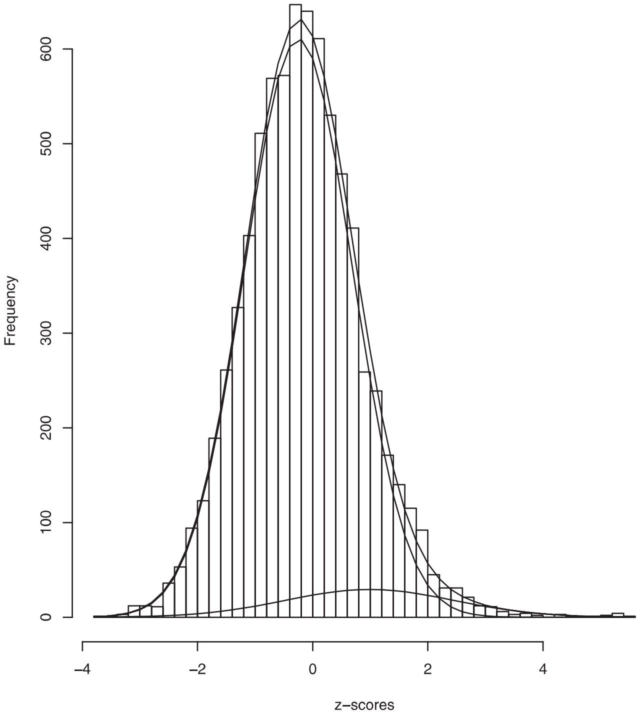

Fig. 3. HIV data: plot of fitted two-component normal mixture model with empirical null and non-null components (weighted respectively by ($\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$ imposed on histogram of *z*-scores.

We considered an empirical null for this dataset by fitting the normal mixture model (16), obtaining as our fit, $\hat{\pi}_0 = 0.93, \hat{\mu}_0 = -0.25, \hat{\sigma}_0^2 = 0.87, \hat{\mu}_1 = 0.99$ and $\hat{\sigma}_1^2 = 2.14$. The empirical null obtained is similar to the fit $\hat{\pi}_0 = 0.917$, $\hat{\mu}_0 = 0.1$ and $\hat{\sigma}_0^2 = 0.54$ obtained by Efron (2005b) for this dataset. Using a fully Bayesian approach to model the gene expressions, Gottardo *et al*. (2006) estimated π_0 to be 0.993, which is similar to our quick and easy estimate of 0.93.

On setting a threshold of $c_o = 0.01$ on the posterior probability of non-differential expression, we find that there is a subset of 15 genes that are differentially expressed, which contains the 12 genes known to be differentially expressed from external information. The implied FDR is 0.002. For $c_o = 0.1$, we obtain that 37 genes are differentially expressed with an implied FDR of 0.03.

The HIV dataset is unique in the examples discussed, as cell lines were used rather than tissue biopsies from patients. Additionally, each of the four microarray experiments within the class (uninfected or infected) used the same pooled RNA, so that each of these were technical replicates rather than individual samples. These factors may explain the decreased dispersion.

11 DISCUSSION

In this paper, we consider the problem of detecting which genes are differentially expressed in multiple classes of tissue samples, where the classes represent various clinical or experimental conditions. The available data consist of the expression levels of typically a very large number of genes for a limited number of tissues in each class. Usually, a test statistic such as the classical t in the case of two classes or the F in case of multiple classes is formed for a test of equality of the class means. The key step in this approach is to transform the observed value of the test statistic for each gene j to a z-score z_j by using the inverse standard normal distribution function of the implied P-value P_j, similar to its use in Efron (2004) and his subsequent papers on this problem. We demonstrate that a two-component normal mixture model is adequate for modelling the empirical distribution of the z_j-scores, where the first component is the standard normal, corresponding to the null distribution of the score, and the second component is a normal density with unspecified (positive) mean and variance, corresponding to the non-null distribution of the score. This model can be used to provide a straightforward and easily implemented assessment of whether a gene is null (not differentially expressed) in terms of its posterior probability of being a null gene. Estimates of this posterior probability can be easily obtained by using the EM algorithm to fit the two-component normal mixture model via ML. As there are multiple local maximizers, consideration has to be given to the choice of starting values for the algorithm. We show that the specification of an initial value $\pi_0^{(0)}$ for the proportion π_0 of null genes completely specifies a starting point for the fitting of the normal mixture model with the theoretical choice of $N(0, 1)$ as the null component. An interval of values for $\pi_0^{(0)}$ can be tried, and a guide to its endpoints is given by values of π_0 obtained by equating the number of z_j values less than a threshold ξ to the expected number under the theoretical $N(0, 1)$ null component. We consider too the case where the theoretical $N(0, 1)$ null is not tenable and an empirical null is adopted with mean and variance estimated from the data. Also, the estimation of the FDR and its control are considered, along with the estimation of other relevant rates such as the FNR. Note that it is not valid to make claims as to the relative superiority of the two models corresponding to the theoretical and empirical nulls on the basis of these error rates, as they are only valid for the model under which they were calculated. Our approach is demonstrated on three real datasets.

Concerning the choice between the use of the theoretical $N(0, 1)$ null and an empirical null, the intent in the first instance is to use the former in modelling the density of the z_j-scores. In some situations as with the HIV dataset above, it will be clear that the use of the theoretical null is inappropriate. In other situations, an informed choice between the theoretical and empirical null components can be made on the basis of the increase in the log likelihood due to the use of an empirical null with its two extra parameters. For this purpose we can use BIC as in the first two examples above.

The reliability of our approach obviously depends on how well the proposed two-component normal mixture model approximates the empirical distribution of the z_j-scores. In the three examples here and in our analyses of other datasets not presented here, this normal mixture model provides an excellent approximation. Its fit can be assessed either by visual inspection of a plot of the fitted normal mixture density versus a histogram of the z_j-scores or, more formally, by a likelihood ratio test for the need for an additional normal density to represent the non-null distribution of the z_j-scores. On a similar note on the adequacy of a two-component normal mixture model, Pounds and Morris (2003) found that a two-component mixture of the uniform (0,1) distribution and a single beta component (with one unspecified unknown parameter) was adequate to model the distribution of the P-values in their analyses. However, it is advantageous to work as proposed here in terms of the z_j-scores, which can be modelled by normal components on the real line rather than working in terms of the P-values.

Finally, we should mention explicitly that it has been assumed throughout that the genes are all independently distributed. Typically in practice, this independence assumption will not hold for all the genes. As cautioned by Qiu *et al*. (2005), care is needed in extrapolating results valid in the case of independence to dependent gene data. In the Supplementary information, we report the results of a few initial simulations that we have performed to investigate the effect of correlation between some of the genes on the distribution of the z-scores. For moderate correlation, the effect was small, while for strong correlation it was found that the use of the empirical null in place of the theoretical $N(0, 1)$ null avoided any marked effect (see Figs 3 and 4 in the Supplementary information). Another approach to handle the effect of strong correlation between the genes would be first to cluster the (normalized) gene profiles on the basis of Euclidean distance, so that highly correlated genes are put in the same cluster. The genes in each of those clusters with highly correlated members can then be represented by a single gene or a linear combination of the member genes (a metagene). To incorporate a priori knowledge that some genes belong to the same pathway, we can take the cluster labels for these genes to be the same in the specification of the complete data in the EM framework for the problem.

Conflict of Interest: none declared.

REFERENCES

Allison,D.B. *et al*. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal*., **39**, 1–20.

Alon,U. *et al*. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

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

Broët,P. *et al*. (2004) A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*, **20**, 2562–2571.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B*, **39**, 1–38.

Do,K-A. *et al*. (2005) A Bayesian mixture model for differential gene expression. *Appl. Stat*., **54**, 627–644.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc*., **99**, 96–104.

Efron,B. (2005a) Selection and estimation for large-scale simultaneous inference. *Technical Report*. Department of Statistics, Stanford University, Stanford, CA, http://www-stat.stanford.edu/brad/papers/Selection.pdf.

Efron,B. (2005b) Local false discovery rates. *Technical Report*. Department of Statistics, Stanford University, Stanford, CA, http://www-stat.stanford.edu/brad/papers/False.pdf

Efron,B. and Tibshirani,R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol*., **23**, 70–86.

Efron,B. *et al*. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc*., **96**, 1151–1160.

Gottardo,R. *et al*. (2006) Bayesian robust inference for differential gene expression in cDNA microarrays with multiple samples. *Biometrics*, **62**, 10–18.

Guo,X. and Pan,W. (2005) Using weighted permutation scorse to detect differential gene expression with microarray data. *J. Bioinformatics Compat. Biol*., **3**, 989–1006.

Hedenfalk,I. *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med*., **344**, 539–548.

Lee,M.-L.T. *et al*. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9838.

Lönnstedt,I. and Speed,T. (2002) Replicated microarray data. *Statist. Sinica*, **12**, 31–46.

McLachlan,G.J. and Krishnan,T. (1997) *The EM Algorithm and Extensions*. Wiley, New York.

McLachlan,G.J. and Peel,D. (2000) *Finite Mixture Models*. Wiley, New York.

Newton,M.A. *et al.* (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

McLachlan,G.J. *et al.* (2004) *Analyzing Microarray Gene Expression Data*. Wiley, Hoboken, NJ.

Newton,M.A. *et al.* (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Pan,W. (2003) On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*, **19**, 1333–1340.

Pan,W. *et al.* (2003) A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics*, **3**, 117–124.

Pawitan,Y. *et al.* (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017–3024.

Ploner,A. *et al.* (2006) Multidimensional local false discovery rate for microarray studies. *Bioinformatics*, **22**, 556–565.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Informatics*, **19**, 1236–1242.

Qiu,X. *et al.* (2005) Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol.*, **4**, No. 1, Article 34..

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, *No. 1*, Article 3..

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser B*, **64**, 479–498.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

van't Wout,A.B. *et al.* (2003) Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4^{+}-T-cell lines. *J. Virol.*, **77**, 1392–1402.

Wilson,E.B. and Hilferty,M.M. (1931) The distribution of chi-square. *Proc. Nat. Acad. Sci. USA*, **28**, 94–100.

Xie,Y. *et al.* (2005) A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, **21**, 4280–4288.

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