*STATISTICA, anno LXVIII, n. 1, 2008*

# LARGE-SCALE SIMULTANEOUS INFERENCE WITH APPLICATIONS TO THE DETECTION OF DIFFERENTIAL EXPRESSION WITH MICROARRAY DATA

G.J. McLachlan, K. Wang, S.K. Ng

## 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, $C_1, ..., C_g$. 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, as proposed initially by McLachlan, Bean, and Ben-Tovim Jones (2006). This model is becoming more widely adopted in the context of microarrays, where one component density corresponds to that of the test statistics 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, its values are transformed to $z$-scores, whose null and non-null distributions can be represented by a single normal each. We explain 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 $\pi_0$ of genes that are not differentially expressed. There is an easy to apply procedure for determining suitable initial values for $\pi_0$ 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 the approach on a data set that has been analyzed previously in the bioinformatics literature.

In the above formulation of the problem, it is assumed that there is a nonzero proportion of the genes that are differentially expressed. We shall consider also an example where there would appear to be no differentially expressed genes. Hence it is advised in general that one should in the first instance carry out a test

of a single normal distribution versus a mixture of two normal components; that is, a test of an empirical null only versus a mixture of an empirical null and non-null normal component.

## 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 expreriments (corresponding to $m$ tissue samples), where we measure the expression levels of $N$ genes in each experiment, the results can be represented by $N \times m$ 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 expressions*

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 oneway 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 oneway 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 nondifferential expression*

In this paper, we focus on a decision-theoretic approach to the problem of finding genes that are differentially expressed, as proposed in McLachlan, Bean, and Ben-Tovim Jones (2006). Their approach is 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 $\pi_i$ 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 $j$th gene is not differentially expressed (that is, belongs to $G_0$) is given by

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

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 nondifferential expression in practice, we need to be able to estimate $\pi_0$, 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) has developed a simple empirical Bayes approach to this problem with minimal assumptions. This problem has been studied since under more specific assumptions, including the work by Newton *et al.* (2001, 2004), Lönnstedt and Speed (2002), Pan et al. (2002), 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 $\pi_0$ 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 the estimation problem. If $\hat{\pi}_0$, $\hat{f}_0(w_j)$, and $\hat{f}_1(w_j)$ denote estimates of $\pi_0$, $f_0(w_j)$, 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

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

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 $c_0$ or by taking the top $N_0$ genes in the ranked list.

### 4.1. *FDR*

Suppose that we select all genes with

$$\hat{\pi}_0(w_j) \leq c_0 . \tag{5}$$

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

$$\widehat{\text{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j) I_{[0,c_0]}(\hat{\tau}_0(w_j))/N_r , \tag{6}$$

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 the zero otherwise.

Similarly, the false nondiscovery rate (FNDR) can be estimated by

$$\widehat{\text{FNDR}} = \sum_{j=1}^{N} \hat{\tau}_1(w_j) I_{[c_0,\infty]}(\hat{\tau}_0(w_j))/(N - N_r) . \tag{7}$$

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

$$\widehat{\text{FPR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j) I_{[0,c_0]}(\hat{\tau}_0(w_j))/\sum_{j=1}^{N} \hat{\tau}_0(w_j) \tag{8}$$

and

$$\widehat{FNR} = \sum_{j=1}^{N} \hat{\tau}_1(w_j) I_{(c_0,\infty)}(\hat{\tau}_0(w_j))/\sum_{j=1}^{N} \hat{\tau}_1(w_j) \tag{9}$$

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 the 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 nondifferential 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).

## 5. USE OF *Z*-SCORES

### 5.1. *Normal transformation*

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

$H_j$: $j$th gene is the not differentially expressed. (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 follow McLachlan *et al.* (2006) and proceed in a similar manner as in Efron (2004) to 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 $\Phi$ 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 $z_j$. The 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 chi-squared 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. 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

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 (McLachlan *et al.*, 2006). 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 $\mu$ and variance $\sigma^2$, 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 data sets that we have analysed, 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, 2005b), for some microarray data sets 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 $\mu_0$ and variance $\sigma_0^2$ 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 describe 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 $\pi_0$, $\mu_1$, and $\sigma_1^2$. This is effected by maximum likelihood 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, it is noted that the maximum likelilhood (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 $\sigma_0^2 = 1$ 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 $\pi_0$, initial values for the other parameters to be estimated, $\mu_1$ and $\sigma_1^2$, 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 $\pi_0$, 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 $\xi$. There is an inherent bias-variance trade-off in the choice of $\xi$. In most cases as $\xi$ grows larger, the bias of $\hat{\pi}_0^{(0)}(\xi)$ grows larger, but the variance becomes smaller.

### 7.2. *Empirical null*

In this case, we do not assume that the mean $\mu_0$ and variance $\sigma_0^2$ of the null distribution are zero and one, respectively, but rather they are estimated in addition to the other parameters $\pi_0$, $\mu_1$, and $\sigma_1^2$. For an initial value $\pi_0^{(0)}$ for $\pi_0$, 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.

## 8. EXAMPLE: BREAST CANCER DATA

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 data set comprised the measurement of $N = 3,226$ genes using cDNA arrays, for $n_1 = 7$ BRCA 1 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 $z_i$-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 $(1\text{-}\hat{\pi}_0)$. It can be seen that this two-component normal mixture model gives a good fit to the empirical distribution of the $z_i$-scores.

In Table 1, we have listed the FDR estimated from (6) for various levels of the threshold $c_0$ in (5). It can be seen, for example, that if $c_0$ 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 $c_0$ is discussed in Efron (2005*b*).

TABLE 1

*Estimated FDR and other error for various levels of the thresohold $c_0$ applied to the posterior probability of nondifferential expression for the breast cancer data, where $N_r$ is the number of selected genes (with theoretical null)*

| $c_0$ | $N_r$ | $\widehat{\text{FDR}}$ | $\widehat{\text{FNDR}}$ | $\widehat{\text{FNR}}$ | $\widehat{\text{FPR}}$ |
|---|---|---|---|---|---|
| 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 | 743 | 0.21 | 0.22 | 0.48 | 0.08 |
| 0.5 | 976 | 0.27 | 0.18 | 0.37 | 0.13 |

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 data set, 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 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.

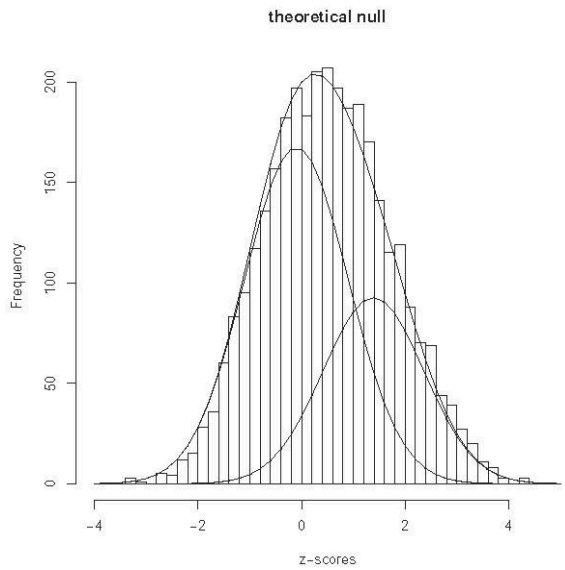


*Figure 1* – Breast cancer data: plot of fitted two-component normal mixture model with theoretical $N(0, 1)$ null and non-null components (weighted respectively by $\hat{\pi}_0$ and (1 - $\hat{\pi}_0$)) imposed on histogram of $z$-scores.

Among other analyses of this data set, $\pi_0$ 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).

Concerning the other type of allocation rates for the choice of $c_0 = 0.1$ (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). Analogous to the miss rate of Taylor *et al.* (2003), we might wish to have an idea of how many false negatives there would be in, say, the next best 57 genes with estimated posterior probability of nondifferential expression greater than $c_0 = 0.1$, which takes one down to the 200th best ranked gene. We can obtain an estimate of this quantity by finding the average of the $\hat{\tau}_1(z_j)$ values for these next 57 genes. In the case of $c_0 = 0.1$, it is 0.89, implying that among the 57 next best genes (all declared to be null genes), approximately 51 are actually non-null.

We also considered the fitting of the two-componet normal mixture model (16) with the null component mean and variance, $\mu_0$ and $\sigma_0^2$, now estimated in addition to $\pi_0$ and the non-null mean and variance , $\mu_1$ and $\sigma_1^2$. As can be seen

from Figure 2, the fit from using the empirical null in place of the $N(0, 1)$ theoretical null is similar to the fit in Figure 1.

In other analyses of this data set, 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 if a thresold of 0.2 is imposed on the posterior probability of nondifferential expression for which the implied FDR is 11% and the FNR is 73%. The corresponding values with the use of the empirical null can be from Table 2 to be 13% and 77% for the FDR and FNR, respectively, with 212 genes declared to be differentially expressed.

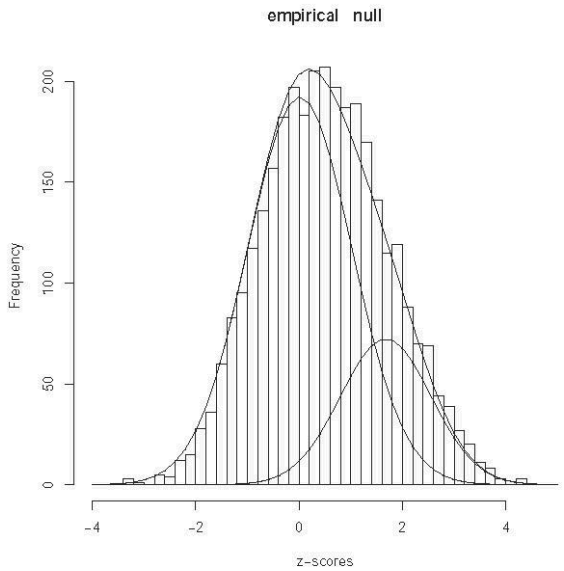


*Figure 2* – Breast cancer data: plot of fitted two-component normal mixture model with empirical null and non-null components (weighted respectively by $\hat{\pi}_0$ and (1 - $\hat{\pi}_0$)) imposed on histogram of $z$-scores.

TABLE 2

*Estimated FDR and other error rates for various levels of the thresohold $c_0$ applied to the posterior probability of nondifferential expression for the breast cancer data, where $N_r$ is the number of selected genes (with empirical null)*

| $c_0$ | $N_r$ | $\widehat{\text{FDR}}$ | $\widehat{\text{FNDR}}$ | $\widehat{\text{FNR}}$ | $\widehat{\text{FPR}}$ |
|---|---|---|---|---|---|
| 0.1 | 62 | 0.07 | 0.23 | 0.93 | 0.00 |
| 0.2 | 212 | 0.13 | 0.20 | 0.77 | 0.01 |
| 0.3 | 343 | 0.17 | 0.18 | 0.64 | 0.02 |
| 0.4 | 504 | 0.23 | 0.15 | 0.51 | 0.05 |
| 0.5 | 644 | 0.28 | 0.13 | 0.41 | 0.07 |

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 $\pi_0$ is greater ($\hat{\pi}_0 = 0.76$). According to the Bayesian Information criterion (BIC), the

empirical null would not be selected in favour of the theoretical $N(0, 1)$ null. The same decision was reached too after we adopted a resembling approach (McLachlan, 1987) to carry out a formal test of a theoretical versus empirical null, using the likelihood ratio test statistic.

## 9. CLUSTER ANALYSIS APPROACH

Another approach to this problem would be make to more assumptions and model the expression level for each gene. Then we can use the model-based procedure EMMIX-WIRE of Ng *et al.* (2006) to cluster the gene profiles. More specifically, we let

$$\boldsymbol{y}_j = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T)^T \tag{22}$$

denote the expression profile for the $j$th gene, where

$$\boldsymbol{y}_i = (y_{ij1}, ..., y_{ijm_i})^T$$

denotes the vector containing the $m_i$ expression levels of the $j$th gene in Class $i$ ($i = 1, 2$). That is, $y_{ijk}$ denotes the expression level of the $j$th gene in the $k$th microarray experiment in the $i$th Class ($i = 1, 2$; $j = 1, ..., N$; $k = 1, ..., m_i$), and $m = m_1 + m_2$.

We model the distribution of the profile vector $\boldsymbol{y}_j$ for the $j$th gene by a $g$-component mixture with each component specified by a linear mixed model. Conditional on its membership of the $h$th component of the mixture, we assume that $\boldsymbol{y}_j$ follows a linear mixed-effects model (LMM),

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta}_h + \boldsymbol{U}\boldsymbol{b}_{hj} + \boldsymbol{V}\boldsymbol{c}_h + \boldsymbol{\varepsilon}_{hj}, \tag{23}$$

where $\boldsymbol{\beta}_h = (\beta_{h1}, \beta_{h2})^T$ is the vector of fixed effects ($h = 1, ..., g$). In (23), $\boldsymbol{b}_{hj} = (b_{h1j}, b_{h2j})^T$ and $\boldsymbol{c}_h$ (a $m$-dimensional vector ) represent the unobservable gene- and cluster-specific random effects, respectively, conditional on membership of the $h$th cluster. The random effects $\boldsymbol{b}_h$ and $\boldsymbol{c}_h$, and the measurement error vectors $(\boldsymbol{\varepsilon}_{h1}^T, ..., \boldsymbol{\varepsilon}_{hm}^T)^T$ are assumed to be mutually independent, where $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are known design matrices of the corresponding fixed or random effects. Here the design matrices $\boldsymbol{X}$ and $\boldsymbol{U}$ are taken to be equal to the $m \times 2$ matrix with the first $m_1$ rows equal to (1, 0) and the next $m_2 = m - m_1$ rows equal to (0, 1), and $\boldsymbol{V}$ is equal to $\boldsymbol{I}_m$, where the latter denotes the $m \times m$ identity matrix. The presence of the random effect $\boldsymbol{c}_h$ for the expression levels of genes in the $h$th component induces a correlation between the profiles of genes within the same cluster.

With the LMM, the distributions of $\boldsymbol{b}_{hj}$ and $\boldsymbol{c}_h$ are taken, respectively, to be multivariate normal $N_2(\boldsymbol{0}, \boldsymbol{B}_h)$ and $N_m(\boldsymbol{0}, \theta_{ch}\boldsymbol{I}_m)$, where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. The presence of the random effect term $\boldsymbol{b}_{hj}$ is to allow for correlation between the tissue samples. If the covariance matrix $\boldsymbol{B}_h$ is diagonal, then it implies that the ex-

pression levels of a gene in different classes are uncorrelated. In an ideal experiment, one would hope that there would be no correlations between the tissue samples, and we could dispense with this random effects term $\boldsymbol{b}_{hj}$ in the model.

The measurement error vector $\boldsymbol{\varepsilon}_{hj}$ is also taken to be multivariate normal $N_m(\mathbf{0}, \boldsymbol{A}_m)$, where $\boldsymbol{A}_h = \text{diag}(\boldsymbol{H}\boldsymbol{\phi}_h)$ is a diagonal matrix constructed from the vector $(\boldsymbol{H}\boldsymbol{\phi}_h)$, where here $\boldsymbol{H} = \boldsymbol{X}$ and $\boldsymbol{\phi}_h = (\sigma_{h1}^2, \sigma_{h2}^2)^T$. That is, we allow the $h$th component variance to be different among the two classes of microarray experiments.

The vector $\boldsymbol{\Psi}$ of unknown parameters can be obtained by maximum likelihood via the EM algorithm, proceeding conditionally on the cluster-specific random effects $\boldsymbol{c}_i$. The E- and M-steps can be implemented in closed form. In particular, an approximation to the E-step by carrying out time-consuming Monte Carlo methods is not required. A probabilistic or an outright clustering of the genes into $g$ components can be obtained, based on the estimated posterior probabilities of component membership given the profile vectors and the estimated cluster-specific random effects $\hat{\boldsymbol{c}}_h (h = 1, ..., g)$.

Before we cluster the gene profiles, we normalized the expression levels in each gene profile so that they have mean zero and standard deviation one. With this normalization of the gene profiles, we fit a $g = 3$ component mixture model, where we let $\beta_{h1}$ and $\beta_{h2}$ denote the fixed effects for the means of the two classes. The clustering of the gene profiles is not invariant under this normalization, but in our experience, it has proved to be a reasonable way to proceed. With this normalization, the intent to find three clusters where (a) for one cluster, the estimate of the fixed effects for the two class means are approximately zero, ($\beta_{11} \approx \beta_{12} \approx 0$), corresponding to the genes that are not differentially expressed; (b) for a second cluster, $\hat{\beta}_{21} < \hat{\beta}_{22}$, corresponding to genes that (before normalization) are upregulated more in Class $C_1$ than in Class $C_2$; (c) for a third cluster, $\hat{\beta}_{31} < \hat{\beta}_{32}$, corresponding to genes that are downregulated more in Class $C_1$ than in Class $C_2$.

On fitting EMMIX-WIRE to the normalized gene profiles, we obtained three clusters in proportions $\hat{\pi}_1 = 0.63$, $\hat{\pi}_2 = 0.14$, $\hat{\pi}_3 = 0.23$ with $\hat{\beta}_1 = (0.06, -0.05)^T$, $\hat{\beta}_2 = (0.56, -0.49)^T$, and $\hat{\beta}_3 = (-0.42, 0.37)^T$. If we take the genes in the first cluster to be the null genes, then our estimate of the proportion of null genes is $\hat{\pi}_0 = 0.63$, which is in general agreement with that obtained above using the $z$-scores.

In Table 3, we have listed the FDR estimated from (6) for various levels of the threshold $c_0$. On comparing this table with Tables 1 and 2, it can be seen that for approximately the same FDR level, we declare more genes to be differentially expressed but with a lower FNR by working with the full data (the gene profiles) rather than the profiles in reduced form as summarized by their $z$-scores. However, the validity of this approach in modelling the full data obviously depends on much stronger distributional assumptions.

TABLE 3

*Estimated FDR and other error rates for various levels of the threshold $c_0$ applied to the posterior probability of nondifferential expression for the breast cancer data, where $N_r$ is the number of selected genes: clustering approach*

| $c_0$ | $N_r$ | $\widehat{\text{FDR}}$ | $\widehat{\text{FNDR}}$ | $\widehat{\text{FNR}}$ | $\widehat{\text{FPR}}$ |
|---|---|---|---|---|---|
| 0.1 | 257 | 0.06 | 0.32 | 0.79 | 0.01 |
| 0.2 | 480 | 0.10 | 0.27 | 0.63 | 0.02 |
| 0.3 | 678 | 0.14 | 0.24 | 0.51 | 0.05 |
| 0.4 | 854 | 0.18 | 0.20 | 0.41 | 0.08 |
| 0.5 | 1048 | 0.23 | 0.17 | 0.32 | 0.12 |

## 10. RESULTS FOR A DIFFERENT VERSION OF THE HEDENFALK DATA

Efron (2004) writes that "there is ample reason to distrust the theoretical null" in the case of the Hedenfalk data, whereas above we have found that the theoretical and empirical null distributions are similar to each other. 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). Thus, the breast cancer data of Hedenfalk *et al.* (2001) that we have analysed above is not the same as analysed in the papers of Efron (2004, 2005a, 2005b).

In Figure 3, we display the histogram of the $z$-scores as obtained by Efron (2004) for this data set, along with the $N(0, 1)$ distribution and the $N(0.05, 2.05)$ distribution with mean and variance equal to the sample mean and variance of the $z$-scores. His $z$-score is defined to be

$$z_j = \mathbf{\Phi}^{-1}(F_{13}(t_j)), \tag{24}$$

where $t_j$ is the pooled two-sample $t$-statistic and $F_{13}$ is its distribution, which is the $t$-distribution with 13 degrees of freedom. Thus non-null genes can have either large positive or large negative values for $z$-scores. If we use the "empirical" distribution $N(0.05, 2.05)$ as the null distribution on its own (without a non-null component) then it can be seen from Figure 3 that no genes would be declared to be differentially expressed.

We now consider the two-component mixture normal approach applied to the same data as analysed in the papers of Efron. We did this by converting his two-sided $z$-scores to our one-sided ones. But before we considered fitting a two-component normal mixture to the latter, we need to address the question of whether we really need a non-null component in our model; that is, whether there are any genes that are differentially expressed ($\pi_0 = 1$). We therefore carried out a test of a single normal distribution with unspecified mean and variance (empirical null) versus a mixture of an empirical null and a non-null component. It was found in accordance with the conclusions of Efron that a single normal distribution suffices.

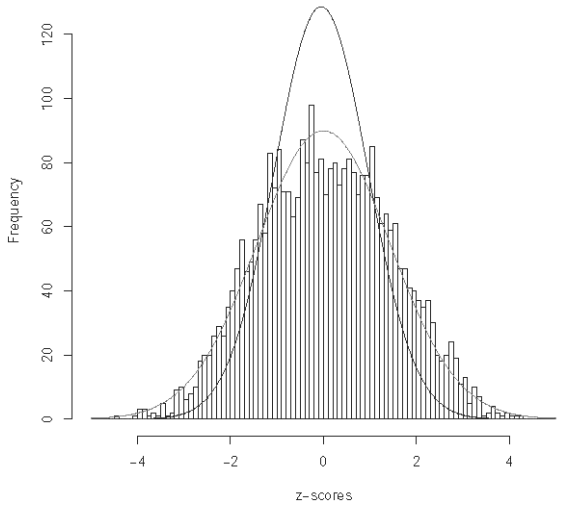


*Figure 3* – Breast cancer data: plot of $N(0, 1)$ distribution and $N(0.05, 2.05)$ imposed on the histogram of $z$-scores as analyzed in Efron's papers.

## 11. DISTRIBUTION

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. Typically, 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 staightforward 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 maximum likelihood. 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 $\pi_0$ of null genes completely specifies a starting point for the fitting of

the normal mixture model with the teoretical 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 $\pi_0$ obtained by equating the number of $z_j$ values less than a threshold $\xi$ 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 the mean and the variance estimated from the data. Also, the estimation of the false discovery rate and its control are considered, along with the estimation of other relevant rates such as the false negative rate. 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.

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, 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 or a resampling approach to assess the *P*-value of a formal test based on the likelihood ratio test statistic. Recent results of the authors suggest that the latter approach is preferable to the use of BIC in this context.

In the version of the Hedenfalk data as analysed the papers by Efron, it appears that there are no genes that are differentially expressed. Hence in general before we proceed to fit a two-component normal mixture model with either a theoretical or an empirical null, the question of whether a single normal distribution is adequate needs to be considered first in situations where it is not obvious that there are some genes present that are differentially expressed.

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. 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 the adoption of the standard normal for the null distribution is equivalent to assuming 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.

*Department of Mathematics* GEOFF. J. MCLACHLAN
*University of Queensland, Australia*

*and ARC Centre of Excellence in Bioinformatics, Institute for Molecular Bioscience, University of Queensland, Australia*

*Department of Mathematics* KENT WANG
*University of Queensland, Australia* SHU KAY NG

## REFERENCES

D.B. ALLISON, G.L. GADBURY, M. HEO, J.R. FERNANDEZ, C.-K. LEE, T.A. PROLLA and R. WEINDRUCH (2002), *A mixture model approach for the analysis of microarray gene expression data.*, "Computational Statistics & Data Analysis", 39, pp. 1-20.

Y. BENJAMINI and Y. HOCHBERG (1995), *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, "Journal of the Royal Statistical Society", *B*57, pp. 289-300.

P. BROËT, A. LEWIN, S. RICHARDSON, C. DALMASSO and H. MAGDELENAT (2004), *A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments*, "Bioinformatics", 20, pp. 2562-2571.

A.P. DEMPSTER, N.M. LAIRD and D.B. RUBIN (1977), *Maximum likelihood from incomplete data via the EM algorithm (with discussion).* "Journal of the Royal Statistical Society", *B*39, pp. 1-38.

K.-A. DO, P. MÜLLER and F. TANG (2005), *A Bayesian mixture model for differential gene expression.* "Applied Statistics", 54, 627-644.

B. EFRON (2004), *Large-scale simultaneous hypothesis testing: the choice of a null hypothesis,* "Journal of the American Statistical Association", 99, 96-104.

B. EFRON (2005a), *Selection and Estimation for Large-Scale Simultaneous Inference,* "Technical Report", Stanford, CA: Department of Statistics, Stanford University, http://www-stat.stanford.edu/~brad/papers/Selection.pdf.

B. EFRON (2005b), *Local False Discovery Rates.* "Technical Report", Stanford, CA: Department of Statistics, Stanford University, http://www-stat.stanford.edu/~brad/papers/False.pdf.

B. EFRON, R. TIBSHIRANI (2002), *Empirical Bayes methods and false discovery rates for microarrays,* "Genetic Epidemiology.", 23, pp. 70-86.

B. EFRON, R. TIBSHIRANI, J.D. STOREY and V.G. TUSHER (2001), *Empirical Bayes analysis of a microarray experiment,* "Journal of the American Statistical Association", 96, pp. 1151-1160.

R. GOTTARDO, A.E. RAFTERY, K.Y. YEUNG and R.E. BUMGARNER (2006), *Bayesian robust inference for differential gene expression in cDNA microarrays with multiple samples,* "Biometrics", 62, to appear.

X. GUO, W. PAN (2005), *Using weighted permutation scorse to detect differential gene expression with microarray data,* "Journal of Bioinformatics and Computational Biology", 3, pp. 989-1006.

I. HEDENFALK *et al.* (2001), *Gene-expression profiles in hereditary breast cancer*, "The New England Journal of Medicine", 344, pp. 539-548.

M.-L.T. LEE, F.C. KUO, G.A. WHITMORE and J. SKLAR (2000), *Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations,* "Proceedings of the National Academy of Science", *USA* 97, pp. 9834-9838.

I. LÖNNSTEDT, T. SPEED (2002) *Replicated microarray data*, "Statistica Sinica", 12, pp. 31-46.

G.J. MCLACHLAN (1987), *On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture,* "Applied statistics", 36, pp. 318-324.

G.J. MCLACHLAN, R.W. BEAN and L. BEN-TOVIM JONES (2006), *A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays,* "Bioinformatics", 22, pp. 1608-1615.

G.J. MCLACHLAN, K.-A. DO and C. AMBROISE (2004), *Analyzing Microarray Gene Expression Data,* Wiley, Hoboken, New Jersey.

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

G.J. MCLACHLAN, D. PEEL (2000), *Finite Mixture Models,* Wiley, New York.

M.A. NEWTON, C.M. KENDZIORSKI, C.S. RICHMOND, F.R. BLATTNER and K.W. TSUI (2001), *On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data,* "Journal of Computational Biology", 8, pp. 37-52.

M.A. NEWTON, A. NOUEIRY, D. SARKAR and P. AHLQUIST (2004), *Detecting differential gene expression with a semiparametric hierarchical mixture method,* "Bioinformatics", 5, pp. 155-176.

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

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

W. PAN, J. LIN and C.T. LE (2003), *A mixture model approach to detecting differentially expressed genes with microarray data. Model-based cluster analysis of microarray gene-expression data,* "Genome Biology", 3, research0009.1-0009.8.

Y. PAWITAN, S. MICHIELS, S. KOSCIELNY, A. GUSNANTO and A. PLONER (2005), *False discovery rate, sensitivity and sample size for microarray studies,* "Bioinformatics", 21, pp. 3017-3024.

A. PLONER, S. CALZA, A. GUSNANTO and Y. PAWITAN (2006), *Multidimensional local false discovery rate for microarray studies,* "Bioinformatics", 22, pp. 556-565.

S. POUNDS, S.W. MORRIS (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, pp. 1236-1242.

X. QIU, L. KLEBANOV and A. YAKOVLEV (2003), *Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes,* "Statistical Applications in Genetics and Molecular Biology", 4, n. 1, Article 34.

G.K. SMYTHY (2004), *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments,* "Statistical Applications in Genetics and Molecular Biology", 3, n. 1, Article 3.

J.D. STOREY (2002), *A direct approach to false discovery rates,* "Journal of the Royal Statistical Society", *B* 64, pp. 479-498.

J.D. STOREY, R. TIBSHIRANI (2003), *Statistical significance for genome-wide studies,* "Proceedings of the National Academy of Sciences", *USA* 100, pp. 9440-9445.

J. TAYLOR, R. TIBSHIRANI and B. EFRON (2005), *The 'miss rate' for the analysis of gene expression data,* "Biostatistitics", 6, pp. 111-117.

V.G. TUSHER, R. TIBSHIRANI and G. CHU (2001), *Significance analysis of microarrays applied to the ionizing radiation response,* "Proceedings of the National Academy of Sciences", *USA* 98, pp. 5116-5121.

A.B. VAN'T WOUT *et al.* (2003), *Cellular gene expression upon human immunodeficiency virus type 1 infection of $CD4^+$-T-cell linear,* "Journal of Virology", 77, pp. 1392-1402.

E.B. WILSON and M.M. HILFERTY (1931), *The distribution of chi-square,* "Proceedings of the National Academy of Sciences", *USA* 28, pp. 94-100.

Y. XIE, W. PAN and A.B. KHODURSKY (2005), *A note on using permutation-based false discovery rate estimate to compare different analysis methods for microarray data,* "Bioinformatics", 21, pp. 4280-4288.

Y. ZHAO, W. PAN (2003), *Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments,* "Bioinformatics", 19, pp. 1046-1054.

SUMMARY

*Large-scale simultaneous inference with applications to the detection of differential expression with microarray data*

An important problem in microarray experiments is the detection of genes that are differentially expressed in agiven mumber of classes. We consider 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. By converting to a $z$-score the value of the test statistic used to test the significance of each gene, we can use a simple two-component normal mixture to model adequately the distribution of this score. In the context of the application of this approach to a well known breast cancer data set, we consider some of the issues associated with the problem of the detection of differential expression, including the case where there is need for the use of an empirical null distribution in place of the standard normal (the theoretical null) and the case where none of the genes might be differentially expressed. We also describe briefly some initial results on a cluster analysis approach to this problem, which attempts to model the joint distribution of the individual gene expressions. This latter approach thus has to make distributional assumptions which are note necessary with the former approach based on the $z$-scores. However, in the case where the distributional assumptions are valid, it has the potential to provide a more powerful analysis.