

Supplementary text for “A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays”

G.J. McLachlan, R.W. Bean, L. Ben-Tovim Jones

1 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, we can obtain an assessment of the P -value P_j via permutation methods. Using just permutations of the class labels for the gene-specific statistic W_j , the P -value for $W_j = w_j$ is assessed as

$$P_j = \frac{\#\{b : |w_{0j}^{(b)}| \geq |w_j|\}}{B},$$

where $w_{0j}^{(b)}$ is the null version of w_j after the b th permutation of the class labels ($b = 1, \dots, B$).

This suffers from a granularity problem, since it estimates the P -value with a resolution of only $1/B$. If we pool over all N genes, then

$$P_j = \sum_{b=1}^B \frac{\#\{b : |w_{0j}^{(b)}| \geq |w_j|, b = 1, \dots, B\}}{NB}$$

The drawback of pooling the null statistics $w_{0j}^{(b)}$ 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).

2 BREAST CANCER DATA

2.1 Empirical Error Rates

Our software package also gives the predicted values of the numbers a, b, c , and d as displayed in Table 1 below. In the present case of $c_o = 0.1$, the predicted values of a, b, c and d are equal to 2094.6, 8.8, 988.4, and 134.2 respectively. The empirical values of the FDR, FNDR, FNR, and FPR can be expressed in terms of these predicted values. For example, the FDR is given by $b/(b + d)$, the FNDR by $c/(a + c)$, and the FNR and FPR by $c/(c + d)$ and $b/(a + b)$, respectively. These empirical rates will be slightly different to those obtained using the formulas above, since the former are based on outright assignment whereas the latter use soft allocation to the groups of differentially (nondifferentially) expressed genes.

2.2 Empirical Null

For this breast cancer data set, we also considered the fitting of the two-component normal mixture model with the null component

Table 1. True versus assigned allocations with respect to the group G_0 of nondifferentially expressed genes and to the group G_1 of differentially expressed genes

	Assigned	
	Null	Non-null
True	Null	a
	Non-null	b
		c
		d

mean and variance, μ_0 and σ_0^2 , estimated in addition to π_0 and the non-null mean and variance, μ_1 and σ_1^2 . We found that this fit with the empirical null in place of the $N(0, 1)$ theoretical null is given by

$$\hat{\pi}_0 = 0.73, \hat{\mu}_0 = 0.09, \hat{\sigma}_0^2 = 1.02, \hat{\mu}_1 = 1.69, \hat{\sigma}_1^2 = 0.82.$$

This fit for the two-component normal mixture density is displayed below, along with the empirical null and non-null normal component densities weighted by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$, respectively. It can be seen that it is similar to the fit using the theoretical null distribution, as given in Figure 1 in the paper. Besides this visual comparison of the fitted normal mixture densities with theoretical or empirical null component, we can perform the likelihood ratio test. That is, calculate twice the increase in the log likelihood in adopting the empirical null over the theoretical $N(0, 1)$ null. Here this increase is only 0.724, which is less than the value of $2 \log n = 5.4$, as required to select the empirical null, using BIC (Bayesian information criterion).

2.3 GO Terms for Uniquely Identified Significant Genes

Of the 143 genes identified by us, 85 (60%) are over-expressed in BRCA1-mutation-positive tumours relative to BRCA2. This supports the findings of Hedenfalk et al. (2001) and also Storey and Tibshirani (2003) where many genes were found to be over-expressed in BRCA1-mutation-positive tumours, in particular those involved in DNA repair and inducing apoptosis. In their paper, Hedenfalk et al. (2001) noted that this suggests that the BRCA1 mutation, known to be associated with high grade often oestrogen receptor negative cancers, leads to a constitutive stress-type state.

On comparing our 143 genes with the 160 identified by Storey and Tibshirani (2003), we found that there were 113 genes in common. Of the 30 excluded genes, 6 were included in the Hedenfalk (2001) set of 176. The GO terms (where known) of the remaining 24 genes are shown in Table 2, and interestingly include several genes involved in DNA repair, cell cycle control and cell death. For example DDB2 (damage specific DNA-binding protein 2) is found to be

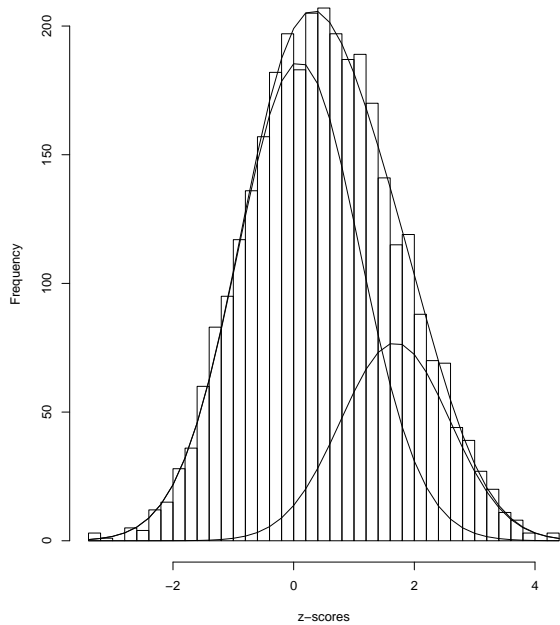


Fig. 1. Breast cancer data: plot of fitted two-component normal mixture model with empirical null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$) imposed on histogram of z -scores.

under-expressed in BRCA1-mutation-positive tumours. This gene is known to be upregulated by BRCA1. The mechanism for DDB2 activation by BRCA1 has conflicting reports in the literature, with evidence for p53-independent activation, as well as other results which suggest that p53 is directly involved with DDB2 activation. Irrespective of this, our findings support that DDB2 up-regulation is suppressed in the BRCA1-mutation.

As mentioned above, we find genes known to induce apoptosis (such as PDC5), to be over-expressed in BRCA1-mutation-positive tumours, supporting the findings of Hedenfalk et al. (2001). On the other hand, we find that HDAC3 and MIF, both suppressors of apoptosis, are under- and over-expressed respectively in BRCA1-mutation-positive tumours. The finding of a negative regulator of apoptosis (MIF) over-expressed in BRCA1-mutation-positive tumours, suggests that there may be more complex mechanisms than accounted for in Hedenfalk et al. (2001).

3 COLON CANCER DATA

For the colon cancer data set, we also considered the fitting of the two-component normal mixture model with the null component mean and variance, μ_0 and σ_0^2 , estimated in addition to π_0 and the non-null mean and variance, μ_1 and σ_1^2 . We found that this fit with the empirical null in place of the theoretical $N(0, 1)$ null is given by

$$\hat{\pi}_0 = 0.53, \hat{\mu}_0 = 0.14, \hat{\sigma}_0^2 = 1.20, \hat{\mu}_1 = 1.83, \hat{\sigma}_1^2 = 2.03.$$

This fit for the two-component normal mixture density is given in Figure 2, along with the empirical null and non-null normal component densities weighted by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$, respectively. The value

Table 2. GO terms for uniquely identified genes for the breast cancer data.

Gene	GO Term
UBE2B, DDB2	DNA repair
UBE2V1	cell cycle
RAB9, RHOC	small GTPase signal transduction
ITGB5, ITGA3	integrin mediated signalling pathway
PRKCBP1	regulation of transcription
HDAC3, MIF	negative regulation of apoptosis
KIF5B, spindle body protein	cytoskeleton organisation
CTCL1	vesicle mediated transport
TNAFIP1	cation transport
HARS, HSD17B7	metabolism

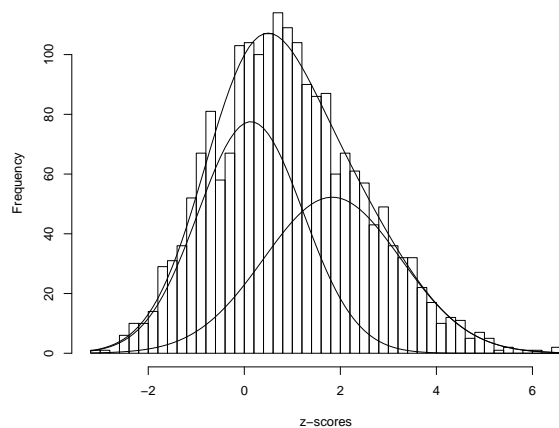


Fig. 2. Colon cancer data: plot of fitted two-component normal mixture model with empirical null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$) imposed on histogram of z -scores.

of twice the increase in the log likelihood here is only 1.8, which is less than the value of $2 \log n = 15.2$, as required to select the empirical null using BIC.

4 HIV DATA

In the example on the HIV data, a gene with a z -score of 2.53 has an estimated posterior probability of nondifferential expression equal to 0.3. But if all genes with this latter probability less than $c_o = 0.3$ were declared to be differentially expressed, then the implied FDR is only 0.14. Quoting just this last number gives an optimistic assessment of the gene's significance, as discussed by Efron (2005b) on a similar result in his analysis of this data set. This is because the FDR is obtained by averaging over genes with much more extreme values of the posterior probability of nondifferential expression than those with a value of 0.3 for this probability. As suggested by Efron (2005b), the local and global FDR measures can be combined.

5 SIMULATION STUDY: CORRELATED GENES

Allison et al. (2002) were interested in looking at the effect of the assumption of independently distributed expression levels of the genes. To this end, they generated gene expression levels for M experiments (with $M/2$ “mice” per experimental group) and for $N = 3000$ genes. The M vectors \mathbf{y}_j of dimension N were generated randomly from a multivariate normal distribution with covariance matrix specified to be

$$\Sigma = \sigma^2 \mathbf{B} \otimes \mathbf{I}_6 \quad (1)$$

and

$$\mathbf{B} = \rho \mathbf{1}_{500} \mathbf{1}_{500}^T + (1 - \rho) \mathbf{I}_{500}. \quad (2)$$

Here $\mathbf{1}_{500}$ denotes the unit vector of length 500 and \mathbf{I}_m is the $m \times m$ identity matrix. For the simulations the common variance was $\sigma^2 = 4$, while the correlation ρ varied over three values of 0 (independence), 0.4 (moderate dependence), and 0.8 (strong dependence). They noted that this covariance structure seems plausible since groups of genes are likely to be coexpressed, but it is unlikely that a particular gene is correlated with all other genes. For 20% of the genes (600 randomly selected), a true mean difference in expression between the two classes of mice was incorporated by adding d to the gene measurements \mathbf{y}_j from $j = \frac{1}{2}M + 1$ through to M . We applied our mixture model approach, using $d = 0, 4, 8$ and $M = 10$. As before, we transformed the pooled t -statistic to a z -score.

We found that the effect of the correlation was only marked in the case of strong correlation ($\rho = 0.8$) for $d = 4$ (that is, when the Mahalanobis distance between the means of the components corresponding to the differentially and nondifferentially expressed genes was not large). In Figure 3, we display the fit of the two-component normal mixture density, along with the theoretical $N(0, 1)$ null and non-null normal component densities weighted by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$, respectively, imposed on a histogram of the simulated (correlated) values of the z -scores in the case of $d = 4$ and $\rho = 0.8$. The corresponding plot with the use of the empirical null is given in Figure 4, where it can be seen that the use of the empirical null in place of the theoretical $N(0, 1)$ null almost reduces the effect of the correlation between the genes on the z -scores. In Figures 5 and 6, we give the corresponding densities for widely separated classes ($d = 8$), where it can be seen that the strong correlation between the genes ($\rho = 0.8$) has little effect on the mixture density of the z -scores with either the theoretical or empirical null component adopted.

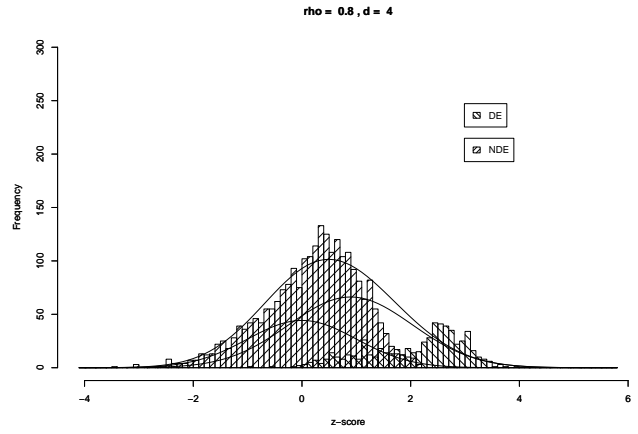


Fig. 3. Simulated correlated data: plot of fitted two-component normal mixture model with theoretical null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$) imposed on histogram of correlated values of z -scores for $\rho = 0.8$ and $d = 4$.

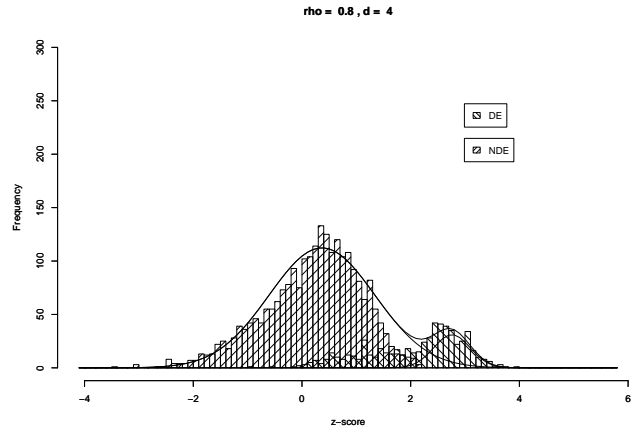


Fig. 4. Simulated correlated data: plot of fitted two-component normal mixture model with empirical null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$) imposed on histogram of correlated values of z -scores for $\rho = 0.8$ and $d = 4$.

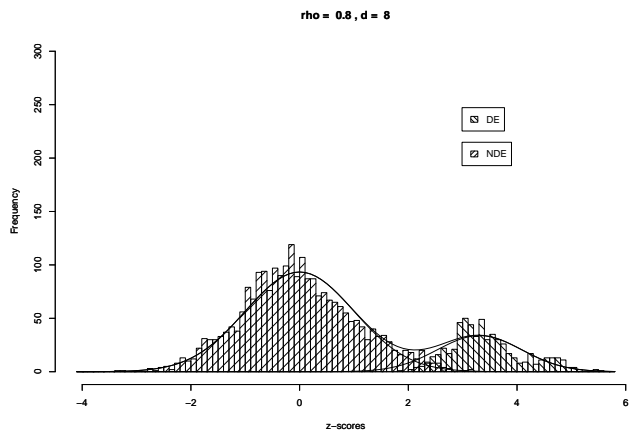


Fig. 5. Simulated correlated data: plot of fitted two-component normal mixture model with theoretical null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$) imposed on histogram of correlated values of z-scores for $\rho = 0.8$ and $d = 8$.

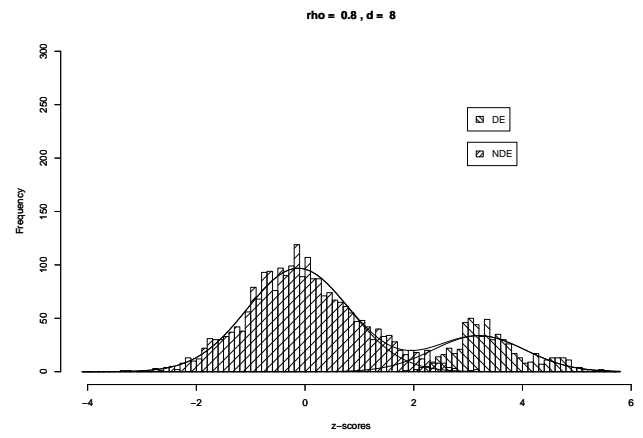


Fig. 6. Simulated correlated data: plot of fitted two-component normal mixture model with empirical null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$) imposed on histogram of correlated values of z-scores for $\rho = 0.8$ and $d = 8$.