Use of Mixture Models in Multiple Hypothesis Testing with Applications in Bioinformatics

Geoffrey J. McLachlan[1] and Leesa Wockner[2]

[1] Department of Mathematics & Institute for Molecular Bioscience, University of Queensland gjm@maths.uq.edu.au
[2] Department of Mathematics, University of Queensland l.wockner@uq.edu.au

Summary. There are many important problems these days where consideration has to be given to carrying out hundreds or even thousands of hypothesis testing problems at the same time. For example, in forming classifiers on the basis of high-dimensional data, the aim might be to select a small subset of useful variables for the prediction problem at hand. In the field of bioinformatics, there are many examples where a large number of hypotheses have to be tested simultaneously tested. For example, a common problem there is the detection of genes that are differentially expressed in a given number of classes. The problem of testing many hypotheses at the same time can be expressed in a two-component mixture framework, using an empirical Bayes approach; see, for example, Efron (2004). In this framework, we present further results as part of an ongoing investigation into the approach of McLachlan et al. (2006) on the adoption of normal mixture models to provide a parametric approach to the estimation of the so-called local false discovery rate. The latter can be viewed as the posterior probability that a given null hypothesis does hold. With this approach, not only can the global false discovery rate be controlled, but also the implied probability of a false negative can be assessed. The methodology is demonstrated on some problems in bioinformatics.

Key words: Large-scale simultaneous testing, Choice of null hypothesis, Local and global false discovery rates, False negative and positive rates, Normal mixture models

1 Introduction

The analysis of very large data sets presents many challenges. One is the need to carry out the simultaneous testing of hundreds or possibly thousands of statistical hypotheses. In situations with many variables, an initial step in many analyses is to reduce the dimension of the problem by selecting a subset of useful variables by consideration of the variables considered separately. For example, in a situation where the problem is to cluster the data at hand, the relevance of a variable might be assessed in terms of its effectiveness in revealing some group structure in the data. After formulating a null and an

alternative hypothesis for each variable to assess its relevance for the problem at hand, we subsequently obtain a P-value for each variable. The question then arises on how to select a useful subset of variables from the many P-values in the case of a high-dimensional data set.

A problem as described above can occur in many modern scientific studies. To present our methodology, we shall focus on the problem in bioinformatics arising in the analysis of microarray experiments, known as the detection of differential expression. With this problem, the aim is to determine which of several thousands genes are differentially expressed between a number of k different classes $C_1, \ldots, C_k$. In the case of $k = 2$, Class C_1 might refer to some women who have a good prognosis following the diagnosis of a disease (such as breast cancer) while the other Class C_2 refers to those women who have poor prognosis, corresponding to the occurrence of distant metastases within 5 years.

In classic situations involving only one single hypothesis test, the aim is to control the probability of making a Type I error; that is, the probability of making a false positive. In situations where multiple (N) hypotheses are under test, one can use the Bonferroni method to control the probability that at least one false positive error will be made. In Table 1, we have listed the possible outcomes from N hypothesis tests.

Table 1. Possible Outcomes from N Hypothesis Tests

	Accept Null	Reject Null	Total
Null True	N_{00}	N_{01}	N_0
Non-True	N_{10}	N_{11}	N_1
Total	$N - N_r$	N_r	N

However, in the current context where N is a very large number, controlling the family-wise error rate (that is, the probability that $N_{01} \geq 1$) is too strict and will lead to missed findings. In our example of the detection of genes differentially expressed between the classes, the goal is to identify as many genes with significant differences as possible, while incurring a relatively low proportion of false positives.

In a seminal paper, Benjamini and Hochberg (1995) introduced a new multiple hypothesis testing error measure called the false discovery rate (FDR), which they defined as

$$\text{FDR} = E\{\frac{N_{01}}{N_r \vee 1}\}, \tag{1}$$

where $N_r \vee 1 = \max(N_r, 1)$. The effect of $N_r \vee 1$ in the denominator of the expectation in (1) is to set $N_{01}/N_r = 0$ when $N_r = 0$. They proposed an FDR-controlling step-up test procedure for independent P-values associated with the N hypotheses.

Other error rates in addition to the FDR are of interest in practice, such as the false non-discovery rate (FNDR) and the false negative rate (FNR), as given empirically in Table 1 by the ratios, $N_{10}/(N - N_r)$ and N_{10}/N_1, respectively. As the FNDR is nearly always quite small since N_{00} is usually much larger than N_{10}, the FNR is generally more informative.

In this paper, we concentrate on a parametric approach to the handling of the P-values to provide a procedure that not only can be used to control the FDR, but also allows the implied FNR to be estimated. 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.

We adopt the parametric approach of McLachlan et al. (2006) that transforms the P_j values via the probit transformation to z-scores. Suppose in the present context of the detection of differential expression P_j denotes the P-value for the test of the null hypothesis

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

Then the z_j-score is given by

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

where Φ denotes the (cumulative) normal distribution function. This transformation is defined so that large positive values of the z_j-score suggest departures from the null hypothesis. Here the P_j values ($j = 1, \ldots, N$) constitute the input for this parametric approach. We do not consider how the P_j values are computed in the first instance. For example, they could be calculated on the basis of the classical t- or F-statistics, depending on whether there are two or multiple classes. Alternatively, the P_j might be calculated via a permutation method.

2 Modelling of Z-Scores

The density of the z_j-score can be modelled by 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 we denote the density of z_j in G_i by $f_i(z_j)$. The unconditional density of Z_j is then given by the two-component mixture model,

$$f(z_j) = \pi_0 f_0(z_j) + \pi_1 f_1(z_j).$$

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

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

In this framework, the gene-specific posterior probabilities provide the basis for optimal statistical inference about differential expression. The posterior probability $\tau_0(z_j)$ 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 π_0, the mixture density $f(z_j)$, and the null density $f_0(z_j)$, or equivalently, the ratio of densities $f_0(z_j)/f(z_j)$. Efron et al. (2004) has developed a simple empirical Bayes approach to this problem with minimal assumptions. We focus on a fully parametric approach using mixtures of normal densities. If the assumptions under which the P-values have been calculated hold, then the null density of Z_j is given by the standard normal density; that is,

$$f_0(z_j) = \phi(z_j;\, \mu_0,\, \sigma_0^2),$$

where $\mu_0 = 0$ and $\sigma_0^2 = 1$. This is known as the theoretical null distribution to distinguish it from the "empirical" null (as termed by Efron (2004)) in situations where the assumptions breakdown. The density $f_1(z_j)$ of z_j under the alternative hypothesis is approximated by a single normal density,

$$f_1(z_j) = \phi(z_j;\, \mu_1,\, \sigma_1^2).$$

In practice, the differentially expressed genes have varying values for the differences between their class means, and so it is somewhat surprising that for the data sets that we have analysed, a single normal distribution has sufficed to model the density of the z-scores for the non-null genes (genes that are differentially expressed). As shown by McLachlan et al. (2006), the two-component normal mixture model

$$f(z_j) = \pi_0 \phi(z_j;\, 0,\, 1) + \pi_1 \phi(z_j;\, \mu_1,\, \sigma_1^2)$$

can be fitted very quickly via the EM algorithm, as in their program called EMMIX-FDR.

The genes can be ranked on the basis of the estimated posterior probabilities $\tau_0(z_j)$, and we can select all genes with

$$\hat{\tau}_0(z_j) \leq c_o \tag{3}$$

to be differentially expressed. McLachlan et al. (2006) have shown how estimates of the implied rates, including the FDR and FNR, can be formed in terms of the $\tau_0(z_j)$ for a specified threshold c_o. In particular, an estimate of the FDR is given by

$$\widehat{FDR} = \sum_{j=1}^{N} \frac{\hat{\tau}_0(z_j) I_{[0,c_0]}(\hat{\tau}_0(z_j))}{N_r}, \tag{4}$$

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

3 Example: Breast Cancer Data

To illustrate the application of this parametric approach to multiple hypothesis testing, we consider the detection of differentially expressed genes for 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 $n_1 = 7$ BRCA1 tumours and $n_2 = 8$ BRCA2 tumours. We display the fitted mixture density in Figure 1.

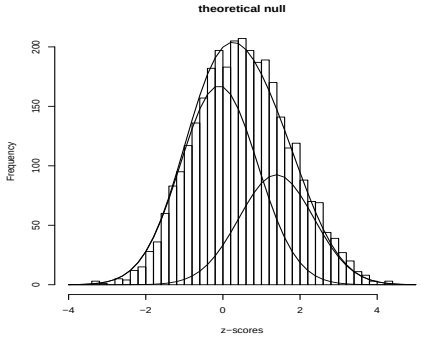


Fig. 1. Breast cancer data: plot of fitted two-component normal mixture model with theoretical $N(0,1)$ null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1-\hat{\pi}_0)$) imposed on histogram of z-scores.

In Table 2, we have listed the FDR estimated from (4) for various levels of the threshold c_o in (3). 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.

4 Empirical Null

As pointed by Efron (2004), for some microarray data sets the normal scores do not appear to have the theoretical null distribution, which is the standard

Table 2. Estimated FDR and other error rates for various levels of the threshold c_o 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_o	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

normal. In this case, Efron has considered the estimation of the actual null distribution called the empirical null as distinct from the theoretical null.

If we adopt an empirical null in our parametric approach, 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. One reason why the theoretical null distribution may not be appropriate is that the assumptions do not hold for the P-value to have a uniform distribution on the unit interval under the null hypothesis. Another reason is that the P-values are not independently distributed due to the expression profiles not being independent for all the genes.

5 Simulation Study

Allison et al. (2002) performed some simulations to investigate the effect of correlation among the genes on their results. They generated data for 10 tissue samples on 3000 genes. Each gene profile was drawn from a 3000-dimensional normal distribution with mean $\mu = 10$ and covariance matrix Σ. In order to mimic the idea that genes which are co-expressed would be correlated, but genes which are not co-expressed would not be correlated, Allison et al. (2002) split the 3000 genes into 6 blocks of size 500. Within each block the correlation between the genes ranged over the three values of ρ: 0 (independence), 0.4 (moderate dependence), 0.8 (strong dependence). This resulted in a covariance matrix of the form

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{B} \otimes \boldsymbol{I}_6, \tag{5}$$

where σ^2 is the common variance and where

$$\boldsymbol{B} = \mathbf{1}_{500}\mathbf{1}_{500}^T\rho + (1-\rho)\boldsymbol{I}_{500},$$

$\mathbf{1}_m$ is a vector of ones length m and $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. Finally, for 20% of randomly selected genes (600 genes), a mean difference of Δ was added to the expression levels for the last 5 tissue samples. Allison et al. (2002) suggested that a value of $\rho = 0.4$ 'tended to produce higher correlations

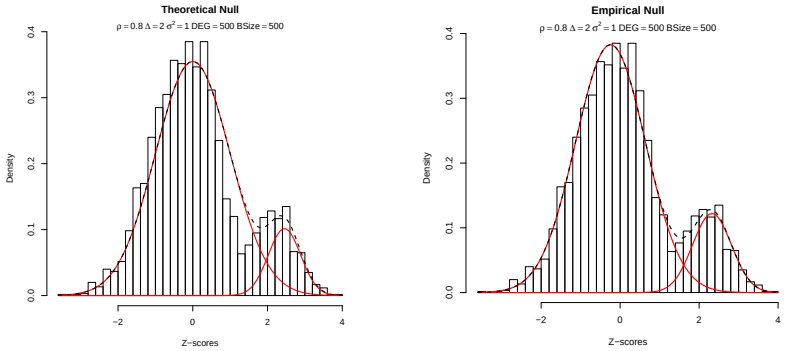
among gene expressions than were present in [their] actual example data set' contrary to previous opinions about the existence of correlations amongst gene expression levels.

Following this example, we generated data for 10 tissue samples from a normal distribution with $\mu = 0$ and correlation matrix as in (5), with $\sigma^2 = 1$ and ρ defined as before. For 500 randomly selected genes (17%) a difference Δ of 1 , 2 or 4 was added to the last five tissue samples.

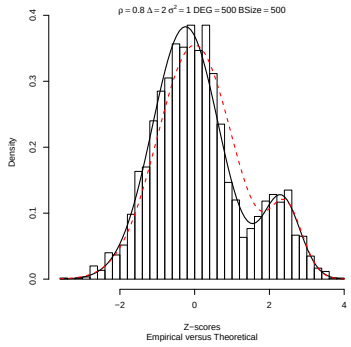
It was demonstrated in the presence of strong correlation between the genes ($\rho = 0.8$) that the empirical null distribution led to a much better fit than with the theoretical null; see Figure 2.

References

1. D.B. Allison, G.L. Gadbury, M. Heo, J.R. Fernndez, C.K. Lee, T.A. Prolla, and R. Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39:1-20, 2002.
2. Y. Benjamini, and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple testing. *Journal of the Royal Statistical Society* B57:289-300, 1995.
3. B. Efron. 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. 2004.
4. B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the Amercian Statistical Association*, 99:96-104, 2004.
5. B. Efron, and R. Tibshirani. Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.*, 23:7086, 2002.
6. B. Efron, R. Tibshirani, J. D. Storey and V. Tusher, Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* 96:11511160, 2001.
7. I. Hedenfalk, D. Duggan, Y. D. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi et al., Gene-expression profiles in hereditary breast cancer, *The New England Journal of Medicine* 344:539548, 2001.
8. M. M-T. Lee, F.C. Kuo, G.A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* 97:9834-9838. 2000
9. G.J. McLachlan, R.W. Bean, and L. Ben-Tovim Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22:1608-1615, 2006.



(a) (b)



(c)

Fig. 2. A simulation study where $\pi_0 = 0.83$ and $\Delta = 2$. (a)-(c) Strong dependence (a) Overall Theoretical Null (dashed) with two (weighted) components (solid) $\hat{\pi}_0 = 0.88$ (b) Overall Empirical Null (dashed) with two (weighted) components (solid) $\hat{\pi}_0 = 0.85$ (c) An overlay of Theoretical (dashed) and Empirical (solid).