Robust semiparametric mixing for detecting differentially expressed genes in microarray experiments

Marco Alfò, Alessio Farcomeni and Luca Tardella
Dipartimento di Statistica, Probabilitá e Statistiche Applicate
Università di Roma "La Sapienza"

31st January 2006

Abstract

An important goal of microarray studies is the detection of genes that show significant changes in observed expressions when two or more classes of biological samples such as treatment and control are compared. Using the *c-fold rule*, a gene is declared to be differentially expressed if its average expression level varies by more than a constant factor c between treatment and control (typically $c = 2$). While often used, however, this simple rule is not completely convincing. We propose to model this filter and define a binary variable at the gene×experiment level, allowing for a more powerful treatment of the corresponding information. We introduce a gene-specific random term controlling for both dependence among genes and variability with respect to the *c-fold* threshold. We make inference via a two-level finite mixture model under a likelihood approach. Then, using the counting distribution we show how parameter estimates can be usefully derived also under a Bayesian nonparametric approach which allows to keep under control some error rate of erroneous discoveries. We illustrate the effectiveness of both proposed approaches through a large-scale simulation study and an real-data application based on the well known dataset introduced by Alon *et al.* (1999).

Keywords: Microarray Data, Up-regulated genes, Mixture Models, Counting Distribution, False Discovery Rate.

1 Introduction and Motivation

An important goal of microarray studies is the detection of genes that show differential behavior in two or more biological situations of interest (Amaratunga and Cabrera (2004),Parmigiani *et al.* (2003)). Lists of (up or down) regulated genes can be identified by many different statistical methods. Supposing no additional covariate has been measured, one possibility is to do a T test (or F test in case of more than two biological conditions) on the expression levels for each gene, and select the significant genes using a correction for multiple testing. Most popular procedures control the False Discovery Rate (FDR) of Benjamini and Hochberg (1995), see for instance Dudoit *et al.* (2003) for a review. Another possibility is to apply clustering methods, as proposed by Alon *et al.* (1999); for a review, see Pollard and van der Laan (2003) and references therein. Though being more flexible than multiple testing methods, clustering methods do not actually provide a formal assessment of the significance of the differences between observed expression levels.

In almost all cases of actual practice some genes are preliminary filtered out from the analysis. For each gene the so called *fold-change* is measured in terms of the ratio between the expressions detected in the two different biological conditions. Genes that have afold change lower than a certain threshold (a value of 2 is commonly used) are not considered for validation, since the difference in expression may be too low to be actually detected by methods like PCR (polymerase chain reaction). This is the so called "2-fold rule", which has been widely used and criticized in the literature; among others, see for instance Tusher *et al.* (2001); Wolfinger *et al.* (2001); Sabatti *et al.* (2002); Gieseg *et al.* (2002). Applying the filter after testing is inefficient, since the power of many multiple testing procedures is not increasing with the number of tests. Some of the discoveries are filtered out and the power of the entire procedure is artificially lead to be lower than it could. On the other hand, it is not possible to apply the filter before doing the tests, since all the test statistics must be considered in the multiple testing procedure to avoid explosion of the number of false positives (see Hochberg and Tamhane (1987)). To bypass some of the aforementioned difficulties we propose an approach to the discovery of differentially expressed genes via a mixture model, which efficiently incorporates the 2-fold filter and provides a formal statement on the regulation of selected genes between the analyzed biological conditions. The proposed approach generalizes the standard "2-fold rule", avoiding to test the average expression level; moreover, it can be employed in both a likelihood and a Bayesian framework, and can be joined with proper control of the estimated FDR (see e.g. Storey *et al.* (2004)). As far as we know, the SAM (Significance Analysis of Microarrays) of Tusher *et al.* (2001) is the only other technique that directly incorporates the c-fold rule and controls the FDR. The main difference between SAM and our techniques is that we model the gene expression at a slide level.

The paper is structured as follows: in Section 2 we introduce our modeling approach together with the adopted notation. In Section 3 we describe a finite mixture approach. In Section 4 we slightly deviate from the finite mixture model leaving unspecified the mixing density. We describe a Bayesian method to make inference on the counting distribution of exceedances. In Section 5 we illustrate how to choose the regulated genes using the proposed modeling strategies. The empirical behavior of the proposed models is investigated through both a simulation study presented in Section 6 and an application to the benchmark dataset discussed by Alon *et al.* (1999). Finally Section 7 gives some concluding remarks.

2 Modeling framework

Let us start assuming we are analyzing a set of G genes, composed of $G_1 \ll G$ genes which are truly differentially expressed, while $G_0 = G - G_1$ are not. The corresponding sets of genes will be denoted with $\mathcal{G}_1$ and $\mathcal{G}_0$, respectively; in real examples G_1/G can be lower than 1%. Let us suppose we have recorded the normalized ratios of i-th gene, $i = 1, \ldots, G$, on n_i slides; on each slide tissues from two different biological conditions were hybridized. In the following we denote with f_{ij} the fold change observed for the i-th gene in the j-th sample, $i = 1, \ldots, G$, $j = 1, \ldots, n_i$. We assume that genes that are not differentially expressed are the same in all slides, thus showing no substantial difference across the analyzed biological conditions. For simplicity, we restrict the discussion to up-regulated genes as if the researcher is filtering out only those genes whose fold change is below a certain threshold c – say $c = 2$. It is obviously straightforward to consider down-regulated genes, and the proposed approach could be also extended, at least in principle, to model jointly down and up-regulated genes.

While the c-fold filter is usually applied on the mean (or median) fold for each gene, usually at a logarithmic scale, we propose here to apply it at a slide level. We transform the original fold change measures into a binary matrix Y, in which y_{ij} is 1 if the fold change of i-th gene is above c (i.e. $f_{ij} > c$) in the j-th sample, and 0 otherwise. It is worth noting that the cut-off c is meaningful to the researchers and usually not completely arbitrary. We show in the following that this transformation allows a powerful, efficient and more robust treatment of the analyzed data exploiting many tools that are usually devised in statistics for binary data. We only need to take into account the special characteristics of the data, in particular the fact that the number of genes is very high and potentially only a small number of them is truly differentially expressed; at the same time, the number of samples is relatively small.

Our method is more suitable for two-color technologies in which a fold change based on a natural matching of biological samples directly measured by putting tissues from the two classes on the same slide. See Schena (2000) for a discussion of the different technologies that can be used in DNA microarrays, and Kerr and Churchill (2001) for issues in experimental design in this setting. Nevertheless, the derivation of the binary matrix Y may yield good results also with other technologies through the derivation of individual fold changes with an appropriate matching of slides, possibly at random..

First, genes are modeled separately and information from different genes is pooled only for non differentially expressed, in the sense that they share a common probability of yielding an expression above the threshold c. Then, genes are jointly modeled in a hierarchical framework, where the probability of being above the threshold for differentially expressed genes is modeled using a common but unknown distribution, which we denote by $F(\cdot)$. The underlying idea is that expression of non differentially expressed genes should differ only by experimental error, while a more heterogeneous behavior is expected for differentially expressed genes.

More formally, we avoid distributional assumptions on the fold changes f_{ij}, $i = 1, \ldots, G$, $j = 1, \ldots, n_i$ and we limit ourselves to directly model the binary outcome

$$Y_{ij} = \begin{cases} 1 & \text{if } f_{ij} > c \\ 0 & \text{otherwise} \end{cases}$$

We denote with

$$q_0 = \Pr(f_{ij} > c \mid i \in \mathcal{G}_0), \qquad j = 1, \ldots, n_i \tag{1}$$

the probability that the $i-$th gene, which is not differentially expressed, yields a fold change f_{ij} above c in the $j-$th sample. On the other hand, let

$$q_i = \Pr(f_{ij} > c \mid i \in \mathcal{G}_1), \qquad j = 1, \ldots, n_i \tag{2}$$

denote the probability that a truly differentially expressed gene yields a fold change over the threshold c. We are implicitly assuming that q_i does not change across slides, i.e., that the slides are exchangeable. Moreover, we assume that $q_i > q_0$ not only for identifiability reasons, but also because it seems natural to expect that truly up-regulated genes are more likely than non up-regulated genes to yield a fold change above c. With the above assumptions Y_{ij} is a Bernoulli variate with marginal probability

$$\Pr(Y_{ij} = 1) = p_i = q_0(G_0/G) + q_i(G_1/G), \quad \forall j = 1, \ldots, n_i. \tag{3}$$

It is usually safe to assume independence at the sample level so that for the i-th gene we would have a likelihood proportional to

$$p_i^{\sum_j y_{ij}}(1-p_i)^{\sum_j (1-y_{ij})}.$$

We can express the above model as a complete data problem. Let z_i be the indicator of the i-th gene to be differentially expressed. We have

$$\begin{aligned} \Pr(Y_{ij} = 1 | Z_i = 1) &= q_i \\ \Pr(Y_{ij} = 0 | Z_i = 0) &= q_0 \\ \Pr(Z_i = 1) &= G_1/G = \pi. \end{aligned}$$

Note that the individual contribution to the likelihood becomes proportional to

$$(G_1)^{z_i}(G_0)^{(1-z_i)}q_0^{\sum_j y_{ij}(1-z_i)}(1-q_0)^{\sum_j (1-y_{ij})(1-z_i)}q_i^{z_i \sum_j y_{ij}}(1-q_i)^{z_i \sum_j (1-y_{ij})}$$

and that our considerations imply (by an immediate application of Bayes theorem) the following statement:

$$P(Z_i = 1 | y_{ij} > c) = \frac{q_i G_1/G}{p_i}.$$

Unfortunately, it is well acknowledged that genes are likely to be dependent, so that we can not combine the G likelihoods directly. This is a usual problem in microarray data analysis. A standard approach is to treat each gene separately, as if one is considering G different models and then make a correction for the multiplicity. Here a good correction may be given by shrinking the posterior parameter estimates.

An EM algorithm can now be used to fit this model. Potential problems can arise due to the number of samples n_i which is usually small when compared to G and consequently the variance of the estimates can be very large. The approach in the next section tries to overcome this problem by linking the estimates at a gene level. The main quantity of interest here is a formal statement on z_i, $i = 1, \ldots, G$, that is, an estimation of the probability of the i-th gene to be differentially expressed. There are some interesting by-products: one can estimate the strength of the effect with q_i, and a general statement on the sensibility of the c-fold test for differential expression is given by the first moment of $(q_i)_{i=1,\ldots,G}$. The quantity $(1-q_0)$ gives a measure of both the specificity of the c-fold test, and the ability in carrying out the experiment: if non-differentially expressed genes are very likely to have high fold changes, probably normalization did not properly clean the data.

3 Likelihood analysis with finite mixture approach

We start by briefly reviewing the approach in a simple case where a common parameter for the up-regulated genes $q_i = q_1$ is assumed. Given the above description of the complete data problem, the marginal density is

$$f(\mathbf{y},\mathbf{z}) = \prod_{i=1}^{G} \left\{ \left[\left(\prod_j q_1^{y_{ij}} (1-q_1)^{(1-y_{ij})} \right) \pi \right]^{z_i} \left[\left(\prod_j q_0^{y_{ij}} (1-q_0)^{(1-y_{ij})} \right) (1-\pi) \right]^{(1-z_i)} \right\}$$

and the corresponding log-likelihood function is therefore defined as follows

$$\begin{aligned} \ell_c(\cdot) &= \sum_{i=1}^{G} \left[z_i \left(\sum_j y_{ij} \log(q_1) + \sum_j (1-y_{ij}) \log(1-q_1) + \log(\pi) \right) \right] + \\ &+ \sum_{i=1}^{G} \left[(1-z_i) \left(\sum_j y_{ij} \log(q_0) + \sum_j (1-y_{ij}) \log(1-q_0) + \log(1-\pi) \right) \right] \end{aligned}$$

Denoting with $\sum_j y_{ij} = k_i$ the number of samples with a fold change over the threshold, the log-likelihood function can be rewritten as:

$$\begin{aligned} \ell_c(\cdot) &= \sum_{i=1}^{G} \{ z_i \left[k_i \log(q_1) + (n_i - k_i) \log(1-q_1) + \log(\pi) \right] \} = \\ &+ \sum_{i=1}^{G} \left[(1-z_i) \left(k_i \log(q_0) + (n_i - k_i) \log(1-q_0) + \log(1-\pi) \right) \right] \end{aligned}$$

As usual, in the E-step of the EM algorithm, we define the log-likelihood for the *observed* data by taking the expectation of the log-likelihood for *complete* data over the unobservable class indicator vector z_i given the observed data $\mathbf{y}$ and the current ML estimates, say $\mathbf{q}^{(t)}$. In other words, at the t-th step, we replace z_i with its conditional expectation

$$w_i = \frac{f_{i1}\pi}{f_{i1}\pi + f_{i0}(1-\pi)}$$

where $f_{i1} = f(\mathbf{y}_i \mid z_i = 1)$, and $f_{i0} = f(\mathbf{y}_i \mid z_i = 0)$. The conditional expectation of the complete log-likelihood given the observed data $\mathbf{y}$ is expressed by the function

$$\begin{aligned} Q^{(t)}(\cdot) &= \mathrm{E}_{\mathbf{q}}^{(t)} \{ \ell_c(\cdot) | \mathbf{y} \} = \sum_{i=1}^{G} \{ w_i \left[k_i \log(q_1) + (n_i - k_i) \log(1-q_1) + \log(\pi) \right] \} + \\ &+ \sum_{i=1}^{G} \{ (1-w_i) \left[k_i \log(q_0) + (n_i - k_i) \log(1-q_0) + \log(1-\pi) \right] \} \end{aligned}$$

Maximizing $Q^{(t)}(\cdot)$ with respect to $\mathbf{q} = (q_0, q_1)$ we obtain the following ML parameter estimates:

$$\widehat{q}_0^{(t)} = \frac{\sum_i (1-w_i^{(t)}) k_i}{\sum_i (1-w_i^{(t)}) n_i}, \qquad \widehat{q}_1^{(t)} = \frac{\sum_i w_i^{(t)} k_i}{\sum_i w_i^{(t)} n_i} \tag{4}$$

while the prior probability estimate is

$$\widehat{\pi}^{(t)} = \frac{\sum_i w_i}{G}$$

which represents a standard result in ML estimation in finite mixtures. It is worth noticing that the constraint $q_1 > q_0$ can be satisfied by simply (post-) ordering the estimated locations. However, the hypothesis that all genes are homogeneous, i.e. $q_i = q_1$, $i = 1, \ldots, G$, can be quite unsatisfactory. If we rather adopt subject-specific parameters q_i, $i = 1, \ldots, G$, we can easily generalize the EM algorithm to endow subject-specific parameters obtaining the following estimates

$$\widehat{q}_i^{(t)} = \frac{w_i^{(t)} k_i}{w_i^{(t)} n_i}.$$

Obviously, these estimates are free of any constraints, while we should rationally assume that $q_i > q_0$ to ensure that truly up-regulated genes are more likely to yield a fold change above c. Also in this case post-estimation ordering can be of help. However ML estimation could be in this case particularly cumbersome since the number of parameters could be very high (often $G \geq 5000$ in practical examples). Moreover, while heterogeneity among genes is preserved by this model, potential dependence among genes is not accounted for. For this purpose, we introduce a discrete mixing distribution on q_i to allow for dependence among genes belonging to the same component, while controlling for potential overdispersion with regards to the homogeneous Binomial model. We consider that, conditionally on $Z_i = 1$, the observed counts of samples yielding a fold change over the threshold follow an overdispersed distribution, such as a finite mixture of K Binomial distributions, with locations defined over the interval $(q_0, 1]$. We define a second-level component indicator, say X_{ik}, with $X_{ik} = 1$ if $Z_i = 1$ and the gene belongs to the $k-$th component of the finite mixture. We denote with

$$\tau_k = \Pr(X_{ik} = 1 \mid Z_i = 1)$$

the corresponding prior probabilities. The marginal density for the complete data is

$$\begin{aligned} f(\mathbf{y}, \mathbf{z}, \mathbf{x}) &= \prod_{i=1}^{G} \left\{ \left[\prod_k \left(\prod_j q_k^{y_{ij}} (1 - q_k)^{(1-y_{ij})} \tau_k \right)^{x_{ik}} \pi \right]^{z_i} \times \right. \\ &\times \left. \left[\left(\prod_j q_0^{y_{ij}} (1 - q_0)^{(1-y_{ij})} \right) (1 - \pi) \right]^{(1-z_i)} \right\} \qquad (5) \end{aligned}$$

The log-likelihood function for the complete data is as follows

$$\begin{aligned} \ell_c(\cdot) &= \sum_{i=1}^{G} \left\{ z_i \left[\sum_{k=1}^{K} x_{ik} \left[k_i \log(q_k) + (n_i - k_i) \log(1 - q_k) + \log(\tau_k) \right] + \log(\pi) \right] \right\} + \\ &+ \sum_{i=1}^{G} \left\{ (1 - z_i) \left[k_i \log(q_0) + (n_i - k_i) \log(1 - q_0) + \log(1 - \pi) \right] \right\}. \end{aligned}$$

As before, in the E-step of the EM algorithm, we define the log-likelihood for the *observed* data by taking the expectation of the log-likelihood for *complete* data over the unobservable class indicator vectors $(\mathbf{z}, \mathbf{x})$ given the observed data $\mathbf{y}$ and the current ML estimates, say $\mathbf{q}^{(t)}$. In other words, at the t-th step, we replace both z_i and x_{ik}, $i = 1, \ldots, G$, $k = 1, \ldots, K$ with their conditional expectations:

$$w_i = \frac{f_{i1}\pi}{f_{i1}\pi + f_{i0}(1-\pi)} \quad v_{ik} = \frac{f_{i1k}\tau_k}{\sum_k f_{i1k}\tau_k}$$

where $f_{i1k} = f(\mathbf{y}_i \mid Z_i = 1, X_{ik} = 1)$. The conditional expectation of the complete log-likelihood given the observed data $\mathbf{y}$ is expressed by the function

$$\begin{aligned} Q^{(t)}(\cdot) &= \sum_{i=1}^{G} \left\{ w_i \sum_k v_{ik} \left[k_i \log(q_k) + (n_i - k_i)\log(1-q) + \log(\tau_k) + \log(\pi) \right] \right\} + \\ &+ \sum_{i=1}^{G} \left\{ (1-w_i) \left[k_i \log(q_0) + (n_i - k_i)\log(1-q_0) + \log(1-\pi) \right] \right\} \end{aligned}$$

Maximizing $Q^{(t)}(\cdot)$ with respect to q_0 and $\mathbf{q}' = (q_1, \ldots, q_K)$ we obtain the following ML estimates for the parameters of the Bernoulli densities

$$\widehat{q}_0^{(t)} = \frac{\sum_i (1 - w_i^{(t)}) k_i}{\sum_i (1 - w_i^{(t)}) n_i}, \qquad \widehat{q}_k^{(t)} = \frac{\sum_i w_i^{(t)} v_{ik} k_i}{\sum_i w_i^{(t)} v_{ik} n_i} \tag{6}$$

while the prior probability estimates are

$$\widehat{\pi}^{(t)} = \frac{\sum_i w_i}{G} \quad \widehat{\tau}_k^{(t)} = \frac{\sum_i w_i v_{ik}}{\sum_i w_i}$$

which mimic previous results. In this case, however, we have not taken into account the K constraints $q_k \in (q_0, 1]$ and post-estimation ordering can not be of any help. To avoid complex maximization procedures, we can proceed as follows: doing a little algebra, we can show that the Q function in equation (6) is equal to a Q function corresponding to a simple finite mixture of $K+1$ Binomial densities. In fact, by writing $z_{ik}^* = z_i x_{ik}$, $w_{ik}^* = w_i v_{ik}$, $1 - z_i = z_{K+1}^*$, $\pi_k^* = \pi\tau_k$ and $\pi_{K+1} = 1 - \pi$, $i = 1, \ldots, G$, $k = 1, \ldots, K$, we obtain

$$\begin{aligned} Q^{(t)}(\cdot) &= \sum_{i=1}^{G} \sum_{k}^{K} w_{ik}^* \left\{ \left[k_i \log(q_k) + (n_i - k_i)\log(1-q) + \log(\pi * \tau k) \right] \right\} + \\ &+ \sum_{i=1}^{G} w_{iK+1}^* \left[(k_i \log(q_{K+1}) + (n_i - k_i)\log(1 - q_{K+1}) + \log(\pi_{K+1})) \right] = \\ &= \sum_{i=1}^{G} \sum_{k=1}^{K+1} \left\{ w_{ik} \sum_k \left[k_i \log(q_k) + (n_i - k_i)\log(1-q_k) + \log(\pi_k^*) \right] \right\} \end{aligned}$$

and this suggests a possible simple approach to estimation and to post-estimation ordering. In fact, constraints $q_k > q_0$ are satisfied by posing $\widehat{q}_0 = \widehat{q}_{K+1} = \min_k(\widehat{q}_k)$. In this case, the ML estimates are those of a standard finite mixture of $K+1$ Binomial densities, namely

$$w_{ik}^{*(t)} = \frac{f_{ik}\pi_k^*}{\sum_k f_i \pi_k^*} \quad \widehat{q}_k^{(t)} = \frac{\sum_i w_{ik}^{*(t)} k_i}{\sum_{i,k} w_{ik}^{*(t)} n_i}$$

and

$$\widehat{z}_i = w_i^* = \sum_{k=1}^{K} w_i v_{ik} = \sum_{k=1}^{K} w_{ik}^* = 1 - w_{K+1}.$$

So far we have considered a finite mixture model for q_i with a fixed number K of components. In fact, it is easy to consider a broader class of mixture models with an unknown number of components, say from a minimum of $K = 1$ to a maximum $K = K_{max}$. Several fitting procedures are then carried out and one can select the appropriate K by using an information criterion such as BIC.

This model approach is quite simple to be implemented and is often robust to model misspecification, but it can suffer from possible drawbacks. In fact, while it is well known that finite mixture models tend to closely fit the observed distribution, in empirical applications they are likely to produce a *smooth* sequence of location estimates. In the empirical context we are analyzing, this would result in substantial problems when selecting up-regulated genes, i.e., in separating one location from the other K. In the absence of any a priori information on the magnitude of q_0 or $\mathcal{G}_0$, which can be based on control or noise spots, stronger constraints like $q_i - q_0 > M > 0$ or larger individual sample sizes n_i are needed to detect a more reliable segmentation of the analyzed genes. A possible alternative is to assume a continuous distribution for the q_i, $i = 1, \ldots, G$, such as a Beta distribution reshaped to take values on the interval $(q_0, 1]$. This choice does not lead to a closed form for the corresponding log-likelihood function

$$\begin{aligned} \ell_c(\cdot) &= \sum_{i=1}^{G} z_i \left\{ \log \left[\int_{q_0}^{1} \left(q_i^{k_i} (1 - q_k)^{(n_i - k_i)} \right) \mathrm{d}F(q_i) \right] + \log(\pi) \right\} + \\ &+ \sum_{i=1}^{G} \left\{ (1 - z_i) \left[k_i \log(q_0) + (n_i - k_i) \log(1 - q_0) + \log(1 - \pi) \right] \right\}. \end{aligned}$$

Therefore, only procedures based on approximations (such as those based on QL approximation or on linearization techniques) or on numerical integration can be employed to obtain ML estimates. However, these techniques often fail with discrete data. For this reason, we now turn to re-express this problem adopting a more suitable approach, based on the counting distribution, which is detailed in the following paragraph.

4 A Bayesian semiparametric approach

A slightly different perspective for making inference could be adopted with a semiparametric model for the analyzed data, by leaving the latent distribution $F(\cdot)$ unspecified and modeling the counting distribution of exceedances. Assuming for simplicity that $n_i = n$ for any $i = 1, \ldots, G$, let h_k be the number of genes that exceed the threshold exactly k out of n times. It is easy to realize from the independence structure that the vector $\mathbf{h} = (h_0, \ldots, h_n)$, referred to as the counting distribution, contains all the relevant experimental evidence. Note that $\sum_{k=0}^{n} h_k = G$. The probability of observing a vector $\mathbf{h} = (h_0, \ldots, h_n)$,

$p(h|F, q_0, G_0)$, is given by

$$\begin{aligned} p(h|F, q_0, G_0) &= \frac{G!}{\prod_k h_k!} \prod_k p_{F,k}^{h_k} \\ p_{F,k} &= \binom{n}{k} \left[(q_0^k (1-q_0)^{n-k}) \frac{G_0}{G} + \frac{G_1}{G} \int_{q_0}^1 q_i^k (1-q_i)^{n-k} \, dF(q_i) \right]. \end{aligned}$$

The resulting F-mixture binomial probability is almost intractable in this form, but one can overcome difficulties by expanding the binomial term displayed as second factor inside the integral. After a little bit of algebra, this leads to the following expression:

$$p_{F,k} = \binom{n}{k} \left[(q_0^k (1-q_0)^{n-k}) \frac{G_0}{G} + \frac{G_1}{G} \sum_{r=k}^{n} (-1)^{r-k} \binom{n-k}{r-k} m_r \right], \qquad (7)$$

where $m_r = \int_{q_0}^1 x^r \, dF(x)$ is the r-th moment of $F(\cdot)$. Now we can turn the seemingly cumbersome problem to a purely parametric one, since $F(\cdot)$ intervenes only through a finite number of its moments. The problem is now tractable, and parameters can be estimated either with a Bayesian approach or by a maximum likelihood. A similar alternative semiparametric approach to model dependent binary data can be found in George and Bowman (1995). A convenient reparametrization of (7) which avoids dealing with a constrained convex parameter space can be obtained by replacing the moments of $F(\cdot)$ with the corresponding canonical moments (Dette and Studden, 1997), which are in one-to-one correspondence and are defined over an unrestricted space, such as $(0,1)^n$. Also in Dette and Studden (1997) it is shown how one can easily compute the invertible mapping, which we denote with $\psi(\cdot)$, from the space of the first n canonical moments $c_1, \ldots, c_n$ to the space of the first n ordinary moments $m_1, \ldots, m_n$. They also prove that the canonical moments are invariant under linear transformations of the random variables with those canonical moments; for this reason, we can conveniently use the easily tractable mapping $\psi(\cdot)$ for the moments of an hypothetical random variable X_i on $[0,1]$, and then apply the mapping $\eta : (0,1)^n \to (q_0, 1)^n$ to get moments of $q_i = X_i * (1-q_0) + q_0$. From the same source (Dette and Studden, 1997) we know that the mapping $\eta(\cdot)$, for the r-th moment, is

$$E[q_i^r] = \sum_{i=0}^{r} \binom{r}{i} q_0^{(r-i)} (1-q_0)^i E[X^i].$$

Hence, $p_{F,k}$ is conveniently re-expressed as

$$p_{F,k} = \binom{n}{k} [(q_0^k (1-q_0)^{n-k}) \frac{G_0}{G} + \frac{G_1}{G} \sum_{r=k}^{n} (-1)^{r-k} \binom{n-k}{r-k} \eta(\psi(c_r))]$$

We propose to derive inference under the Bayesian paradigm. In order to do that, we just need to specify a convenient prior distribution over the parameter space and then derive the posterior distribution conditionally on the observed data. As usual, no closed form expression can be obtained and hence we propose to approximate the posterior distribution by a Markov Chain Monte Carlo sampling scheme (Gilks *et al.* (1998)). In fact, we have used an hybrid sampler called ARMS (Gilks and Wild, 1992), which combines a Gibbs sampling scheme with a Metropolis-Hastings routine to draw from full conditionals. This provides an automatic Metropolis-within-Gibbs sampler, avoiding the need for the usual fine tuning of Metropolis-Hastings proposal parameters. In a Bayesian context, we need to specify priors on the parameters $(G_0/G, q_0, c_1, \ldots, c_N)$; if prior information on such parameters is available, this should be used. For instance, it should be customary to use for G_0/G a prior which puts most of the mass close to 1, like a $Beta < a, 1 >$ where a is large (say $a \geq 10$); nevertheless, default priors for other parameters can be used either flat uniform priors or Jeffreys' prior for the ordinary moments. More details on how to obtain such default priors in terms of ordinary moments and canonical moments can be found in Tardella (2002). A little extra detail must be added here since one needs also to consider the mapping $\eta(\cdot)$ and hence the corresponding Jacobian. In fact, $\frac{\partial E[q_i^r]}{\partial m_s} = \binom{r}{s}(1-q_0)^s q_0^{r-s}$ if $r \geq s$ and 0 otherwise. The Jacobian is the determinant of the matrix whose (r,s)-th entry is given by the expression above, and can be computed in a closed form as $(1-q_0)^{\sum_{k=1}^n i} = (1-q_0)^{n(n+1)/2}$. Of course, we could have used a maximum likelihood approach instead of deriving Bayesian inference, but we have eventually preferred this last option because the maximization routine used in terms of canonical moments produced less stable and reliable results than those obtained for the Bayesian analysis through the MCMC simulation machinery.

5 Selection of up-regulated genes

We now describe how one can use the modeling approaches in Section 3 and 4 to select up-regulated genes.

With respect to the approach described in Section 3, once the complete data problem for the finite mixture model has been setup, one can rely on the posterior probabilities

$$Pr\{Z_i = 1 | data\} \qquad i = 1, \ldots, G$$

and select genes for which the above probability is greater than 0.5. One can also use the finite mixture structure to cluster up-regulated genes in further K subgroups; this can be done adopting a maximum a posteriori (MAP) approach on the probabilities of component membership X_{ik}, $k = 1, ..., K$.

When turning to the modeling approach of Section 4, we have devised a different method which is based directly on the counting distribution. After fitting the model, one can choose those genes which exceed the threshold at least $\bar{k}$ times with an appropriate value of $\bar{k}$. In order to select $\bar{k}$ one can try to keep under control some error rate (such as FDR) of the resulting procedure. Recall that the FDR of Benjamini and Hochberg (1995) is the expected proportion of false discoveries over the number of selected genes, if any, and zero otherwise. In fact, for any fixed k we can estimate the number of falsely selected genes as follows

$$\widehat{FDR}_k = \frac{\sum\limits_{i=k}^{S} \widehat{G}_0 \binom{S}{i} \widehat{q}_0^i (1-\widehat{q}_0)^{S-i}}{\sum\limits_{i=k}^{S} h_i}. \tag{8}$$

Then, we select the smallest k for which the FDR is estimated to be below a pre-specified level α can be chosen, $\bar{k} = \min_{k=1,\ldots,n} \{k : FDR_k \leq \alpha\}$. We will make the resulting procedure clearer with a simple numerical example at the beginning of next section.

6 Applications

In this section we will give an illustration of FDR control and the Bayesian approach to gene selection we have just proposed. Then, we will show a large simulation study in order to check the validity of our proposals, and in particular to verify the comparative performance of both methods. Finally, a benchmark dataset is discusses to give more insight.

6.1 Numerical illustration of the selection rule for the Bayesian approach

We generated one simulated dataset as follows: we took $G = 2000$, of which $G_0 = 1800$ not differentially expressed, with a probability of exceeding the threshold equal to $q_0 = 0.05$. The probability q_i corresponding to the $G_1 = 200$ genes with differential expression is generated as $(1 - q_0)X_i + q_0$, where X_i is a Beta variate with parameters 0.5 and 0.5. All genes are observed in $n_i = n = 4$ replicates and using this setup we have obtained a vector of counts $\mathbf{h} = (1503, 332, 72, 38, 55)$.

After deriving MCMC approximations for the joint posterior distributions of all parameters we got the posterior means of parameters of interest as shown in Table 1, which are actually close to the true values used for generating the data.

	True Value	Posterior Mean
G_1	200	216
q_0	0.05	0.046
$E[q_i]$	0.5	0.497
$E[q_i^2]$	0.375	0.341
$E[q_i^3]$	0.312	0.269
$E[q_i^4]$	0.273	0.228

Table 1: True values and posterior means for the simulated example

We can now illustrate with this simple numerical example the rationale used to filter out genes that have to be considered over expressed. One could have chosen a number of around $\widehat{G}_1 = 216$ genes to be declared significantly differentially expressed. In that case one would have taken all the 165 (72+38+55) genes which exceed the threshold c two or more times, and discarded the remaining ones. However, this would yield a high number of false discoveries. In fact we know from the simulated data that 38 among these 165 genes are not truly differentially expressed and this means that by choosing 165 genes we would get a False Discovery Proportion (the true number of false discoveries over the number of rejections) of 0.23. To bound the proportion of false discoveries we decided to aim at a proper control over the expected False Discovery Proportion, namely, the FDR. Hence we suggest to use the posterior expected number of non differentially expressed genes, which is equal to $\widehat{G}_0 = 2000 - 216 = 1784$. Each non-regulated gene is expected to be over the threshold c with probability $\widehat{q}_0 = 0.046$. Therefore, we expect $\widehat{G}_0 \binom{S}{k} \widehat{q}_0^k (1 - \widehat{q}_0)^k$ genes to be over the threshold exactly k times. In this case, we get a vector of expected false discoveries $\mathbf{h}^* = (1440, 278, 20, 1, 0)$. This allows us to estimate the posterior expected false discovery proportion for each k as the ratio of the reverse cumulative sums of $\mathbf{h}^*$ and $\mathbf{h}$, as $\widehat{FDR} = (0.869, 0.60, 0.12, 0.01, 0)$. Hence, an FDR controlling procedure consists in fixing $\bar{k} = 3$ and selecting those 93 (38+55) genes exceeding the threshold 3 or 4 times.

6.2 Comparison of the procedures based on simulated data

We have studied the comparative performance of our proposed methods by generating $B = 1000$ simulated datasets, which mimic a classical microarray experiment. Again, we have considered $G = 2000$ genes, of which only $G_1 = 200$ up-regulated. The fold changes have been simulated according to $\log(f_{ij}) \sim N(0, \sigma^2)$, where σ^2 is such that $q_0 = \int_2^\infty \phi(x/\sigma)\ dx$ and $\phi(\cdot)$ is the pdf of a standard normal random variable. This leads to fix $\sigma^2 = 1.478$. The fold changes for the differentially expressed genes are generated according to $\log(f_{ij}) \sim N(\mu_i, \sigma^2)$, where μ_i is such that $\int_2^\infty \phi(\frac{x-\mu_i}{\sigma})\ dx = q_i$. The remaining simulation set-up is analogue to the previous section. We have used the simulated datasets to infer on the over-expressed genes. The reference approaches have been based on calculating p-values from a standard one sample t-test on the generated log-fold changes. The experiment has been carried over for $B = 1000$ replications and the corresponding average (actual) false discovery proportion (the *false discovery rate*), average false non-discovery proportion (the *false non-discovery rate*) and number of selected genes (R) have been recorded. The False Non-discovery Rate (FNR) is a Type II error rate defined in Genovese and Wasserman (2002) as the proportion of false negatives over the number of non-rejected hypotheses; and we denote by $\bar{f}_{i.}$ the geometric average of the fold change for the i-th gene.

Table 2 gives an overall picture of comparative performances. SAM is the technique proposed in Tusher *et al.* (2001) with no threshold for the fold change, while SAM-2 is the same technique in which a c-fold rule is applied with $c = 2$. Somewhat surprisingly one can see a low expected number of rejections given by the Benjamini and Hochberg (1995) (BH) correction. This can be explained as a consequence of the small number of replicates and the fact that the signal is weak (some of the q_is are very close to q_0). Table 2 also shows that the method based on the counting distribution, among the methods controlling the FDR at level 0.05, is the most powerful (i.e. lowest FNR). It is worth noticing that the method succeeds in having a lower FNR *and* a lower realized FDR. The finite mixture approach, while yielding good results in terms of power with a low FNR, does not succeed in controlling the FDR. This can be explained by the low ($n = n_i = 4$) number of samples per individual (i.e. gene) and therefore by the limited experimental information which is available to the researcher. We are not claiming the reference approaches are dominated by the approaches based on finite mixtures or the counting distribution; in fact, the data are generated from a specific model, even if the simulation design is likely to resemble quite well a real microarray setting.

Table 3 shows the results of a different simulation setting where $q_i = (1 - q_0)X + q_0$ and $X \sim Beta < 3, 0.5 >$; in this case, the signal is stronger. With the same number of samples but with a distribution of the q_is more concentrated and asymmetric towards the unit boundary, the Bayesian method still reaches the best performance in terms of FNR within those controlling the FDR. Similar comments apply to the figures resulting from the finite mixtures approach.

	FDR	FNR	R
Discrete Mixture	0.172	0.044	143.96
Non-parametric method	0.009	0.058	90.03
SAM	0.0487	0.079	47.29
SAM-2	0.0467	0.079	47.165
Uncorrected Testing	0.441	0.048	204.37
Uncorrected Testing, only $\bar{f}_{i.} > 2$	0.246	0.047	150.46
BH corrected Testing	0.048	0.099	1.66
BH corrected Testing, only $\bar{f}_{i.} > 2$	0.025	0.099	1.57
Bonferroni corrected Testing	0.036	0.099	0.47
Bonferroni corrected Testing, only $\bar{f}_{i.} > 2$	0.018	0.099	0.45

Table 2: Comparison of methods in simulation, $q_i \sim Beta < 0.5, 0.5 >$, $n = 4$

	FDR	FNR	R
Discrete Mixture	0.117	0.005	216.22
Non parametric method	0.005	0.015	173.20
SAM	0.0338	0.062	84.46
SAM-2	0.031	0.062	84.49
Uncorrected Testing	0.330	0.010	273.71
Uncorrected Testing, only $\bar{f}_{i.} > 2$	0.169	0.010	220.61
BH corrected Testing	0.045	0.092	19.43
BH corrected Testing, only $\bar{f}_{i.} > 2$	0.015	0.092	18.80
Bonferroni corrected Testing	0.023	0.099	1.14
Bonferroni corrected Testing, only $\bar{f}_{i.} > 2$	0.009	0.099	1.11

Table 3: Comparison of methods in simulation, $q_i = (1 - q_0)X + q_0, X \sim Beta < 3, 0.5 >$, $n = 4$

When we move to simulate larger samples sizes per individual such as $n = 8$ things change substantially as shown by results reported in Tables 4 and 5. The available experimental information is now enough to let the finite mixture approach reach better results, which mimic those obtained through the counting distribution.

Even if the method has not been defined to control the FDR, the finite mixture approach is now a reliable competitor of the semi-parametric approach based on the counting distribution; the FNR is even lower in one case and the estimated FDR is comparable. Similar results in fact are detailed in Table 5, showing results obtained with a different simulation scheme with $q_i = (1 - q_0)X + q_0$ and $X \sim Beta < 3, 0.5 >$ (the signal is stronger).

	FDR	FNR	R
Discrete Mixture	0.007	0.047	111.20
Non-parametric method	0.007	0.049	108.71
SAM	0.023	0.057	93.03
SAM-2	0.018	0.057	93.00
Uncorrected Testing	0.380	0.030	236.21
Uncorrected Testing, only $\bar{f}_{i.} > 2$	0.198	0.030	181.49
BH corrected Testing	0.048	0.049	113.81
BH corrected Testing, only $\bar{f}_{i.} > 2$	0.023	0.049	110.82
Bonferroni corrected Testing	0.001	0.083	37.76
Bonferroni corrected Testing, only $\bar{f}_{i.} > 2$	0.000	0.083	37.73

Table 4: Comparison of methods in simulation, $q_i \sim Beta < 0.5, 0.5 >$, $n = 8$

	FDR	FNR	R
Discrete Mixture	0.003	0.005	191.66
Non parametric method	0.048	0.004	218.43
SAM	0.023	0.020	166.50
SAM-2	0.016	0.020	166.233
Uncorrected Testing	0.310	0.001	288.58
Uncorrected Testing, only $\bar{f}_{i.} > 2$	0.149	0.006	233.78
BH corrected Testing	0.022	0.004	201.86
BH corrected Testing, only $\bar{f}_{i.} > 2$	0.022	0.004	196.77
Bonferroni corrected Testing	0.000	0.058	89.26
Bonferroni corrected Testing, only $\bar{f}_{i.} > 2$	0.000	0.058	89.26

Table 5: Comparison of methods in simulation, $q_i = (1-q_0)X + q_0, X \sim Beta < 3, 0.5 >$, $n = 8$.

6.3 Comparison of procedures via classification of colon tumor

Cancer classification has greatly improved during last few years, thanks to the development of more general approaches for class discovery or class prediction. The approach to cancer classification based on gene expression monitoring by DNA microarrays has been firstly described and applied to human acute leukemia by Golub *et al.* (1999). The availability of class discovery procedures which automatically separate acute Myeloid leukemia from acute Lymphoblastic leukemia raised a tremendous interest in the field of microarray data statistical analysis. Since this seminal paper, various proposals have been introduced demonstrating the feasibility of cancer classification based solely on gene expression monitoring. In this context, an important goal is to have reliable statistical methods which help the researcher to correctly classify the analyzed tissues using the available gene expression profiling. Data referred to as *Colon Tumor Data* (CTD) come from Alon *et al.* (1999), and represent a well known benchmark dataset for genes selection and discrimination. The dataset consists of 2000 genes recorded on 62 individuals, 22 safe and 40 ill of colon cancer.

With this dataset we aim at showing that improving on the standard practice of filtering out genes with our proposed procedures instead of the c-fold rule leads not only to a better control and balance of the error rates (FDR and FNR) but also to a better classification performance and better chance of successful discoveries in the genetic field.

We have selected a training set of 30 samples, 15 from normal patients and 15 from ill patients. So, for each of the $G = 2000$ genes a set of $n = 15$ fold changes are computed, with the same matching design considered in *Bioconductor* http://www.bioconductor.org/. This leaves us with a test set of 32 samples, of which 7 are normal and 25 tumoral. This set is then used to estimate the classification error.

In statistical classification it is customary to select only a subset of variables which will be used to build the classifier. In our context, we can think of selecting only those genes which result as significantly differentially expressed through a statistical analysis.

Table 6 shows, for each technique, the number of selected genes and the corresponding estimated classification error when classification is performed using the k-Nearest Neighbor Classifier (Cover and Hart (1967)), with $k = 1$.

Filtering rule	Filtered Genes	Classification Error
Finite Mixture Approach	479	0.125
Non Parametric Approach	73	0.125
SAM	117	0.156
SAM-2	117	0.156
Uncorrected Testing	428	0.156
Uncorrected Testing, only $\bar{f}_{i.} > 2$	23	0.281
BH corrected Testing	77	0.187
BH corrected Testing, only $\bar{f}_{i.} > 2$	13	0.219
Bonferroni corrected Testing	11	0.219
Bonferroni corrected Testing, only $\bar{f}_{i.} > 2$	1	0.219

Table 6: Number of selected genes and estimated classification error for the Alon *et al.* (1999) Colon Tumor Data

If the k-NN classifier is applied with other values of k the proposed methods still outperform the other competitors. In fact, when $k = 3$, the classification error of our proposals does not improve, and is equal to those achieved by SAM, SAM-2, Uncorrected testing and BH corrected testing.

Note however that the non-parametric approach always achieves the smallest classification error with the lowest number of genes (73). Also consider that, since we are not working with the average fold change, there is not a complete overlap with the other methods. For instance, only 35 of the 77 genes selected by the BH method are included in the set of 73 genes chosen by our non-parametric method.

7 Concluding remarks

In this paper we discuss models for the detection of differentially expressed genes in a microarray experiment when two biological conditions are compared. We use a gene-specific random term controlling for variability among genes with respect to the probability of yielding a fold change value over a certain threshold. We propose two different methods: a two-level finite mixture representation of the binary matrix resulting from the dichotomization of the fold changes and a nonparametric approach based on the counting distribution. The first one is based on maximum likelihood inference which is easy to implement through an EM-type algorithm. The second one is based on the counting distribution, i.e. on modeling the number of exceedances and inference is derived under a Bayesian approach. The performance of the two different selection rules has been discussed by analyzing a well known benchmark dataset (Alon *et al.*, 1999), and by performing a large-scale simulation study. Both approaches always exhibit promising results, with the substantial difference that the finite mixture approach needs a higher sample size (more tissues to be compared) to achieve a sufficiently low FDR, while the approach based on the counting distribution seems to guarantee a good FDR control even with a small number of tissues. These promising results encourage us to extend the proposals so that over-expressed as well as under-expressed genes can be dealt with simultaneously. Another aspect which deserves to be more thoroughly investigated concerns the robustness of the selection procedures to the choice of the threshold c. These insights will be pursued in a forthcoming paper.

References

U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**, 6745–6750.

D. Amaratunga and J. Cabrera (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley.

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 (Ser. B)*, **57**, 289–300.

T. Cover and P. Hart (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, **IT-13**, 21–27.

H. Dette and W.J. Studden (1997). *The theory of canonical moments with applications to statistics, probability, and analysis*. Wiley.

S. Dudoit, P.J. Shaffer, and J.C. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

C.R. Genovese and L. Wasserman (2002). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society (Ser. B)*, **64**, 499–518.

E. Olusegun George and Dale Bowman (1995). A full likelihood procedure for analysing exchangeable binary data. *Biometrics*, **51**, 512–523.

M.A. Gieseg, T. Cody, M.Z. Man, S.J. Madore, M.A. Rubin, and E.P. Kaldjian (2002). Expression profiling of humal renal carcinomas with functional taxonomic analysis. *BMC Bioinformatics*, **3**.

W. R. Gilks and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

W. R. (ed.) Gilks, S. (ed.) Richardson, and D. J. (ed.) Spiegelhalter (1998). *Markov Chain Monte Carlo in Practice*. Chapman & Hall Ltd.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, H. Mesirov, J.P. andColler, M.L. Loh, J.R. Downing, M.A. Caligiuri, C. D. Bloomfied, and E.S. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Y. Hochberg and A.C. Tamhane (1987). *Multiple Comparisons Procedures*. Wiley.

M. Kerr and G. Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–202.

G. Parmigiani, E.S. Garret, R. Irizarry, and S.L. Zeger (2003). *The analysis of gene expression data: methods and software*. Springer.

K.S. Pollard and M.J. van der Laan (2003). A method to identify significant clusters in gene expression data. *Tech. Rep. 107*, Division of Biostatistics, UC Berkeley.

C. Sabatti, S.L. Karsten, and D.H. Geschwind (2002). Thresholding rules for recovering a sparse signal fom microarray experiments. *Math. Biosci.*, **176**, 17–34.

M. Schena (2000). *Microarray Biochip Technology*. Eaton.

J.D. Storey, J.E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society (Ser. B)*, **66**, 187–205.

L. Tardella (2002). A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika*, **89, 4**, 807–817.

V.G. Tusher, R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 5116–5121.

R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **8**, 625–637.