An Article Submitted to

Statistical Applications in Genetics and Molecular Biology

Manuscript 1565

A Three Component Latent Class Model for Robust Semiparametric Gene Discovery

Marco Alfo'[*] Alessio Farcomeni[†]

Luca Tardella[‡]

[*] Sapienza - Università di Roma, marco.alfo@uniroma1.it

[†] Sapienza - Università di Roma, alessio.farcomeni@uniroma1.it

[‡] Sapienza - Università di Roma, luca.tardella@uniroma1.it

A Three Component Latent Class Model for Robust Semiparametric Gene Discovery*

Marco Alfo', Alessio Farcomeni, and Luca Tardella

Abstract

We propose a robust model for discovering differentially expressed genes which directly incorporates biological significance, i.e., effect dimension. Using the so-called c-fold rule, we transform the expressions into a nominal observed random variable with three categories: below a fixed lower threshold, above a fixed upper threshold or within the two thresholds. Gene expression data is then transformed into a nominal variable with three levels possibly originated by three different distributions corresponding to under expressed, not differential, and over expressed genes. This leads to a statistical model for a 3-component mixture of trinomial distributions with suitable constraints on the parameter space. In order to obtain the MLE estimates, we show how to implement a constrained EM algorithm with a latent label for the corresponding component of each gene. Different strategies for a statistically significant gene discovery are discussed and compared. We illustrate the method on a little simulation study and a real dataset on multiple sclerosis.

KEYWORDS: differentially expressed genes, effect size, microarray data, mixture model

*We wish to thank two anonymous referees and the associate editor for comments that helped improve the original paper.

1 Introduction

In microarray studies (Amaratunga and Cabrera, 2004) where the main goal is gene discovery, selected genes must fulfill two requirements: *statistical* significance, that is, generalizability of the estimated difference to the population of patients, and *biological* significance, that usually translates into a lower bound for the effect size (Kirk, 2007).

Statistical significance is usually addressed by means of hypothesis testing and p-values, but in microarray experiments one must take into account the multiplicity issue. There are currently a lot of alternative ways for defining an error rate and keeping it under control and one can look at Westfall and Young (1993), and Farcomeni (2008) for a review of recent developments. Many modern procedures privilege in a first stage the control of the False Discovery Rate (FDR), as conceived in Benjamini and Hochberg (1995), and discard only at a second stage those genes with a fold-change between two experimental/biological conditions too close to 1. Formally, it is requested that the ratio between the average expression in one experimental condition is at least c times the average in the other condition, where typically $c = 2$, see e.g. Tusher, Tibshirani, and Chu (2001), Sabatti, Karsten, and Geschwind (2002).

However, applying the criterion on the fold change after testing is inefficient, since it may deflate the power of the multiple testing procedure and even lead to loss of control (see the simulation study in Section 5); testing after discarding small effects is data snooping and may lead to a loss of control of the considered Type I error rate.

Therefore, a possible solution is to address simultaneously statistical significance (i.e., generalizability of the effect to the population) and biological significance (i.e., sufficiently large effect size). The importance of including the effect size into the selection criterion is illustrated also by Yao, Rakhade, Li, Ahmed, Krauss, Draghici, and Leob (2004), who compare “selected” and “validated” genes. Also Zhu, Hero, Qin, and Swaroop (2005) link statistical significance with effect size. Alfò, Farcomeni, and Tardella (2007) and van de Wiel and Kim (2007) do this in the two-classes case with a paired design, briefly discussing generalizations. Other techniques are available that use the c-fold rule together with significance assessment, the most well-known being probably Significance Analysis of Microarrays (SAM) (Tusher et al., 2001).

Alfò et al. (2007) focus on discovering up regulated genes as a two-class problem, merging down regulated with not differentially expressed genes. The discovery of down regulated genes can be implemented with a separate similar strategy. The procedure obviously loses power when merging two clusters of genes. van de Wiel and Kim (2007) incorporate the c-fold rule into FDR estimation, but not in the adopted statistical model, where the observed expressions are simply averaged. In this work, unlike van de Wiel and Kim (2007) and the majority of works on gene discovery where an average fold change is computed and modelled, we show how to extend the approach in Alfò et al. (2007) by directly modelling the observed expression values at a sample level, incorporating the c-fold rule into the model. We believe that use of the average fold change is not robust, especially in a setting where, even after filtering and normalization, systematic noise may still be present in the data. This problem can be easily seen to be less severe when one discretizes at raw expression level, as it is done in this paper and in Alfò et al. (2007).

We focus on experimental settings with the presence of paired cases under two different biological conditions, considering three classes of genes (up, down and not differentially expressed) simultaneously. A strong flexibility is given by the fact that genes are allowed to have an arbitrary set-specific distribution, since we only model the probability that the expression is above or below a threshold. Our approach is somehow linked to the Bayesian mixture models of Lewin, Bochkina, and Richardson (2007), to which we refer the reader also for a review of other existing methods.

The paired design we consider commonly arises in case-control studies with matching as in our motivating example which is based on an original study on multiple sclerosis conducted by the Center for Experimental Neurological Therapy of Sapienza, University of Rome. The study involves 13 couples of homozygotic Italian twins, who are discordant by disease and differential gene expression is measured co-hybridating twins' sera on the same array. The main aim of the data analysis resides in discovering genes which are differentially expressed under the two experimental conditions (diseased, healthy).

The paper is organized as follows: in Section 2 we describe the proposed modeling approach. In Section 3 we show how to derive maximum likelihood estimates. In Section 4 we show how one can guarantee statistical significance by controlling an estimate of the FDR. Biological significance is guaranteed by fixing appropriate expression thresholds.

The method is illustrated by a simulation study in Section 5 and on the twin data set in Section 6. In Section 7 we conclude with a brief discussion.

2 Modeling Framework

We start up assuming that there are m different samples for which a set of G gene expressions have been jointly measured under two different experimental conditions and the corresponding fold-changes have been properly normalized. We denote the subsets of up- and down-regulated genes by $\mathscr{G}_1$ and $\mathscr{G}_{-1}$, with size G_1 and G_{-1} respectively. The set of "uninteresting" genes is denoted by $\mathscr{G}_0$, its cardinality is G_0, where $G_{-1}+G_0+G_1 = G$ and typically $G_0 >> G_1 + G_{-1}$.

In the following, we will denote with f_{ij} the fold change observed for the i-th gene in the j-th sample, $i = 1,\ldots,G$, $j = 1,\ldots,m$. Many authors agree on the fact that the expression measurement scale is often just a comparative scale with no absolute universal reference. Sometimes it might be appropriate to adjust the scale in terms of a reference fold-change based on measurements taken from blank spots or house-keeping gene spots. We assume there are two fixed thresholds c_l and c_u representing lower and upper thresholds used to derive a decision on whether a gene is differentially expressed at a sufficient biologically relevant extent.

Following Alfò et al. (2007), we transform the observed fold change measures, say $f = \{f_{ij}\}$, $i = 1,\ldots,G$, $j = 1,\ldots,m$ into a discrete matrix Y, with generic element y_{ij}, defined according to the following rule:

$$y_{ij} = \begin{cases} \;\;1 & \text{if} \quad \log f_{ij} > \log c_u \\ -1 & \text{if} \quad \log f_{ij} < \log c_l \\ \;\;0 & \qquad \text{otherwise} \end{cases}$$

so that data become an S-valued matrix Y, where $S = \{-1,0,1\}$ corresponding to the i-th gene being, in the j-th slide, over an upper threshold c_u ($y_{ij} = 1$), below a lower threshold c_l ($y_{ij} = -1$) or within the two thresholds (c_l, c_u) ($y_{ij} = 0$).

Conventionally the rule with $1/c_l = c_u = c = 2$ is used to consider genes as functionally important but there is sometimes the need to raise the threshold as large as $c = 10$ and sometimes to keep it as small as $c = 1.5$ (Cheng, Fabrizio, Ge, Longo, and Li, 2007, Millien, Beane, Lenburg, Tsao, Lu, Spira, and Ramirez, 2008).

Our aim is to model this *discrete* outcome Y, by considering that each gene may be drawn from one of the three subsets of genes ($\mathscr{G}_{-1}, \mathscr{G}_0, \mathscr{G}_1$).

Let us denote with Z_{ik} the latent variable indicating whether the i-th gene belongs to the k-th subset $\mathscr{G}_k$, $k \in \{-1,0,1\}$ and

$$\begin{aligned} \theta_{u|k} &= \Pr(f_{ij} > c_u \mid i \in \mathscr{G}_k) = \Pr(y_{ij} = 1 \mid Z_{ik} = 1) \\ \theta_{l|k} &= \Pr(f_{ij} < c_l \mid i \in \mathscr{G}_k) = \Pr(y_{ij} = -1 \mid Z_{ik} = 1) \\ \theta_{0|k} &= 1 - \theta_{u|k} - \theta_{l|k} = \\ &= \Pr(c_l < f_{ij} < c_u \mid i \in \mathscr{G}_k) = \Pr(y_{ij} = 0 \mid Z_{ik} = 1) \end{aligned}$$

the (conditional) probabilities for the i-th gene in the k-th set to yield a fold change respectively over the upper threshold, below the lower threshold or within the two thresholds. For each k, the usual simplex constraints hold:

$$\theta_{u|k} + \theta_{0|k} + \theta_{l|k} = 1, \qquad k \in \{-1,0,1\}$$

Let us further denote with π_k the prior probability that the i-th gene belongs to the k-th set $\mathscr{G}_k, i = 1, \ldots, G, k \in \{-1, 0, 1\}$.

The main idea behind our approach is that up regulated genes will show $y_{ij} = 1$ more often than not differentially expressed genes, not differentially expressed genes will show $y_{ij} = 1$ more often than down regulated genes. A similar argument can be developed with respect to $y_{ij} = -1$. Not differentially expressed genes have, by definition, a true fold change which is expected to be 1, since expression levels for ill patients should be approximately equal to those of the healthy ones. In this case, observed departures from equality are assumed to be due to random sampling errors, and therefore should not be persistent over experiments. These considerations lead to the constraints which are described in the following.

Indeed it seems natural to expect that up-regulated genes $i \in \mathscr{G}_1$, have a conditional probability ($\theta_{u|1}$) of yielding a fold change above c_u, which is higher than the corresponding probability for non up-regulated genes (i.e. either $i \in \mathscr{G}_0$ or $i \in \mathscr{G}_{-1}$). Taking into account a symmetric argument for the conditional probability of exceeding the lower threshold from below, we will assume:

$$\theta_{u|1} \geq \theta_{u|0} \geq \theta_{u|-1} \tag{1}$$

$$\theta_{l|-1} \geq \theta_{l|0} \geq \theta_{l|1}. \tag{2}$$

The set of constraints can be represented graphically as in Figure 1 both in terms of probabilities and in terms of cumulative probabilities. In fact, when we naturally order the three nominal categories, the set of constraints can be interpreted as a stochastic dominance of the conditional distributions corresponding to each component so that one can write (1) and (2) as follows

$$\theta_{-1} \prec \theta_0 \prec \theta_1$$

Constraints (1)-(2) are also useful in parameter estimation of the proposed finite mixture model since they are used to identify each component.

One could also discuss alternative meaningful constraints. For example up-regulated genes could have a conditional probability of yielding a fold change above c_u, $\theta_{u|1}$ which is higher than the probabilities of yielding a fold change within thresholds or below the lower threshold c_l, respectively $\theta_{0|1}$ and $\theta_{l|1}$. Using the adopted notation, these constraints would be written as follows:

$$\theta_{u|1} > \max\{\theta_{0|1}, \theta_{l|1}\}$$

The same argument applies to down-regulated genes; therefore, we may have:

$$\theta_{l|-1} > \max\{\theta_{0|-1}, \theta_{u|-1}\}$$

However, attention is needed since these constraints may be too strong, especially in those contexts where the number of experiments s is low and we will not pursue this idea further and limit ourselves to assume (1) and (2) in what follows.

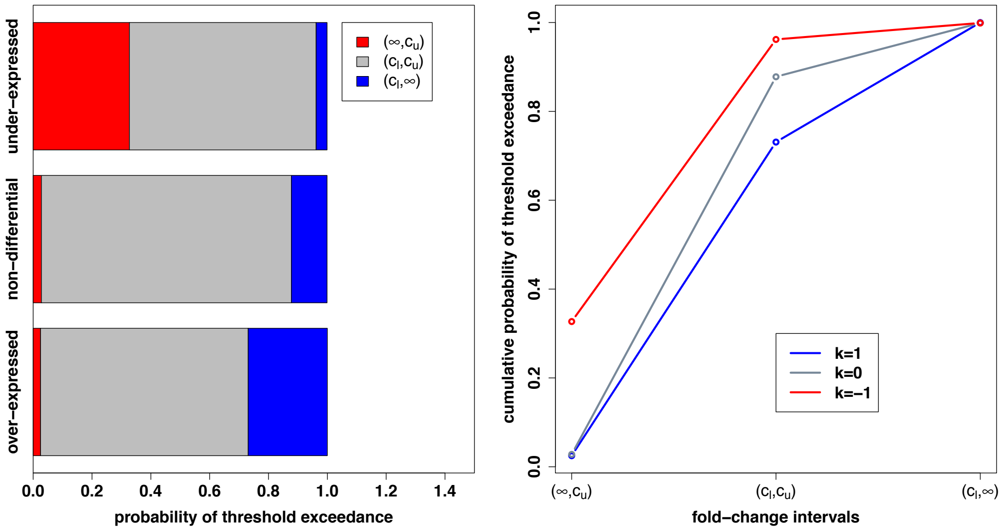

Figure 1: Conditional probabilities of threshold exceedance for the groups of underexpressed, non differentially expressed and overexpressed genes, as estimated for the Twin data set

3 Constrained ML Parameter Estimation

To fix ideas and notation, we first write down the likelihood function and review parameter estimation in the usual finite mixture context without constraints on prior and/or conditional probabilities, assuming conditional independence among experiments.

Augmenting the data with unobserved (latent) labels z_{ik}, indicating whether or not the i-th gene comes from the k-th set, the likelihood function for the *complete* data can be written as follows:

$$\mathrm{L}_c(\Psi) = \prod_{i=1}^{G}\prod_{k}\prod_{j=1}^{m}\left[f(y_{ij} \mid \theta_k)\pi_k\right]^{z_{ik}}$$

where $\Psi = (\Theta, \pi) = \{(\theta_k, \pi_k), k \in \{-1,0,1\}\}$, $\theta_k = \left(\theta_{l|k}, \theta_{0|k}, \theta_{u|k}\right)$ and z_{ik} represent the k-th component label. Besides,

$$f(y_{ij} \mid \theta_k) = (\theta_{u|k})^{\max(y_{ij},0)}(\theta_{l|k})^{-\min(y_{ij},0)}(1-\theta_{u|k}-\theta_{l|k})^{(1-|y_{ij}|)}$$

Hence, the log-likelihood function for complete data can be written as follows:

$$\begin{aligned}\{\ell_c(\Psi)\,|Y\} \quad &\propto \quad \sum_{i,j,k} z_{ik}\log(\pi_k) + \sum_{i,j,k} z_{ik}\left[\max(y_{ij},0)\log(\theta_{u|k}) - \min(y_{ij},0)\log(\theta_{l|k})\right.\\ &+ \quad \left.(1-|y_{ij}|)\log(1-\theta_{u|k}-\theta_{l|k})\right]\end{aligned}$$

In the E-step of the EM algorithm, we define the log-likelihood for *observed* data by calculating the expectation of the log-likelihood for complete data over the unobservable component label vector z_i, given the observed data Y and the current model parameter estimates, say $\Psi^{(t)}$. In other words, at the t-th iteration, z_{ik} is replaced by the conditional expectation (E-step):

$$\mathrm{E}(Z_{ik} \mid y_i, \Psi^{(t-1)}) = w_{ik}^{(t)} = \frac{\pi_k^{(t)} f(y_i; \theta_k^{(t-1)})}{\sum_k \pi_k^{(t)} f(y_i; \theta_k^{(t-1)})}$$

where

$$f(y_i; \theta_k^{(t-1)}) = \prod_{j=1}^{s} f(y_{ij} \mid \theta_k^{(t-1)})$$

Thus $w_{ik}^{(t)}$ represents the posterior probability that the i-th unit belongs to the k-th set, given the observed data and the current parameter estimates. The conditional expectation of the complete log-likelihood given the observed data Y is expressed by the function:

$$\begin{aligned} Q(\Psi|\Psi^{(t-1)}) &= E_{\Theta^{(t-1)}}\{\ell_c(\cdot)|Y\} \propto \sum_{i,j,k} w_{ik}^{(t-1)} \log(\pi_k) + \\ &+ \sum_{i,j,k} w_{ik}^{(t-1)} \left[\max(y_{ij},0)\log(\theta_{u|k}) - \min(y_{ij},0)\log(\theta_{l|k})\right. \\ &+ \left.(1-|y_{ij}|)\log(1-\theta_{u|k}-\theta_{l|k})\right] \end{aligned}$$

The modelling assumptions define a latent class model for a nominal observed variable; estimation can be simply accomplished by using a standard EM algorithm, leading to the following parameter estimates at the M-step:

$$\begin{aligned} \hat{\theta}_{u|k}^{(t)} &= \frac{\sum_i \sum_j \max(y_{ij},0) w_{ik}}{\sum_i \sum_j w_{ik}} \\ \hat{\theta}_{l|k}^{(t)} &= \frac{\sum_i \sum_j -\min(y_{ij},0) w_{ik}}{\sum_i \sum_j w_{ik}} \\ \hat{\theta}_{0|k}^{(t)} &= 1 - \hat{\theta}_{u|k} - \hat{\theta}_{l|k} \\ \hat{\pi}_k^{(t)} &= \frac{\sum_i w_{ik}}{G} \end{aligned}$$

The E and M steps are iterated until convergence.

Constrained maximum likelihood estimation, given the modeling assumptions, can be accomplished by transforming the Fisher-scoring algorithm proposed by Lang and Agresti (1994) into an active-set method. For instance, Vermunt (1999) shows how to transform a simple uni-dimensional Newton-type algorithm for ML estimation with equality constraints into an active-set method. In the present context, the standard EM algorithm should be modified as follows: at each M-step, the inequality constraints which are no longer necessary are de-activated (i.e. if $\hat{\theta}_{u|1} > \hat{\theta}_{u|0}$ the corresponding constraint is simply removed at the present iteration), while the ones which are violated are *activated*, see e.g. Gill, Murray, and Wright (1981), defining equality constraints for the parameter estimates at the current iteration. Obviously the ”activation” step does not break down the monotone increasing behaviour of the loglikelihood, see also Ingrassia and Rocci (2007) for a more thorough discussion.

4 Gene Selection Strategies

In order to select genes, we use a threshold for statistical significance, based on the posterior class probabilities estimated through the EM algorithm.

The estimate of the posterior probability that the i-th gene is in one of the three sets is denoted with $w_{ik} = \Pr(Z_{ik} = 1 \mid y_i, \hat{\Psi}), k \in \{-1,0,1\}$. This probability is conditional on the observed data Y and the MLE $\hat{\Psi}$.

The i-th gene can be assigned to the set corresponding to the highest estimated posterior probability, using then a simple Maximum a Posteriori (MaP) rule. In practice, different probability thresholds may be used to give more conservative lists of *potential* differentially expressed genes. A natural question is how to calibrate the probability threshold and thus the effectiveness of the cutoff. In order to select the probability threshold one can try to keep under control some error rate (such as the FDR) of the resulting procedure, see for instance McLachlan, Do, and Ambroise (2004) and Efron (2007). Recall that the FDR is the expected proportion of false discoveries over the number of selected genes, or zero if no gene is selected. Therefore, along the lines of Newton, Noueiry, Sarkar, and Ahlquist (2004), we may use the following estimates

$$\widehat{FDR}_{[\tau]} = \frac{\sum_{i:(1-\hat{w}_{i0})>\tau} \hat{w}_{i0}}{\text{card}\{i : (1-\hat{w}_{i0}) > \tau\}} \tag{3}$$

where card$\{\mathscr{H}\}$ is the cardinality of the set $\{\mathscr{H}\}$. The threshold probability τ can be selected as the largest τ for which the estimated FDR is below a pre-specified level α, i.e.

$$\hat{\tau}_\alpha = \max\left\{\tau : \widehat{FDR}_{[\tau]} \leq \alpha\right\}.$$

In practice, one need not evaluate all the 2^G possible groups of genes, but at most G groups after ordering with respect to w_{i0}. Genes corresponding to w_{i0} such that $(1-w_{i0}) > \hat{\tau}$ are classified as differentially expressed.

We stress that we are controlling an estimate of the FDR, which depends on the correct estimation of posterior probabilities from the latent class model. The situation is hence different from a probabilistic control of the realized FDR as in the original Benjamini and Hochberg (1995) approach. Nevertheless, as will be seen in simulations, the estimate is often good enough that the realized FDR is also controlled at least on average. We also note that the setting is classification rather than hypothesis testing. In a formal hypothesis testing framework one should summarize the null hypothesis for the i-th gene as $H_{0i} : Z_{i0} = 1$.

We could also be a bit more strict and request not only that genes in G_0 are not classified as differentially expressed, but also that up-regulated genes are not misclassified as down-regulated and the other way around. This is usually referred to as directional, or Type III, error rate control, and in practice reduces to substituting (3) with

$$\widehat{FDR}_{[\tau]} = \frac{\sum_{i:(1-\hat{w}_{i0})>\tau} \hat{w}_{i0} + \sum_{i:(1-\hat{w}_{i0})>\tau \& \hat{w}_{i1}>\hat{w}_{i-1}} \hat{w}_{i-1} + \sum_{i:(1-\hat{w}_{i0})>\tau \& \hat{w}_{i-1}>\hat{w}_{i1}} \hat{w}_{i1}}{\text{card}\{i : (1-\hat{w}_{i0}) > \tau\}}, \tag{4}$$

which leads to a bit more conservative procedure. For a deeper discussion on the directional FDR refer for instance to Sarkar and Zhou (2008). When the directional errors are not of equal importance, it is straightforward to modify (4) so that different thresholds are used to declare up and down regulation.

5 Simulation Study

In this section we apply and compare our method on synthetic data, with a simulation scheme similar to the one discussed by van de Wiel and Kim (2007).

We simulate samples of $G = 4000$ genes with observed (log) expression levels defined by:

$$\log f_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, G \, j = 1, \ldots, m;$$

where i and j index genes and tissue samples, respectively, and $m = 10$ or $m = 50$. Mean log-ratios are drawn from a mixture of three normal distributions:

$$\mu_i \sim \frac{1}{16}\text{N}(-2.0, 0.4^2) + \frac{7}{8}\text{N}(0, 0.5^2) + \frac{1}{16}\text{N}(2.0, 0.4^2)$$

where the distributions correspond to down-, non-, and up-regulated genes. We sample from different error distributions:

1. $\varepsilon_i \sim \mathrm{N}(0, 0.424^2),\ i = 1, \ldots, G$;
2. $\varepsilon_i \sim \mathrm{N}(0, \sigma_i^2) \sigma_i^2 \sim \mathrm{U}(0.212, 0.636),\ i = 1, \ldots, G$;
3. $\varepsilon_i \sim \mathrm{N}(0, 0.\sigma_i^2) \sigma_i^2 \sim \mathrm{Gamma}(\alpha, \beta),\ \alpha = 12.00,\ \beta = 28.30\ i = 1, \ldots, G$, i.e. Gamma distribution with mean and variance equal to those of 2.;
4. $\varepsilon_i \sim \mathrm{N}(0, 0.212^2), i \in G_{-1}$, $\varepsilon_i \sim \mathrm{N}(0, 0.636^2), i \in G_1$, $\varepsilon_i \sim \mathrm{N}(0, 0.424^2), i \in G_0$
5. $\varepsilon_i \sim t_4, i = 1, \ldots, G$, a Student t distribution with 4 degrees of freedom

Weaponshow repeat data generation $B = 250$ times. Each time we fit our model based on discretizing the fold change, for different values of $c_u = 1/c_l$, and proceed to gene discovery by either using a simple Maximum a Posteriori (*MaP*) allocation rule (which is not bound to control the FDR), and to a MaP allocation conditional on controlling the estimated FDR at level $\alpha = 0.1$. The FDR is estimated as in (3). The estimate can be seen as an estimate of expected proportion of falsely selected genes over the number of selected genes. The actual realized FDR is indeed recorded for each generated dataset and finally averaged to get an estimate of the FDR of the procedure. Power is measured through the realized False Non-Discovery Rate (FNR), i.e., the proportion of differentially expressed genes not declared significant over the number of genes not declared significant. Hence, the smaller the FNR, the higher is power. We report the average realized FDR (*FDR*), the average realized FNR (*FNR*). When we use our FDR estimation method, we also report the average estimated FDR (*eFDR*), the standard deviation of the realized FDR (*sdFDR*) and the proportion of times the realized FDR exceeds the nominal FDR level $\alpha = 0.1$ ($p(\mathrm{FDR} > \alpha)$) in our 250 iterations.

In order compare our approach with a possible common practice competitor, we also fit a three latent class model on the continuous fold change, using both a MaP selection rule and an expected FDR controlled rule with c-fold rule applied after classification as it commonly happens. As with our discretized approach, the finite mixture model gives us posterior probabilities for each of the three latent classes, thus we can still base FDR estimation on (3).

A further comparison is made with a non parametric approach, using the rank-based Wilcoxon test. After computing significance levels through the Wilcoxon test, we formally control the FDR using the Benjamini and Hochberg (1995) correction. Once again, the c-fold rule is applied after testing.

Tables 1 and 2 show the results for our discretized model, while tables 3 and 4 the results for the finite mixture model on the continuous outcome. Table 5 shows the results for the non-parametric approach based on Wilcoxon test, when $m = 10$.

Table 1: Simulation results, discrete model, $m = 10$ samples, for different values of the thresholds $c_u = 1/c_l$. The FDR controlled selection aims at controlling the expected FDR at level $\alpha = 0.1$. Reported FDR (FNR) are actual false discovery (non discovery) rates obtained as Monte Carlo average based on $B = 250$ simulations. eFDR is the average of the 250 estimates of FDR based on (3).

	MaP selection		FDR controlled selection				
Case 1	Normal Errors						
$\log_2(c_u)$	FDR	FNR	FDR	eFDR	sdFDR	FNR	$p(\text{FDR} > \alpha)$
1.0	0.429	0.003	0.103	0.100	0.012	0.012	0.33
1.2	0.254	0.017	0.091	0.099	0.007	0.013	0.17
1.5	0.088	0.034	0.088	0.099	0.004	0.025	0.09
2.0	0.045	0.054	0.084	0.099	0.001	0.038	0.03
Case 2	Normal Errors with uniform variance						
1	0.437	0.003	0.107	0.100	0.015	0.013	0.37
1.2	0.254	0.036	0.094	0.099	0.011	0.014	0.17
1.5	0.123	0.044	0.093	0.099	0.008	0.032	0.14
2	0.059	0.052	0.091	0.098	0.004	0.039	0.04
Case 3	Normal Errors with gamma variance						
1	0.443	0.004	0.108	0.099	0.014	0.009	0.41
1.2	0.253	0.007	0.095	0.099	0.008	0.005	0.21
1.5	0.105	0.016	0.092	0.099	0.004	0.014	0.15
2	0.049	0.058	0.089	0.098	0.002	0.049	0.06
Case 4	Heteroschedastic Normal Errors						
1	0.448	0.006	0.117	0.100	0.012	0.012	0.53
1.2	0.254	0.005	0.113	0.099	0.009	0.003	0.47
1.5	0.101	0.018	0.094	0.099	0.004	0.011	0.15
2	0.045	0.058	0.086	0.099	0.002	0.051	0.07
Case 5	Student t errors						
1	0.564	0.007	0.108	0.099	0.013	0.010	0.39
1.2	0.432	0.009	0.099	0.099	0.013	0.011	0.29
1.5	0.112	0.008	0.091	0.100	0.015	0.009	0.21
2	0.079	0.009	0.091	0.099	0.003	0.009	0.07

From the tables it can be seen that our FDR controlling method (i) yields on average a Type I error rate below the fixed nominal $\alpha = 0.1$ for different choices of the (c_l, c_u) thresholds with a slight approximate overestimation only for the lowest threshold and, more importantly, (ii) achieves approximately the same error rate independently of the true error distribution. This is one of the expected consequence of the coarsening approach we propose, which is somewhat robust with respect to the error distribution. The MaP strategy does not necessarily control the FDR (and does not even attempt to do so), even if using a higher cutoff for the effect size increases the odds of a lower FDR. The only appreciable change of performance

Table 2: Simulation results, discrete model, $m = 50$ samples, for different values of the thresholds $c_u = 1/c_l$. The FDR controlled selection aims at controlling the expected FDR at level $\alpha = 0.1$. Reported FDR (FNR) are actual false discovery (non discovery) rates obtained as Monte Carlo average based on $B = 250$ simulations. eFDR is the average of the 250 estimates of FDR based on (3).

	MaP selection		FDR controlled selection				
Case 1	Normal Errors						
$\log_2(c_u)$	FDR	FNR	FDR	eFDR	sdFDR	FNR	$p(\text{FDR} > \alpha)$
1.0	0.395	0.000	0.112	0.096	0.007	0.001	0.48
1.2	0.170	0.000	0.097	0.096	0.007	0.001	0.32
1.5	0.087	0.001	0.097	0.099	0.007	0.001	0.28
2.0	0.065	0.004	0.096	0.091	0.006	0.001	0.21
Case 2	Normal Errors with uniform variance						
1	0.507	0.007	0.112	0.095	0.007	0.016	0.49
1.2	0.337	0.013	0.107	0.094	0.005	0.023	0.34
1.5	0.115	0.019	0.095	0.094	0.005	0.020	0.20
2	0.069	0.036	0.097	0.091	0.004	0.014	0.14
Case 3	Normal Errors with gamma variance						
1.0	0.482	0.006	0.119	0.096	0.007	0.018	0.37
1.2	0.313	0.008	0.098	0.095	0.004	0.016	0.26
1.5	0.107	0.019	0.097	0.095	0.004	0.019	0.21
2.0	0.096	0.024	0.096	0.094	0.004	0.024	0.13
Case 4	Heteroschedastic Normal Errors						
1	0.401	0.000	0.098	0.096	0.009	0.001	0.29
1.2	0.170	0.000	0.095	0.096	0.004	0.001	0.24
1.5	0.048	0.000	0.097	0.095	0.004	0.001	0.17
2	0.044	0.000	0.095	0.094	0.004	0.001	0.08
Case 5	Student t errors						
1	0.513	0.004	0.100	0.093	0.001	0.004	0.27
1.2	0.305	0.006	0.100	0.092	0.001	0.006	0.23
1.5	0.082	0.005	0.095	0.099	0.001	0.001	0.14
2	0.038	0.004	0.094	0.098	0.001	0.005	0.06

in increasing the number of samples m from 10 to 50 is the shrinkage of the distribution of the realized false discovery proportion around the expected nominal level. Indeed the average FDR control was already satisfactory with $m = 10$. The reduced standard deviation of the actual FDR yields a less distant (on average) actual FDR with respect to the nominal target. The increase in sample size m do not affect dramatically the proportion of simulations where the actual FDR exceeds the nominal value 0.1 which for our method is always below 0.5.

We stress that the FDR controlling procedure is always preferable: with fixed lower cutoffs τ, the observed Type I error rate for the MaP strategy can be

Table 3: Simulation results, finite mixture model with Gaussian components, $m = 10$ samples, for different values of the thresholds $c_u = 1/c_l$. The FDR controlled selection aims at controlling the expected FDR at level $\alpha = 0.1$. Reported FDR (FNR) are actual false discovery (non discovery) rates obtained as Monte Carlo average based on $B = 250$ simulations. eFDR is the average of the 250 estimates of FDR. The FDR is estimated as in (3).

	MaP selection		FDR controlled selection				
Case 1	Normal Errors						
$\log_2(c_u)$	FDR	FNR	FDR	eFDR	sdFDR	FNR	$p(\text{FDR} > \alpha)$
1.0	0.680	0.000	0.112	0.100	0.004	0.001	0.58
1.2	0.683	0.000	0.118	0.100	0.003	0.001	0.71
1.5	0.689	0.000	0.125	0.100	0.003	0.001	0.88
2.0	0.696	0.000	0.152	0.100	0.020	0.001	0.92
Case 2	Normal Errors with uniform variance						
1	0.752	0.011	0.100	0.100	0.011	0.059	0.35
1.2	0.746	0.003	0.114	0.100	0.008	0.037	0.68
1.5	0.745	0.002	0.139	0.100	0.008	0.029	0.95
2	0.748	0.002	0.177	0.100	0.010	0.026	0.99
Case 3	Normal Errors with gamma variance						
1	0.751	0.010	0.103	0.100	0.009	0.059	0.43
1.2	0.744	0.003	0.115	0.100	0.007	0.055	0.62
1.5	0.744	0.002	0.142	0.100	0.009	0.038	0.99
2	0.747	0.002	0.178	0.100	0.012	0.019	0.99
Case 4	Heteroschedastic Normal Errors						
1	0.681	0.000	0.092	0.100	0.004	0.001	0.21
1.2	0.684	0.000	0.118	0.100	0.003	0.001	0.57
1.5	0.690	0.000	0.135	0.100	0.003	0.001	0.83
2	0.697	0.000	0.181	0.100	0.020	0.001	0.98
Case 5	Student t errors						
1	0.741	0.002	0.107	0.100	0.020	0.010	0.46
1.2	0.738	0.002	0.151	0.100	0.021	0.007	0.88
1.5	0.738	0.002	0.206	0.100	0.006	0.005	0.98
2	0.741	0.002	0.245	0.100	0.013	0.004	1.00

arbitrarily high; with higher cutoffs, the FNR of the FDR controlling procedure is expected to be lower.

When we compare the performance of our approach with the finite mixture model on the continuous log-fold change, we see that, in the latter case, FDR is out of control almost in all cases, both when $m = 10$ and when $m = 50$. The proportion of simulations where the actual FDR exceeds the nominal value 0.1 is often above 0.5, and for some values of the thresholds it is close to 1. This is the negative effect of discarding discoveries by the c-fold filter after testing: the numerator of the FDR

Table 4: Simulation results, finite mixture model with Gaussian components, $m = 50$ samples, for different values of the thresholds $c_u = 1/c_l$. The FDR controlled selection aims at controlling the expected FDR at level $\alpha = 0.1$. Reported FDR (FNR) are actual false discovery (non discovery) rates obtained as Monte Carlo average based on $B = 250$ simulations. eFDR is the average of the 250 estimates of FDR. The FDR is estimated as in (3).

	MaP selection		FDR controlled selection				
Case 1	Normal Errors						
$\log_2(c_u)$	FDR	FNR	FDR	eFDR	sdFDR	FNR	$p(\text{FDR} > \alpha)$
1.0	0.672	0.000	0.094	0.100	0.003	0.001	0.38
1.2	0.676	0.000	0.110	0.100	0.003	0.001	0.62
1.5	0.683	0.000	0.116	0.100	0.003	0.001	0.71
2.0	0.688	0.000	0.147	0.100	0.005	0.001	0.91
Case 2	Normal Errors with uniform variance						
1.0	0.755	0.015	0.097	0.100	0.013	0.060	0.31
1.2	0.745	0.004	0.114	0.100	0.008	0.036	0.64
1.5	0.744	0.002	0.140	0.100	0.009	0.019	0.94
2.0	0.747	0.002	0.166	0.100	0.011	0.009	0.99
Case 3	Normal Errors with gamma variance						
1.0	0.754	0.014	0.102	0.100	0.013	0.059	0.47
1.2	0.745	0.003	0.134	0.100	0.008	0.036	0.89
1.5	0.744	0.002	0.161	0.100	0.009	0.018	0.97
2.0	0.747	0.002	0.187	0.100	0.011	0.013	0.99
Case 4	Heteroschedastic Normal Errors						
1.0	0.672	0.000	0.100	0.100	0.003	0.001	0.37
1.2	0.677	0.000	0.103	0.100	0.003	0.001	0.39
1.5	0.683	0.000	0.115	0.100	0.003	0.001	0.47
2.0	0.689	0.000	0.147	0.100	0.007	0.001	0.79
Case 5	Student t errors						
1.0	0.694	0.000	0.108	0.100	0.004	0.001	0.51
1.2	0.695	0.000	0.136	0.100	0.003	0.001	0.81
1.5	0.700	0.000	0.165	0.100	0.004	0.001	0.98
2.0	0.709	0.000	0.231	0.100	0.008	0.001	1.00

is decreased, but, more importantly, also the denominator; thus leading to a final loss of control. It can be easily illustrated that also applying the c-fold rule *before* testing leads to the same problems. The non-parametric approach shows the same issues, unless a higher threshold is used. When the threshold is set to 1.5 or 2, the FDR is somehow too conservatively controlled and this yields a large FNR.

The comparative power performance of the competing methods in terms of FNR may not be fair since it would require the competing methods to yield a controlled FDR (at least approximately). Indeed in the majority of cases FDR

Table 5: Simulation results, gene selection carried out with a Wilcoxon test, $m = 10$ samples, for different values of the thresholds $c_u = 1/c_l$. The FDR is controlled at level $\alpha = 0.1$ using Benjamini-Hochberg correction. Results are based on $B = 250$ iterations.

	FDR control			
Case 1	Normal Errors			
$\log_2(c_u)$	FDR	sdFDR	FNR	$p(\text{FDR} > \alpha)$
1.0	0.273	0.013	0.001	1.00
1.2	0.128	0.013	0.004	0.99
1.5	0.028	0.007	0.017	0.00
2.0	0.001	0.002	0.067	0.00
Case 2	Normal Errors with uniform variance			
1.0	0.481	0.012	0.006	1.00
1.2	0.320	0.016	0.013	1.00
1.5	0.125	0.015	0.031	0.95
2.0	0.012	0.006	0.067	0.00
Case 3	Normal Errors with gamma variance			
1.0	0.482	0.011	0.006	1.00
1.2	0.320	0.015	0.013	1.00
1.5	0.126	0.015	0.031	0.96
2.0	0.011	0.006	0.067	0.00
Case 4	Heteroschedastic Normal Errors			
1.0	0.273	0.015	0.001	1.00
1.2	0.127	0.012	0.004	0.98
1.5	0.030	0.008	0.017	0.00
2.0	0.001	0.002	0.067	0.00
Case 5	Student t errors			
1.0	0.370	0.012	0.025	1.00
1.2	0.287	0.015	0.026	1.00
1.5	0.158	0.016	0.035	1.00
2.0	0.044	0.015	0.069	0.00

control is not achieved with our competitors. We only point out very few cases where the average FDR and the proportion of times the realized FDR violates the nominal control is approximately the same, which happens for instance when the lowest thresholds $c_l = c_u$ is equal to 1 or 1.2 and for $m = 10$. In Case 5 (Student's t) comparing Table 1 and Table 3 the FDR and FNR are approximately the same and there is no apparent loss of power. Indeed our method gives a slightly higher FNR when the continuous fold-change competitor is well specified (i.e., data are simulated from the same parametric assumption used for the finite mixture model), as in Case 4, but we remark that in real data applications the data generating distribution is not known. For a threshold $c_l = c_u = 1.2$ the FDR and

FNR comparison shows the superiority of our method when data are simulated with an error distribution as in Case 2 or Case 3.

6 Twin Data

In this section we analyze the twin data set, a study involving 13 couples of homozygotic Italian twins, who are discordant by disease. On each slide, two samples of mRNA have been put, one from each twin of the 13 couples, with red dye assigned to the ill twin and green dye assigned to the healthy one. The expression levels of 8570 genes have been recorded and normalized. A fold change between the ill twin and the healthy one has been computed for each couple and each gene, resulting in a 8570 × 13 data matrix.

Due to the very limited sample size, and the likely small genetic variability, standard multiple testing approaches, including SAM at FDR control level of 10%, lead to identification of no differentially expressed genes.

We employed the proposed modeling approach to cluster genes coming from the twin dataset, by setting $\alpha = 0.1$, and considering $B = 1000$ starting values for the parameter vector Θ:

$$\Theta = \begin{pmatrix} \theta_{u|-1} & \theta_{u|0} & \theta_{u|1} \\ \theta_{0|-1} & \theta_{0|0} & \theta_{0|1} \\ \theta_{l|-1} & \theta_{l|0} & \theta_{l|1} \end{pmatrix} = \begin{pmatrix} \xi^2 & \xi & 1 \\ \xi & 1 & \xi \\ 1 & \xi & \xi^2 \end{pmatrix}$$

where $\xi \in (0,1)$ represents a random starting value drawn from a Uniform distribution on $(0,1)$. When we set $c_u = 1/c_l = 2$, the maximum likelihood estimates we finally obtain are:

$$\begin{aligned} \hat{\Theta} &= \begin{pmatrix} \hat{\theta}_{u|-1} & \hat{\theta}_{u|0} & \hat{\theta}_{u|1} \\ \hat{\theta}_{0|-1} & \hat{\theta}_{0|0} & \hat{\theta}_{0|1} \\ \hat{\theta}_{l|-1} & \hat{\theta}_{l|0} & \hat{\theta}_{l|1} \end{pmatrix} = \begin{pmatrix} 0.025 & 0.028 & 0.327 \\ 0.706 & 0.850 & 0.635 \\ 0.269 & 0.121 & 0.037 \end{pmatrix} \\ \pi &= (\pi_{-1}, \pi_0, \pi_1) = (0.0007, 0.9980, 0.0013). \end{aligned}$$

By adopting a MaP strategy, i.e. allocating each gene to the mixture component with the highest posterior probability, we obtained 7 up-regulated and 2 down-regulated genes, while using a strategy based on controlling estimated FDR at level $\alpha = 0.1$, 2 and 0, respectively.

In Figure 2 we show a histogram of the estimated posterior probabilities of belonging to the set of *non differentially expressed* genes. We can see that for a very small number of genes, this posterior probability is very low: even though very few genes are selected, we have only a small uncertainty related to their classification. As far as the starting values are concerned, obviously our starting strategy is not fail-safe and needs to be checked in each empirical case. However, by fine-tuning the convergence threshold and selecting a different set of starting points (completely random in a small scale study), we have not observed any significant change in the posterior probability estimates and/or in the conclusion that, if any, only few over-expressed genes are present. Finally, one of the two up-regulated genes selected has been later undergone a PCR validation and has then been confirmed to be a true discovery, while the other selected genes have not undergone any biological validation yet.

To get further insight on the analysis, we employed the proposed model by using the same starting values strategy outlined before, with a different choice of lower and upper thresholds, namely, by setting $c_u = 1/c_l = 1.5$. This procedure could help us detecting a larger set of genes to be subsequently validated through more specific tools. In this case, the maximum likelihood estimates for component specific parameters and prior probabilities are:

$$\begin{aligned}
\hat{\Theta} &= \begin{pmatrix} \hat{\theta}_{u|-1} & \hat{\theta}_{u|0} & \hat{\theta}_{u|1} \\ \hat{\theta}_{0|-1} & \hat{\theta}_{0|0} & \hat{\theta}_{0|1} \\ \hat{\theta}_{l|-1} & \hat{\theta}_{l|0} & \hat{\theta}_{l|1} \end{pmatrix} = \begin{pmatrix} 0.0003 & 0.0638 & 0.3631 \\ 0.6046 & 0.8461 & 0.6360 \\ 0.3952 & 0.0901 & 0.0009 \end{pmatrix} \\
\hat{\pi} &= (\hat{\pi}_{-1}, \hat{\pi}_0, \hat{\pi}_1) = (0.0025, 0.9964, 0.0011).
\end{aligned}$$

which are quite different (especially for what concerns the Θ parameters) from the estimates obtained by setting $c_u = 1/c_l = 2$. By adopting a simple MaP strategy, we obtained 32 down-regulated and 20 up-regulated genes, while using the proposed strategy controlling estimated FDR at the level $\alpha = 0.1$, 6 (down-regulated) and 3 (up-regulated), respectively. The two sets of up-regulated genes discovered by varying the c_u and c_l values share the same 2 up-regulated genes.

7 Discussion and Further Research

We proposed a modeling approach to flexibly and reliably implement gene discovery with microarray data. The approach is semiparametric in that it relies on data coarsening and avoids working with average expression values. As illustrated by the small simulation study, coarsening makes the approach reliable in terms of FDR control and robust with respect to the error distribution.

Our model simultaneously selects biologically and statistically significant genes and does not involve averaging out the fold changes. The main advantage of this feature is illustrated in the application, where biological variability for each

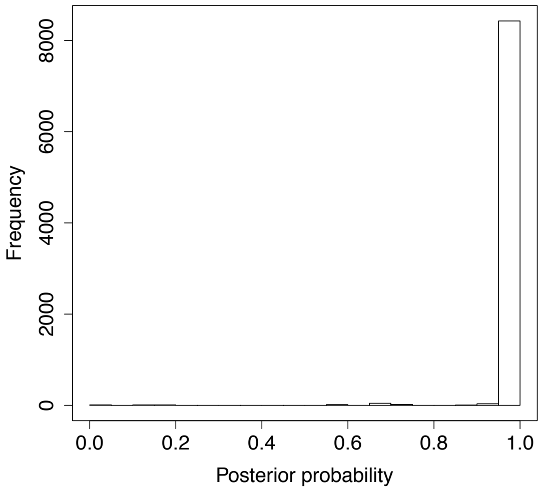

Figure 2: Distribution of posterior probability of belonging to the set of non differentially expressed genes for the Twin data set. Thresholds $c_u = 1/c_l = 2$

paired twin couple is likely to be much smaller than the variability among twin couples. We believe that in a situation like this averaging out the fold change may easily lead to overlook important information as actually discovered with the PCR confirmation in the twins' data.

The biological thresholds c_l and c_u are often readily available for routinely practiced microarray experiments with standard chips. Nevertheless there is no universal rule suitable for all experiments since microarray data can be considered as a proxy to detect only comparative measures which, of course, depend on the measurement scale adopted. Indeed microarray experiments are usually affected by specific experimental conditions as well as other artifacts. Different combinations of preprocessing steps are often used to remove them determining a case-specific final measurement scale of the normalized data used for the statistical analysis. This often results in an artificial expansion or compression of fold-change values which may possibly act nonlinearly on the observed range within and between experimental conditions. Hence, we do not give precise recommendations on the choice of c_l and c_u but suggest performing a sensitivity analysis by varying the thresholds. Sensitivity analysis can prevent overlooking interesting genes.

We remark that the final judgement on the choice of thresholds is mainly a matter of biological meaning rather than statistical properties. For the real data set one should consider biological significance. This can be done by direct PCR validation in case the selected list is short as in our Twin Data case study or by a preliminary understanding of the biological themes in long lists of selected genes via gene-enrichment and functional annotation analysis.

In the present formulation our approach can deal only with paired data. The unpaired case is a direct extension which can be performed along the lines of van de Wiel and Kim (2007, Section 7), simply by averaging out replicates and subtracting averages for the two treatments.

Finally, we note that our model allows for dependence within groups of genes belonging to one or another latent state. A final possibility for further work resides in exploring the effects of using a further general random effect shared by all genes, which could be used to capture general dependence.

References

Alfò, M., A. Farcomeni, and L. Tardella (2007): "Robust semiparametric mixing for detecting differentially expressed genes in microarray experiments," *Computational Statistics & Data Analysis*, 51, 5253–5265.

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

Benjamini, Y. 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.

Cheng, C., P. Fabrizio, H. Ge, V. D. Longo, and L. M. Li (2007): "Inference of transcription modification in long-live yeast strains from their expression profiles," *BMC Genomics*, 8, 219.

Efron, B. (2007): "Size, power and false discovery rates," *Ann. Statist.*, 35, 1351–1377.

Farcomeni, A. (2008): "A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion," *Statistical Methods in Medical Research*, 17, 347–388.

Gill, P., W. Murray, and M. Wright (1981): *Practical Optimization*, Accademic Press.

Ingrassia, S. and R. Rocci (2007): "Constrained monotone EM algorithms for finite mixture of multivariate gaussians," *Computational Statistics & Data Analysis*, 51, 5339 – 5351.

Kirk, R. (2007): "Effect magnitude: a different focus," *Journal of Statistical Planning and Inference*, 137, 1634–1646.

Lang, J. and A. Agresti (1994): "Simultaneously modeling joint and marginal distributions of multivariate categorical responses," *Journal of the American Statistical Association*, 89, 625–632.

Lewin, A., N. Bochkina, and S. Richardson (2007): “Fully bayesian mixture model for differential gene expression: Simulations and model checks,” *Statistical Applications in Genetics and Molecular Biology*, 6, 36.

McLachlan, G., K.-A. Do, and C. Ambroise (2004): *Analyzing Microarray Gene Expression Data*, Hoboken, New Jersey: Wiley.

Millien, G., J. Beane, M. Lenburg, P. N. Tsao, J. Lu, A. Spira, and M. I. Ramirez (2008): “Characterization of the mid-foregut transcriptome identifies genes regulated during lung bud induction,” *Gene Expr. Patterns*, 8, 124–139.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004): “Detecting differential gene expression with a semiparametric hierarchical mixture method,” *Biostatistics*, 5, 155–176.

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

Sarkar, S. K. and T. Zhou (2008): “Controlling Bayes directional false discovery rate in random effects model,” *Journal of Statistical Planning and Inference*, 138, 682–693.

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

van de Wiel, M. and K. Kim (2007): “Estimating the false discovery rate using nonparametric deconvolution,” *Biometrics*, 63, 806–815.

Vermunt, J. (1999): “A general class of nonparametric models for ordinal categorical data,” *Sociological Methodology*, 29, 187–223.

Westfall, P. H. and S. S. Young (1993): *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley.

Yao, B., S. Rakhade, Q. Li, S. Ahmed, R. Krauss, S. Draghici, and J. Leob (2004): “Accuracy of cDNA microarray methods to detect small gene expression changes induced by neuregulin on breast epithelial cells,” *BMC Bioinformatics*, 5, 99.

Zhu, D., A. Hero, Z. Qin, and A. Swaroop (2005): “High throughput screening of co-expressed gene pairs with controlled false discovery rate (FDR) and minimum acceptable strength (MAS),” *Journal of Computational Biology*, 12, 1029–1045.