Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 15

Semi-Parametric Differential Expression Analysis via Partial Mixture Estimation

David Rossell[*] Rudy Guerra[†]
Clayton Scott[‡]

[*] Institute for Research in Biomedicine of Barcelona, rosselldavid@gmail.com
[†] Rice University, rguerra@rice.edu
[‡] University of Michigan, Ann Arbor, clayscot@umich.edu

Semi-Parametric Differential Expression Analysis via Partial Mixture Estimation*

David Rossell, Rudy Guerra, and Clayton Scott

Abstract

We develop an approach for microarray differential expression analysis, i.e. identifying genes whose expression levels differ between two or more groups. Current approaches to inference rely either on full parametric assumptions or on permutation-based techniques for sampling under the null distribution. In some situations, however, a full parametric model cannot be justified, or the sample size per group is too small for permutation methods to be valid.

We propose a semi-parametric framework based on partial mixture estimation which only requires a parametric assumption for the null (equally expressed) distribution and can handle small sample sizes where permutation methods break down. We develop two novel improvements of Scott's minimum integrated square error criterion for partial mixture estimation [Scott, 2004a,b]. As a side benefit, we obtain interpretable and closed-form estimates for the proportion of EE genes. Pseudo-Bayesian and frequentist procedures for controlling the false discovery rate are given. Results from simulations and real datasets indicate that our approach can provide substantial advantages for small sample sizes over the SAM method of Tusher et al. [2001], the empirical Bayes procedure of Efron and Tibshirani [2002], the mixture of normals of Pan et al. [2003] and a t-test with p-value adjustment [Dudoit et al., 2003] to control the FDR [Benjamini and Hochberg, 1995].

KEYWORDS: microarrays, partial mixture, differential expression, density estimation

*We thank David W. Scott for his useful comments.

1 INTRODUCTION

The development of high-throughput biotechnology, such as microarrays, has made possible the collection of large amounts of genomic data. In turn, we face new challenges in applying existing statistical methods or developing novel ones to properly analyze this new generation of biological data. One problem in bioinformatics that has attracted considerable interest is gene differential expression analysis, *i.e.*, the comparison of gene expression between groups defined by treatments or biological conditions. For example, the apo-AI gene experiment (Callow *et al.*, 2000) discussed in Section 5 compares expression levels between apo-AI knock out mice and inbred control mice. The problem of biological interest is to determine which genes are differentially expressed (DE) between these two groups and which are equally expressed (EE). Statistically, the goal is to detect as many DE genes as possible while not having too many false positive genes, *i.e.*, genes declared DE that are, in fact, EE.

More formally, suppose that expression levels for n genes are measured and normalized to account for systematic biases (Dudoit *et al.*, 2002b). To discriminate EE and DE genes we compute a test statistic $\mathbf{X}$ for each gene; for example, a t-statistic comparing mean log-ratios (of red-to-green intensity measurements) between two groups. Other statistics are discussed by Efron *et al.* (2001), Efron and Tibshirani (2002), Tusher *et al.* (2001) and Smyth (2004).

The n observed values of the statistic $\mathbf{x}_1, \ldots, \mathbf{x}_n$ may be viewed as identically distributed (and possibly dependent) realizations of a common marginal mixture density

$$f(\mathbf{x}) = w f_0(\mathbf{x}) + (1-w) f_1(\mathbf{x}),\ \mathbf{x} \in S_x \subseteq \Re^p, \tag{1.1}$$

where w is the proportion of EE genes, and f_0 and f_1 are the densities of the test statistic for EE and DE genes, respectively.

The statistical challenge is to estimate some or all of the components of this mixture (or functions thereof) in order to draw inferences about the genes under consideration. Dudoit *et al.* (2002b) review some approaches based on computing t-tests for each gene and adjusting the raw p-values for multiple comparisons. Tusher *et al.* (2001) introduced the significance analysis of microarrays (SAM), which obtains raw p-values through permutations and uses them to compute (local) q-values, which can be used to control the FDR. Efron and Tibshirani (2002) and Efron (2004) proposed a non-parametric empirical Bayes approach and Pan *et al.*

[0] C.S. was partially supported by an NSF Vertical Integration of Research and Education grant (No. 0240058) as a postdoctoral fellow at Rice University.

(2003) proposed modeling f_0 and f via mixtures of normals. Newton *et al.* (2001), Kendziorski *et al.* (2003), Newton and Kendziorski (2003) and Newton *et al.* (2004) introduce parametric empirical Bayes hierarchical models in which the parameters arise from a mixture of distributions. Do *et al.* (2005) formulate a fully Bayesian non-parametric mixture model based on permutations that provides posterior probabilities. Storey (2007) developed an extension of the Neyman-Pearson theory of hypothesis testing and proposed the Optimal Discovery Procedure, a method that has some optimality properties although it requires estimating some unknown quantities from the data.

All of the methods mentioned above either make full distributional assumptions (Newton *et al.*, 2001; Kendziorski *et al.*, 2003; Newton and Kendziorski, 2003; Newton *et al.*, 2004) or rely on resampling methods to sample under f_0 (Efron and Tibshirani, 2002; Efron, 2004; Pan *et al.*, 2003; Do *et al.*, 2005; Storey, 2007). In some situations, however, these approaches can be difficult to justify. Models for the DE distribution f_1 may be difficult to estimate when very few genes are DE. Permutation methods, on the other hand, need more than a handful of microarrays per group in order to avoid a coarse representation of the test statistic under the null hypothesis. With two groups of three subjects each there are only 10 distinct permutations of the microarrays. Yet sample size is often limited by cost, time, or subject availability, and hence methods are needed to analyze differential expression when sample sizes are small and full parametric models are not appropriate.

In this paper we propose a semi-parametric approach that imposes no structure on f_1 and is especially useful for small sample sizes. It builds on the work of partial mixture estimation by Scott (2004a,b), which is the problem of estimating w and the parameters defining f_0 in (1.1), given a sample from the mixture f. Recently, this approach was used in wavelet applications to denoise signals while relaxing some of the distributional assumptions that are typically made by other methods (Scott, 2006). We develop two improved variants of Scott's original L$_2$E approach to partial mixture estimation which we call *weighted L$_2$E* (WL$_2$E) and *fixed-component WL$_2$E*, respectively. The latter variant provides closed-form and interpretable estimates of the proportion of EE genes.

Our approach requires making only two assumptions. First, we assume a parametric form for f_0. Second, we assume that the test statistics are identically distributed across all EE genes (possibly with dependence). Specifying a parametric family for f_0 is often not unreasonable because EE data are typically much more abundant and better behaved than DE data. For some variants of our approach there is the third implicit assumption that most genes are EE. This is not the case for our *fixed-component* variant of WL$_2$E. Obtaining identically distributed statistics can be achieved via appropriate data pre-processing or normalization procedures, as we illustrate with real data in Section 5. For more detail on normalization procedures see Dudoit *et al.* (2002b).

In the next section we describe the L_2E, WL_2E and fixed-component WL_2E criteria for partial mixture estimation. In Section 3, we describe the application of these algorithms to differential expression analysis. To adjust for multiple testing we present frequentist and Bayesian methods that control the false discovery rate (FDR) at a desired level. In Section 4 we show with simulated data and two real datasets that our method outperforms several existing approaches. A discussion is offered in the concluding section. We provide R code for our methods at `http://rosselldavid.googlepages.com`. Throughout this work we used R version 2.6.1.

2 PARTIAL MIXTURE ESTIMATION

In general, any test statistic that we choose to test for differential expression can be modeled as a mixture of the form presented in (1.1). Suppose that we are willing to assume some parametric form for f_0, but we do not want to impose any restrictions on f_1. In many situations some parametric choices come as a natural assumption; for example, the marginal distribution of many statistics is approximately normal as the number of measurements increases. Also note that if we expect most of the genes to be EE, we can assess the parametric assumption. For example, we can assess normality graphically with a qq-normal plot.

Partial mixture estimation is the problem of estimating w and f_0 only, without estimating the remaining components of the mixture. In Section 2.1 we review the original L_2E criterion (Scott, 2004a,b) for partial mixture estimation. In Section 2.2 we develop a new criterion to obtain partial mixture fits, WL_2E, that improves some shortcomings of L_2E. In Section 2.3 we consider the possibility of treating f_0 as known and only estimating w, and we discuss how it can further improve inference. We distinguish partial mixture estimation from standard robust estimation (Hampel *et al.*, 1986; Huber, 1981), which seeks to estimate f_0 but not w. Robust estimators such as M-estimators typically perform well when the DE component f_1 is well separated from f_0. This is often not the case in practice, however, and simultaneously estimating w can improve the estimate of f_0. Furthermore, most procedures to control the FDR (frequentist or Bayesian) require an estimate of w (Section 3). Over-estimating w can result in an overly conservative procedure leading to too many false-negatives, while under-estimation can result in too many false-positives.

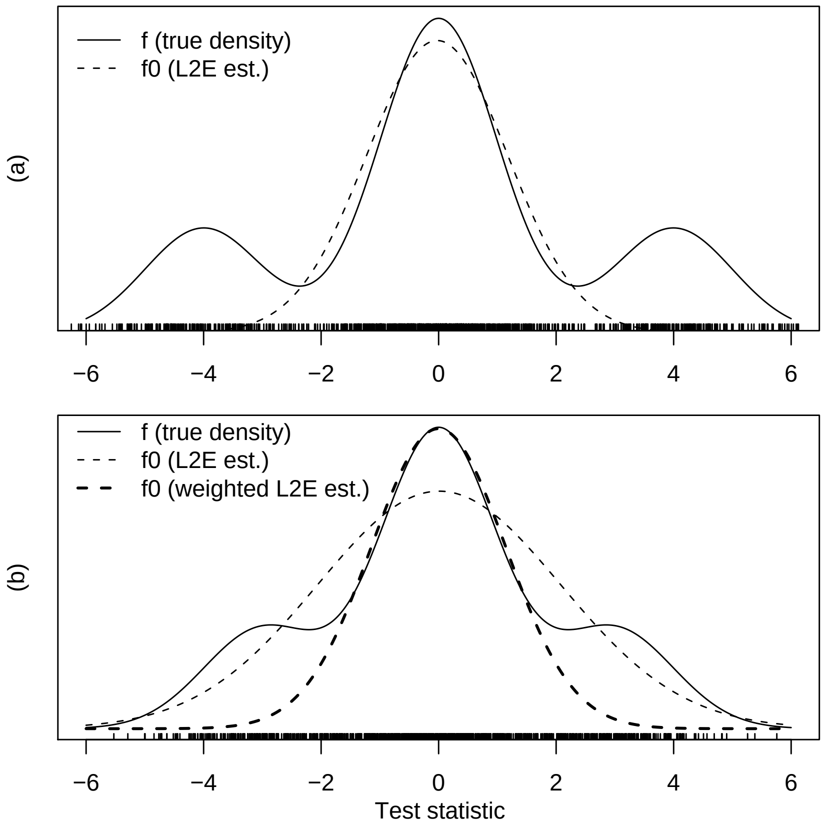

Figure 1: L_2E fit example. (a): $f = .6N(0,1) + .2N(-4,1) + .2N(4,1)$. L_2E estimate: $.68N(0.02, 1.20)$. (b) $f = .6N(0,1) + .2N(-3,1) + .2N(3,1)$. L_2E estimate: $.97N(.03, 1.97)$. WL_2E estimate: $.74N(0.01, 1.25)$. The vertical segments on the x axis indicate the generated test statistic values

2.1 L_2E CRITERION

Before formally defining the approach we illustrate the idea with an example that mimics a differential expression setup. We simulated $n = 1000$ test statistic values, 60% corresponding to EE genes that follow a $N(0,1)$ distribution, 20% to under-expressed genes following a $N(-4,1)$, and 20% to over-expressed genes following a $N(4,1)$. This represents an ideal scenario where there is a clear separation between EE and DE genes. As shown in Figure 1(a) the L_2E fit provides a local estimate of the overall distribution around 0, thereby effectively estimating the distribution of the EE genes. The estimate $\hat{w} = 0.68$ indicates that the L_2E fit finds 32% of the data as not arising from f_0. The true percentage of DE genes is 40%.

To emphasize that as a parametric distribution f is indexed by a (multi-dimensional) parameter, $\boldsymbol{\theta}$, we denote it as $f_{\boldsymbol{\theta}}$. Scott (2001) proposed the L$_2$E criterion for parametric density estimation, finding that it is asymptotically efficient and robust to departures from the assumed model and to the presence of outliers. Note that for partial mixture estimation purposes the sample size is the number of genes, and hence asymptotic considerations become relevant. Scott (2004a,b) used the L$_2$E criterion to estimate f about its central mass with a partial mixture component $wf_{\boldsymbol{\theta}}$. The approach seeks to minimize the integrated squared difference or L_2 distance between the true density f and its local approximation $wf_{\boldsymbol{\theta}}$:

$$\int_{S_x} \left(wf_{\boldsymbol{\theta}}(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x} = w^2 \int_{S_x} f_{\boldsymbol{\theta}}^2(\mathbf{x}) d\mathbf{x} - 2w \int_{S_x} f_{\boldsymbol{\theta}}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} + C, \quad (2.1)$$

where C is a constant free of w and $\boldsymbol{\theta}$ and hence can be ignored in the optimization. If $f(\cdot)$ belongs to the assumed parametric family, *i.e.* $f(\mathbf{x}) = f_{\boldsymbol{\theta}_0}(\mathbf{x})\ \forall \mathbf{x} \in S_x$ for some $\boldsymbol{\theta}_0$, then (2.1) achieves it minimum at $\hat{w} = 1$, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

In differential expression analysis a common assumption is that most genes are EE. This assumption corresponds to f_θ being the largest component of the mixture model. By requiring the estimated component to integrate to w instead of 1, f_θ will tend to approximate the largest component of f instead of blurring all mixture components together. Of course, the approximation will depend on the degree of separation between the components. The WL$_2$E method developed in Section 2.2 is less dependent on the components being well-separated. The fixed-component WL$_2$E from Section 2.3 treats f_θ as fixed and estimates only w, hence being robust by design to the degree of separation.

The first integral in (2.1) has a closed form for several common distributions, including the multivariate normal and t (Wand and Jones, 1995). The second integral is the expected value of $f_{\boldsymbol{\theta}}(\mathbf{X})$ when $\mathbf{X}$ arises from the mixture density in (1.1), and it can be approximated by its corresponding sample average since the mathematical form of f_θ is assumed known.

The L$_2$E partial mixture estimate is thus obtained by minimizing

$$w^2 \int f_{\boldsymbol{\theta}}(\mathbf{x})^2 d\mathbf{x} - \frac{2w}{n} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}(\mathbf{x}_i) \quad (2.2)$$

with respect to w and $\boldsymbol{\theta}$. It is important to note that the criterion in (2.2) does not require the genes to be independent. When $f_{\boldsymbol{\theta}}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance Σ, the criterion simplifies to (Wand and Jones, 1995)

$$\frac{w^2}{2^d \pi^{d/2} |\Sigma|^{1/2}} - \frac{2w}{n} \sum_{i=1}^{n} f_{(\boldsymbol{\mu},\Sigma)}(\mathbf{x}_i). \quad (2.3)$$

To find $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ and $\hat{w}$ that minimize this function we use the R function `nlmin` for general nonlinear minimization. In general, (2.2) may have local optima. In our experience with univariate test statistics we have always been able to avoid local optima by initializing $w = 1$ and $\boldsymbol{\theta}$ to the maximum likelihood estimate, although multivariate statistics may require more care.

The normality assumption may be replaced with other parametric models. For example, one could model $f_{\boldsymbol{\theta}}$ with a multivariate t distribution and estimate its location, scale and degrees of freedom from the data. We explored the t model but because of the relatively heavier tails, f_θ tended to capture more DE genes and thus lead to a higher false-negative rate. However, we did find the t model useful when the value of $\boldsymbol{\theta}$ is specified in the fixed-component WL$_2$E (Section 2.3).

2.2 WEIGHTED L$_2$E CRITERION

As seen in the example of Figure 1, when the EE genes provide values of the test statistic well separated from those of the DE genes, the local estimate obtained via the L$_2$E criterion can capture the behavior of the EE genes quite well. However, when the separation is not so clear problems can arise. To illustrate this point we generated $n = 1000$ test statistic values, 60% representing EE genes from a $N(0, 1)$, 20% under-expressed from a $N(-3, 1)$ and 20% over-expressed from a $N(3, 1)$. As shown in Figure 1(b), the L$_2$E estimate covers most of the support of f and thus fails to capture its central mass around zero. L$_2$E estimates the null proportion w as $\hat{w} = 0.97$, while the true value is $w = 0.6$.

To overcome this problem, we propose a new criterion. Suppose $\hat{\boldsymbol{\theta}}$ is the L$_2$E estimate minimizing (2.2). We now seek to minimize a *weighted L_2* distance

$$\int f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) \left(w f_{\boldsymbol{\theta}}(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x}. \tag{2.4}$$

The weighting factor $f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})$ assumes the initial L$_2$E estimate was “reasonably close,” and places more emphasis on correctly learning the density in the region with the highest probability density. This process can then be repeated, using the updated estimate to specify weighting for a new fit, until the process converges to a fixed point. In our experience convergence is usually achieved within 4 or 5 iterations. We call this criterion *weighted L$_2$E* (WL$_2$E) for partial mixture estimation.

As in Section 2.1, we may expand the integrated weighted squared error as

$$w^2 \int_{\Re^n} f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) f^2_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{x} - 2w \int_{\Re^n} f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) f_{\boldsymbol{\theta}}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} + C \tag{2.5}$$

As before, the first term in (2.5) has a closed form expression for some distributions, in particular for the normal family, while the second term can be approximated by a sample mean. Under the assumption of normality, the WL$_2$E partial mixture estimating criterion is to minimize

$$w^2 \frac{\exp\left\{-\frac{\hat{\boldsymbol{\mu}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}}{2} - \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} + \frac{1}{2}\mathbf{m}'V^{-1}\mathbf{m}\right\}}{(2\pi)^p\,|\hat{\Sigma}|^{\frac{1}{2}}|\Sigma||V|^{\frac{1}{2}}} - \frac{2w}{n}\sum_{i=1}^{n} f_{\hat{\boldsymbol{\mu}},\hat{\Sigma}}(\mathbf{x}_i) f_{\boldsymbol{\mu},\Sigma}(\mathbf{x}_i) \quad (2.6)$$

where minimization is with respect to $(w, \boldsymbol{\mu}, \Sigma)$, where $\mathbf{m} = \left(\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}} + 2\Sigma^{-1}\boldsymbol{\mu}\right)$ and $V = \hat{\Sigma}^{-1} + 2\Sigma^{-1}$.

Returning to the example in Figure 1(b), using WL$_2$E yields $(\hat{\mu}, \hat{\sigma}, \hat{w}) = (0.01, 1.30, 0.76)$, which provides a better local approximation to the true value $(\mu, \sigma, w) = (0, 1, 0.6)$ than the initial L$_2$E estimates $(\hat{\mu}, \hat{\sigma}, \hat{w}) = (0.03, 1.97, 0.97)$. Now we can repeat the process by using the current weighted estimates to weight again and obtain updated estimates. We repeat this process until the change in the parameter estimates is smaller than 1% in square norm. We obtain the final estimate of $(\hat{\mu}, \hat{\sigma}, \hat{w}) = (0.01, 1.25, 0.74)$.

2.3 FIXED-COMPONENT WL$_2$E

In some situations it is reasonable to assume that $\boldsymbol{\theta}$ is (practically) known, and therefore only w needs to be estimated. For example, if x_i is a two-sample t-test statistic or a moderated t-test statistic (Smyth, 2004) it may be reasonable to assume that $f_{\boldsymbol{\theta}}$ is given by a Student's t distribution with known degrees of freedom ν.

In the presence of a large number of DE genes (*i.e.* outliers) L$_2$E or WL$_2$E may result in inflated estimates of the variance and of w. See Figures 1(a) and (b) as an example. It could also happen that EE genes are regarded as outliers and hence that the variance and w are under-estimated. Fixing f_θ protects against both these possibilities.

In this section we derive simple closed-form expressions to estimate w via WL$_2$E when fixing $f_{\boldsymbol{\theta}}$, including the important cases of the multivariate normal and multivariate t distributions. We also obtained estimators via non-weighted L$_2$E, but they seemed to be slightly outperformed by their WL$_2$E counterparts, so we do not describe them here. First consider the normal case. Fixing $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$ and $\hat{\Sigma} = \Sigma$ in (2.6), we have a quadratic function in w with a minimum (the second derivative is positive) at

$$\hat{w} = 3^{p/2}\frac{1}{n}\sum_{i=1}^{n}\exp\{-(\mathbf{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\}. \quad (2.7)$$

That is, $\hat{w}$ is equal to the average squared normal density function multiplied by a constant that accounts for the dimensionality of the problem. In our example of Figure 1(b), (2.7) gives $\hat{w} = 0.62$, which is much closer to the true value of $w = 0.6$ than the estimate $\hat{w} = 0.76$ obtained in Section 2.2.

Now consider the case in which $f_{\boldsymbol{\theta}}$ is assumed to be a multivariate t with location $\boldsymbol{\mu}$, scale Σ and known degrees of freedom ν. Simple integration allows one to obtain an expression analogous to (2.6), *i.e.* quadratic in w with positive second derivative. Taking the derivative with respect to w and setting equal to zero gives the minimum:

$$\hat{w} = \frac{\Gamma(\nu/2)}{\Gamma(\frac{\nu+p}{2})} \frac{\Gamma\left(\frac{3\nu+3p}{2}\right)}{\Gamma\left(\frac{3\nu+2p}{2}\right)} \frac{1}{n} \sum_{i=1}^{n} \left(1 + \frac{1}{\nu}(\mathbf{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right)^{-(\nu+p)}. \tag{2.8}$$

In Sections 4 and 5 below we apply this result with $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = \mathrm{I}$. Again, $\hat{w}$ is equal to the average squared null density times a constant adjusting for the dimensionality of the problem. In fact, it is straightforward to show that for any choice of null distribution, $f_{\boldsymbol{\theta}}$, the estimator for w takes the form

$$\hat{w} = \frac{\mathrm{E}_{\hat{f}}(f_{\boldsymbol{\theta}}(\mathbf{X})^q)}{\mathrm{E}_{f_{\boldsymbol{\theta}}}(f_{\boldsymbol{\theta}}(\mathbf{X})^q)}, \tag{2.9}$$

where q is a positive integer, $\mathrm{E}_g(h(\mathbf{X}))$ denotes the expectation of $h(\mathbf{X})$ with $\mathbf{X}$ distributed according to the density g, and $\hat{f}$ is the empirical distribution of $\mathbf{X}$. Using the L_2E criterion corresponds to $q = 1$, whereas the WL_2E criterion used in (2.7) and (2.8) corresponds to $q = 2$. Having closed-form and interpretable expressions to estimate the proportion of equally expressed genes is appealing and useful (see, e.g., Pounds and Morris (2003) or Langaas *et al.* (2005)). Knowledge about w is important as most frequentist and empirical Bayes procedures use an estimate of w to control the FDR at a desired level, whereas Bayesian procedures find posterior probabilities of differential expression by marginalizing with respect to the posterior distribution of w (Section 3).

3 DIFFERENTIAL EXPRESSION ANALYSIS

The goal of differential expression analysis is to detect as many DE genes as possible while controlling the number of false positives. We adopt the false discovery rate (FDR) as a measure of false-positives. The FDR is defined in a frequentist sense as the expected proportion of genes labeled as DE that are actually EE, setting FDR=0 when no genes are called as DE. The expectation is defined with respect to repeated sampling of the data. The Bayesian FDR is also defined as the expected value of this same proportion, but the expectation is taken with respect to the posterior distribution of the parameters in the model (Müller *et al.*, 2007). Algorithm 1 details the use of partial mixture estimation for differential expression analysis.

Algorithm 1. Partial mixture estimation for differential expression analysis

1. Compute a test statistic x_i for all genes $i = 1, \ldots, n$, *e.g.*, difference between 2 group means or moderated t-test statistics (Smyth, 2004).
2. Fit a partial mixture by WL$_2$E to obtain $\hat{w}$ and $f_{\hat{\boldsymbol{\theta}}}(\mathrm{x})$. Alternatively, treat f_θ as known and estimate only w as in Section 2.3.
3. Classify each gene as DE or EE at some specified FDR level.

There are several variants of the algorithm, depending on the choices made at each step.

Remark 1: In step 1 the approach can work with any number of sensibly chosen test statistics. However, it's very important that for EE genes the test statistic (approximately) follow the parametric form $f_{\boldsymbol{\theta}}$. As already mentioned, independence is not assumed.

Remark 2: In step 2, estimating $\boldsymbol{\theta}$ can be more flexible than treating it as fixed, but it can make the procedure less resistant to outliers (see results in Sections 4 and 5).

Remark 3: Step 3 offers a choice between Bayesian and frequentist approaches. An empirical Bayes approach requires estimating the overall density f. We have found kernel density estimators to perform well, as long as the tails of $f_{\boldsymbol{\theta}}$ are not too thick. If $f_{\boldsymbol{\theta}}$ has thick tails, then genes with extreme test statistic values will have a large ratio, $f_{\hat{\boldsymbol{\theta}}}(\mathrm{x}_i)/\hat{f}(\mathrm{x}_i)$, so that their probability of DE will be small. In our application when $f_{\boldsymbol{\theta}}$ is normal we use the usual normal kernels as implemented in the R function `density` (default bandwidth); when $f_{\boldsymbol{\theta}}$ is a Student's t we use a Cauchy kernel as implemented in the R function `akj` from the `quantreg` library version 4.10 (Koenker (2007)). Under the Bayesian FDR genes are declared as DE as explained in Section 3.1. If frequentist control of the FDR is desired we compute raw p-values for the observed x_i using $f_{\hat{\boldsymbol{\theta}}}$ as the null distribution. Genes are declared as DE based on p-values adjusted by some method for controlling the FDR.

In the remainder of this section we elaborate on both Bayesian and frequentist FDR determination.

3.1 BAYESIAN CONTROL OF THE FDR

To estimate the Bayes FDR we use an empirical Bayes approach. Let $v_i = 1 - w\frac{f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{f(\mathbf{x}_i)}$ be the posterior probability that gene i is differentially expressed conditional on w, $\boldsymbol{\theta}$ and f (and the data). Both L$_2$E and WL$_2$E partial mixture fits provide estimates for w and $\boldsymbol{\theta}$, while f is typically easy to estimate from the observed test statistics using standard methods like kernel density estimation or a mixture of normals. Plugging in these estimates provides the pseudo-posterior probabilities $\hat{v}_i$. We use the term *pseudo* to emphasize the difference with a fully Bayesian approach, which would compute v_i by averaging with respect to the posterior distribution of $(w, \boldsymbol{\theta}, f)$.

Let d_i be an indicator for declaring gene i as DE: $d_i = 1$ for DE and $d_i = 0$ for EE. We compute the Bayesian FDR (Genovese and Wasserman (2002)), denoted $\widetilde{\text{FDR}}$ as:

$$\widetilde{\text{FDR}} = \frac{\sum_{i=1}^{n} d_i(1-\hat{v}_i)}{\sum_{i=1}^{n} d_i}. \tag{3.1}$$

The denominator in (3.1) is just the number of genes declared DE, and the numerator is the posterior expected number of false positives. Efron *et al.* (2001) proposed declaring DE those genes with $\hat{v}_i$ greater than a certain threshold, $d_i = I(\hat{v}_i > t)$, but they leave the choice of t to the user. Müller *et al.* (2004) studied this problem from a decision theoretic point of view and found that to minimize the Bayesian false negative rate while controling for $\widetilde{\text{FDR}}$ one must choose the smallest t such that $\widetilde{\text{FDR}} \leq \alpha$. This threshold is easy to find in practice since $\widetilde{\text{FDR}}$ is constant between all order statistics $\hat{v}_{(i)}$ and $\hat{v}_{(i+1)}$, and it is valid under any kind of dependence structure.

Of course, this method of controlling the FDR is dependent on the assumed model being true. Alternatively, it is possible to find the optimal threshold non-parametrically by permutations under the null hypothesis. Storey (2007) uses a permutation method to find the optimal threshold for his ODP test statistic. The validity of such a permutation-based approach is questionable when the sample size is very small, which is not uncommon in microarray studies. Since the focus of this paper is to provide an approach suitable when permutation-based procedures are not adequate, we do not pursue this issue farther.

3.2 FREQUENTIST CONTROL OF THE FDR

If the test statistic is 1-dimensional one can compute the raw p-value for gene i as the tail probability

$$p - value = \int_{z>|x_i-\hat{\mu}|} f_{\hat{\boldsymbol{\theta}}}(z)dz.$$

Multivariate $\mathbf{x}_i$ can often be reduced to a 1-dimensional test statistic through a function $g(\mathbf{x}_i)$. In some instances the integral can be computed either in closed form or via numerical approximation. The simulations in Section 4 required evaluating the tails of normal distributions, which was accomplished via the R function `pt`. More generally, the integral can be approximated by resampled values $z_1 = g(\mathbf{x}_1^*), \ldots, z_B = g(\mathbf{x}_B^*)$, where $\mathbf{x}_j^*$ are independent draws from $f_{\hat{\boldsymbol{\theta}}}$, and counting the proportion of z_j that are greater than $\mathbf{x}_i$. B should be large enough to obtain a reliable estimate. With modern computing power it is fairly common to perform at least 10,000 draws.

To control for an overall FDR$\leq \alpha$, in our experiments we employ a modification of the p-value adjustment of Benjamini and Hochberg (1995) (BH). Benjamini and Hochberg (1995) proved that using $\tilde{\alpha} = \alpha/w$ instead of α as the desired FDR controls the FDR below α. We simply plug in $\hat{w}$ for w. Of course, when the proportion of EE genes is large ($w \approx 1$) it becomes equivalent to the original BH criterion. Even though our L$_2$E approaches make no assumptions about the dependency among genes, the FDR method for p-value adjustment does assume either independence or some form of positive association.

4 SIMULATION STUDY

We compare via simulation the performance of our partial mixture estimation approach to that of four popular methods in a two group setup. We control the FDR level at 0.05. To ensure reproducibility of our findings, the R code used for the simulation is available at `http://rosselldavid.googlepages.com`.

4.1 COMPETING METHODS

In general, the WL$_2$E criterion performed better than the L$_2$E criterion, so we restrict attention to the former. We consider three variants of our approach defined by the choices taken in steps 1-3 of Algorithm 1.

(a) WL$_2$E-PP. Let the test statistic x_i be the difference between the two group means, and fit a partial normal mixture ($f_{\boldsymbol{\theta}}$= normal) via WL$_2$E. Obtain a list of DE genes using pseudo-posterior (PP) probabilities (Section 3.1).

(b) WL$_2$E-BH. Same as WL$_2$E, but obtain the list of DE genes adjusting p-values via our modified BH whereby $\tilde{\alpha} = \alpha/\hat{w}$ (Section 3.2).

(c) WL$_2$E-EBayes. Let the test statistic x_i be the moderated t-test statistic (Smyth, 2004) as implemented in the R functions `lmFit` and `eBayes` in the `limma` package version 2.1 (Smyth, 2005). This moderated t-statistic is a hybrid between the classical frequentist t-statistic construction and a Bayesian version. In place of the standard sample variance used in the denominator the moderated t has a posterior variance and there are added degrees of freedom to reflect borrowed information from other genes. Fit a fixed-component partial t mixture ($f_{\boldsymbol{\theta}}$ = standard t with ν degrees of freedom) via WL$_2$E (2.8), and declare genes as DE based on posterior probabilities (Section 3.1). The function `eBayes` from `limma` estimates the degrees of freedom, $\hat{\nu}$, In simulations $\hat{\nu}$ was sometimes estimated as $\hat{\nu} = \infty$, which decreased slightly the quality of our fit. In practice we restricted $\hat{\nu}$ to be ≤ 25.

Choosing the non-standardized difference between group means as the test statistic x_i (WL$_2$E-PP and WL$_2$E-BH) is appropriate in situations where, after normalization, the variability of x_i is roughly constant across all genes $i = 1, \ldots, n$. See Section 5 for an illustration on how to satisfy this assumption via normalization. When this assumption cannot be met it is more sensible to use standardized differences between group means, *e.g.* using moderated t-statistics as in WL$_2$E-EBayes. Smyth (2004) showed that under the null hypothesis of EE with normally distributed data, the moderated t-test statistic follows a Student's t distribution with augmented degrees of freedom. This is the motivation for fitting a partial t component under WL$_2$E-EBayes. The naming WL$_2$E-EBayes arises from the similarities between this variant and related empirical Bayes methods (Efron and Tibshirani, 2002; Smyth, 2004).

The four competing methods are (a) significance analysis of microarrays (SAM, Tusher *et al.* (2001)); (b) empirical Bayes (EBayes, Efron and Tibshirani (2002)), (c) mixture of normals (MixNor, Pan *et al.* (2003)) and (d) a simple two-sample test with BH p-value adjustment (t-test BH).

(a) **SAM.** Based on a standard t-test statistic with a modified denominator in that there is an offset by a constant added to the standard deviation. We used the implementation in the R function `sam` from the library `siggenes`, version 1.13.2 (Schwender, 2007). Raw p-values are computed by repeatedly permuting group labels to obtain the null distribution of the t statistic. The p-values are then used to compute q-values with the function `sam`. The q-value for a given gene is defined as the pFDR (a modified definition of the FDR) that would be expected if that gene were declared DE along with all genes having a more extreme test statistic value (Storey, 2003). Declare DE all genes with a q-value $\leq$0.05.

(b) **EBayes.** Based on a standard t-test statistic with a modified denominator in that there is an offset by a constant added to the standard deviation. We used the implementation in the R function `ebam` from the library `siggenes`. Estimate f_0/f and w based group label permutations (function `ebam`). Compute pseudo-posterior probabilities of DE to control the FDR (Section 3.1).

(c) **MixNor.** Here we use the same t-statistic as with the above EBayes approach. Estimate w using the function `ebam`. Estimate f by fitting a normal mixture with 3 components to the observed test statistics. Parameters are estimated with the function `em` in the R library `Mclust`, version 3.1.2 (Fraley and Raftery, 2003, 2006). To estimate f_0 permute group labels and fit a single normal component. Compute pseudo-posterior probabilities of DE to control the FDR (Section 3.1).

(d) ***t*****test BH.** Compute the Welch t-test statistic with approximately normal data (see Section 4.2) and the Wilcoxon test statistic otherwise. Here we attempt to reproduce what a data analyst might do if they could perfectly assess normality. Compute raw p-values using the asymptotic normal distribution and adjust them via regular BH.

Three of the four alternatives, SAM, EBayes and MixNor, use permutations to compute raw p-values, whereas t-test BH assumes asymptotic normality for all genes. In contrast, our (three) partial mixture approaches make distributional assumptions only for the EE genes, which in many cases tend to follow a normal distribution.

4.2 DATA SIMULATION

The simulation focuses on the comparison of n=5000 genes between 2 groups when we have m=3, 5 or 10 microarrays per group. The equality of group sample sizes is for convenience; our approach is applicable in non-balanced situations, as well. We defined two groups as follows. Group 1 was composed entirely of EE genes with log-ratios following a standard normal distribution centered at zero with standard deviation $\sigma = 1$. Group 2 included a mixture of equally-, over-, and under-expressed genes according to specified proportions. The differential expression analysis should detect the over- and under-expressed genes defined in Group 2. To generate the data we consider three possibilities for the mixture weights (EE, OE, UE) in Group 2: (0.8,0.1,0.1), (0.95,0.025,0.025) and (0.95,0.04,0.01). Second, we generate log-ratios for Group 2 according to one of the four scenarios described in Table 1. In the normal-normal scenario all expression values arise from a normal distribution: EE genes $\sim$ N(0,1), OE $\sim$ N(2,1), UE $\sim$ N(-2,1). In the normal-uniform scenario we generate Group 2 either from a normal or a uniform distribution: EE genes $\sim N(0,1)$, OE $\sim U(0,5)$, UE $\sim U(-5,0)$. We chose the uniform

	EE	over-expr.	under-expr.
normal-normal	$N(0,1)$	$N(2,1)$	$N(-2,1)$
normal-uniform	$N(0,1)$	$U(0,5)$	$U(-5,0)$
uniform-uniform	$U(-1,1)$	$U(0,4)$	$U(-4,0)$
t-t	$t_5(0,1)$	$t_5(2,1)$	$t_5(-2,1)$

Table 1: Simulation scenarios. Distribution used to generate expression values.

distribution because it represents a strong departure from normality. Other authors have also used the uniform distribution in their modeling choices (e.g., Parmigiani *et al.* (2002)). The last two scenarios do not satisfy the L_2E parametric assumption of normality for f_θ. In the uniform-uniform case the EE genes arise from a uniform distribution and in the t-t case from t-distributions with 5 degrees of freedom. We chose a t with low degrees of freedom to represent a situation in which there may be appreciable overlap among the three components of expression.

To focus on distributional assumptions, mixing proportions, and component separations we consider the independent case only. The 3 choices of mixing proportions and the 4 distributional scenarios give rise to 12 distinct simulation configurations. We also analyzed some additional scenarios in which the three expression components had substantial overlap [e.g., EE $\sim$ N(0,1), OE $\sim$ N(1,1), UE $\sim$ N(-1,1)] to explore the signal-to-noise limits. Not surprisingly, in these circumstances all methods performed (data not shown) very poorly.

Once the expression data were generated, we used all the methods described in Section 4.1 to obtain a list of DE genes. The FDR was estimated by repeatedly generating data and computing the average proportion of genes in this list that were false positives. The power was estimated as the average proportion of DE genes that were in the list. The number of repetitions was large enough to ensure that the width of the 95% confidence intervals for the power and FDR estimates was ≤ 0.01. In most scenarios 100 repetitions were sufficient.

4.3 RESULTS

We now present the findings of the simulation. Figure 2 provides the estimated power for each of the considered methods. Power is computed as the percentage of DE genes that were indeed declared to be DE. Figure 3 reports the corresponding estimated FDR. To reproduce the exact numerical results, see the R code available at `http://rosselldavid.googlepages.com`.

In general, we observed that the WL_2E-PP and WL_2E-BH variants of our partial mixture algorithm perform very similarly. These two WL_2E approaches were the

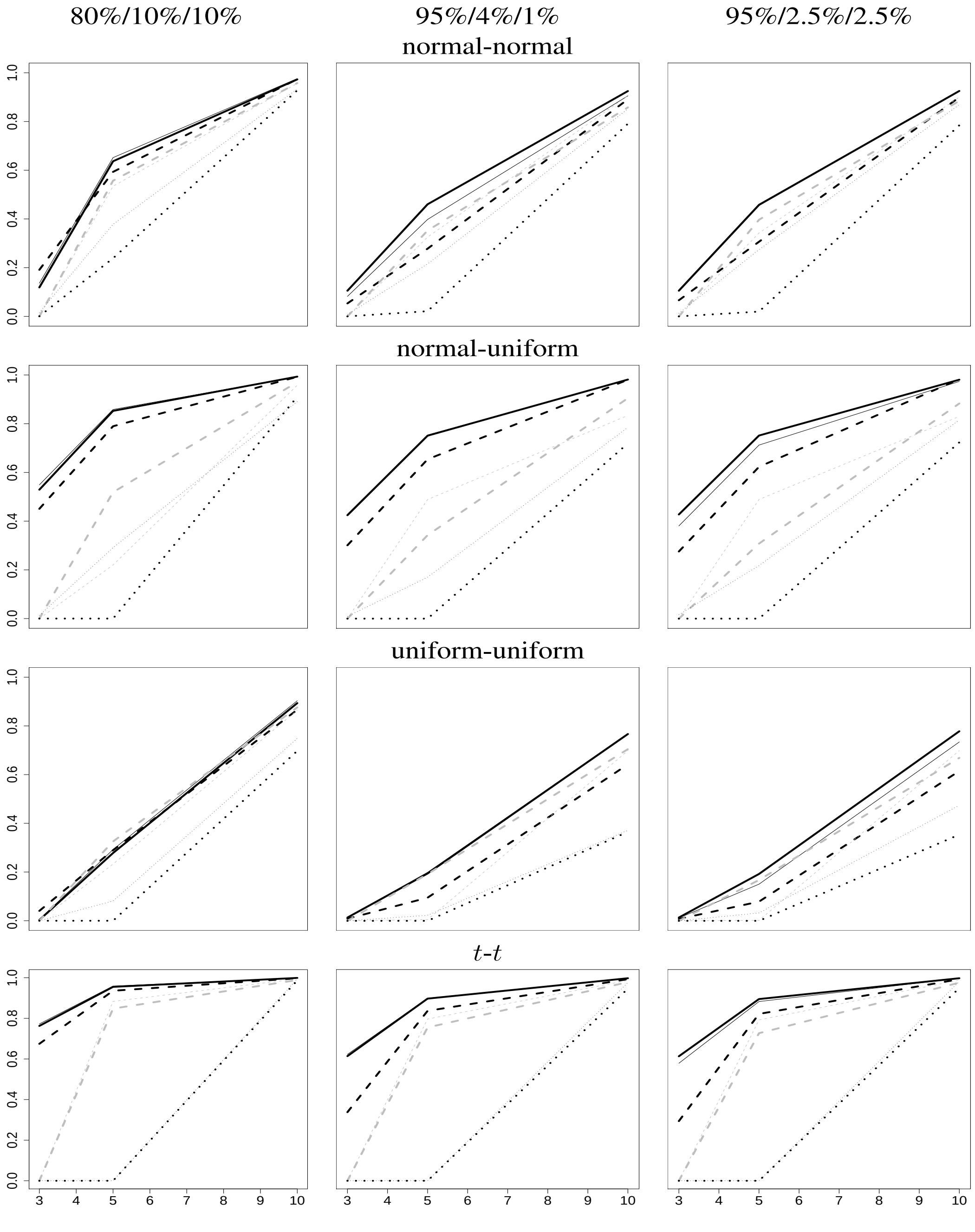

Figure 2: Power for simulation study (y-axis) at 3, 5 and 10 observations per group (x-axis). Lines interpolate the power at 3, 5 and 10 observations. Thin solid: WL$_2$E-PP. Thick solid: WL$_2$E-BH. Black dashed: WL$_2$E-EBayes. Thin dashed gray: SAM. Thick dashed gray: MixNor. Thin dotted: EBayes. Thick dotted: t-test BH. All methods attempting to control FDR below 0.05.

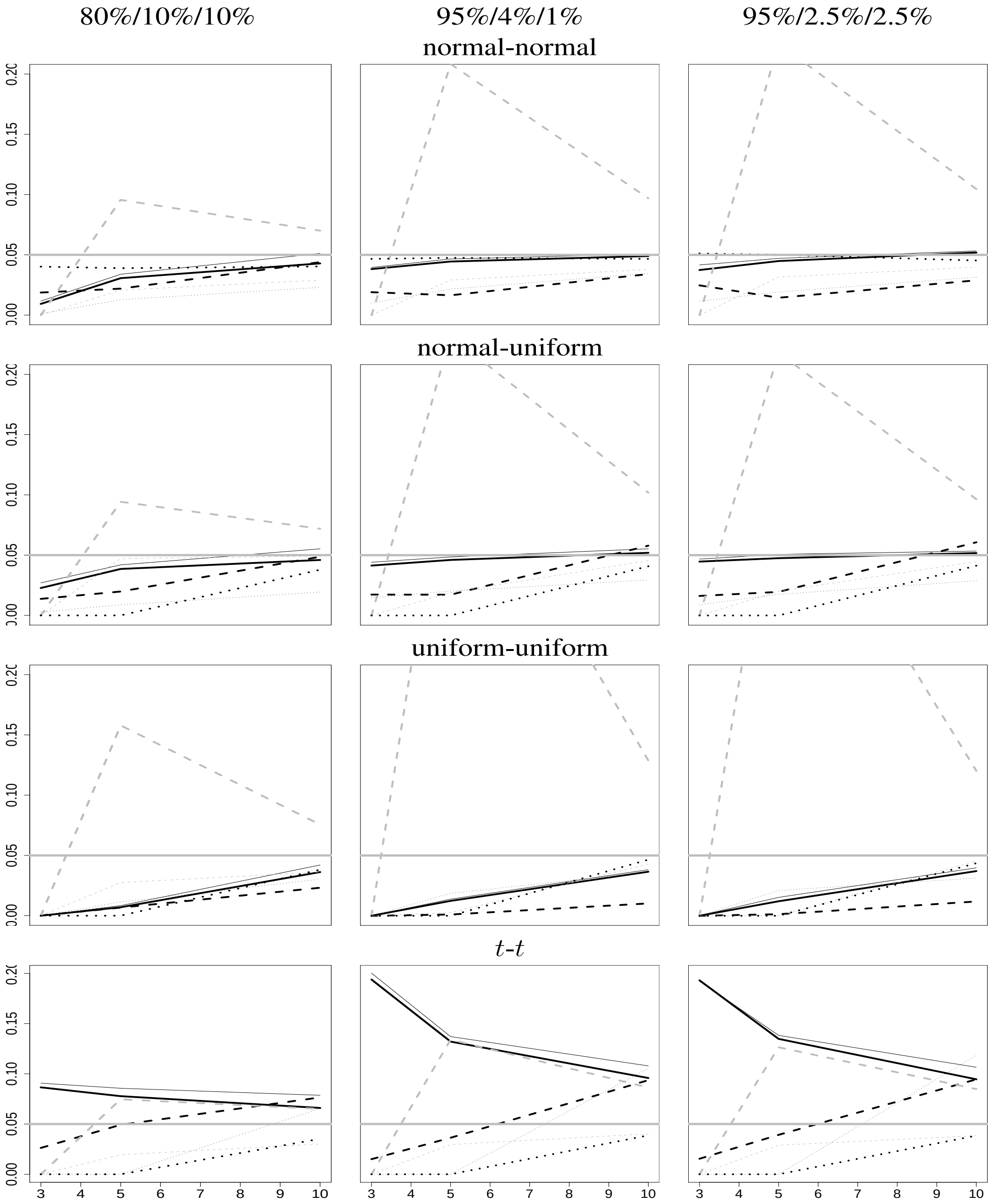

Figure 3: FDR for simulation study (y-axis) at 3, 5 and 10 observations per group (x-axis). Lines interpolate the FDR at 3, 5 and 10 observations. Thin solid: WL_2E-PP. Thick solid: WL_2E-BH. Black dashed: WL_2E-EBayes. Thin dashed gray: SAM. Thick dashed gray: MixNor. Thin dotted: EBayes. Thick dotted: t-test BH. All methods attempting to control FDR below 0.05 (thick horizontal line indicates 0.05).

most powerful under almost all conditions in the normal-normal, normal-uniform and uniform-uniform cases, with very significant advantages being observed for group sample sizes of 3 and 5. The fixed-component WL$_2$E-EBayes also performed quite well. When the sample size was 3 the power of all competitors was virtually zero in all scenarios, whereas WL$_2$E achieved a power between 12%-19% in the normal-normal and 45%-55% in the normal-uniform scenario. In all these situations WL$_2$E controlled the FDR below the desired 5% level, including the uniform-uniform where the normality assumption is violated. The one exception occurred in the normal-uniform case where the FDR level sometimes reached 6%.

In the t-t scenario WL$_2$E-BH and WL$_2$E-PP again were the most powerful but they presented an FDR well above 5%, especially for smaller sample sizes, while WL$_2$E-EBayes had better FDR levels but still above 5% in some situations. In this scenario the competing methods often failed to control the FDR as well. A possible explanation is that the heavier tails of the t-distribution generate observations that are regarded as outliers by a partial normal component, and are therefore tagged as arising from DE genes, whereas the fixed-component estimate is more resistant to these outliers. We conducted additional simulations with $m = 100$ and found the FDR below 5%. At this sample size the t test statistic likely follows a standard normal distribution and the effect of outliers is minimized, corroborating our explanation.

SAM was the best among the competitors, exceeding their power while controlling the FDR below the desired level in most scenarios. MixNor had good power but its FDR was often well above 5%. EBayes and t-test BH were the only two methods that presented an FDR below 5% in all scenarios, although the loss in power was sometimes substantial. For instance, both methods had virtually zero power to detect any genes for a sample size of 3 observations per group. EBayes tended to outperform t-test BH in terms of power, especially in the normal-normal and normal-uniform scenarios.

Power increased sharply with sample size. In the case of our three partial mixture methods this could be due to the clearer separation of the mixing components in (1.1). For SAM, EBayes and MixNor, another possible explanation. The estimation of the null distribution based on permutations is not accurate when $m = 3$, which likely results in a loss of power.

5 CASE STUDIES

We analyzed real data from an apolipoprotein-AI (apoAI) experiment presented by Callow *et al.* (2000) and from a leukemia study of Golub *et al.* (1999). The methods we used to assess differential expression were the three variants of WL$_2$E, SAM, EBayes, MixNor and t-test BH, all set to control the FDR at 0.05. Dudoit *et al.* (2003) analyzed the apoAI dataset with several methods and found eight DE genes out of 6384. The original analysis of Golub's data was based on a neighbourhood-based method that found 1000 DE genes out of 6817. The apoAI study represents the case in which few DE genes are expected, whereas the true proportion of DE genes in the leukemia study is probably relatively large.

5.1 APOLIPOPROTEIN EXPERIMENT

5.1.1 DESCRIPTION

The apoAI experiment concerned lipid metabolism and atherosclerosis susceptibility in mice. The experiment compared gene expression between 8 apoAI knock-out mice and 8 inbred control mice. cDNA was obtained from mRNA by reverse transcription and hybridized to 6384 probes on a glass microarray. Pooled cDNA from the 8 control mice was used as a common reference sample for all hybridizations (knock-out and control). We obtained the data from `http://www.stat.berkeley.edu/users/terry/zarray/Data/ApoA1/rg_a1ko_morph.txt`.

For two groups of 8 observations (microarrays) each there are 12,870 possible permutations, which should be sufficiently large to accurately estimate the null distribution, f_0. Both SAM and EBayes construct permutation based null distributions.

5.1.2 NORMALIZATION AND MODEL CHECKING

We normalized the gene expression intensities in two steps. First, we corrected for chip printing effects using the `maNormNN` method as implemented in the `nnNorm` package for the R software (see package documentation for details). For WL$_2$E-PP and WL$_2$E-BH we use the simple difference between group means as a test statistic, x_i. We do not divide by its estimated standard error as is done in standard t-test as it is well known that the variance of log-ratios is a function of their magnitude. What is needed is a gene-specific variance estimate. Therefore, we perform a second step to ensure that the test statistic is identically distributed across genes as required by the L$_2$E approach. Denote the sum of the two group means for the i^{th} gene as A_i. Following the MA-plot idea (Dudoit *et al.*, 2002a) we obtain a `lowess` local least squares fit of x_i on A_i and calculate the residuals e_i, $i = 1 \ldots n$. We then regress the squared residuals e_i^2 on A_i via `lowess`. The fitted values at A_i give gene-specific estimates of the residual variance. The variance-stabilized test statistic is obtained by dividing e_i by the square root of its estimated variance (the fitted value). In both `lowess` fits the smoothing parameter is chosen by minimizing the mean absolute

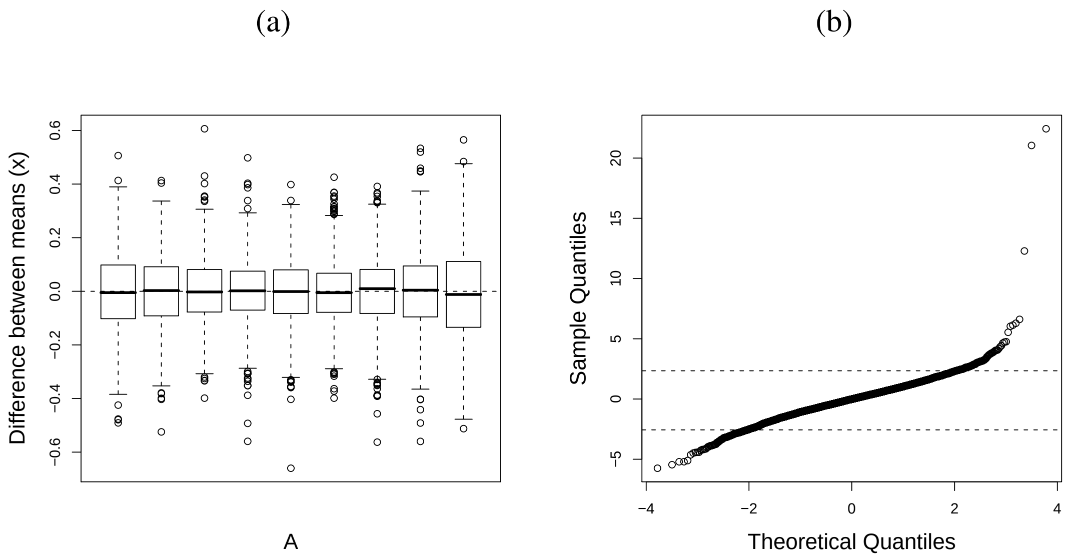

Figure 4: Assessing partial mixture assumptions for apo AI dataset. (a): mean intensity values (A) are categorized in 10 groups according to the observed quantiles. (b): normal quantile plot of the test statistic. Horizontal lines contain a proportion $\hat{w}$ of the data.

error by cross-validation.

Figure 4(a) displays the distribution of x_i for different values of A_i after the variance stabilization procedure. The mean and variance of x_i are roughly constant for all A_i values, suggesting that the partial mixture assumption of the test statistic being identically distributed is not unreasonable.

The WL$_2$E fit estimates are $\hat{\mu} = 0$, $\hat{\sigma} = 0.13$ and $\hat{w} = 0.96$. That is, it indicates that 96% of the test statistic values arise from a normal distribution and the remaining 4% are considered anomalies. The normality of the test statistic is assessed in Figure 4(b), which presents a qq-normal plot using the weighted L$_2$E estimates. The horizontal lines contain $\hat{w}$ =96$\%$ of the test statistic values; observations outside the region delimited by the lines are not considered to arise from equally expressed genes. The normality assumption is plausible since departure from normality is observed mainly in the tails.

We also computed the moderated t-test statistics for the WL$_2$E-EBayes variant of our algorithm (Section 4.1). The augmented degrees of freedom were estimated as 18, which is slightly higher than the 14 degrees of freedom that the classical t-test statistic would have. Equation (2.8) provides the estimate $\hat{w} = 0.95$. To assess the distributional assumptions of the moderated t-test statistic we produced plots

	WL_2E PP	WL_2E BH	WL_2E EBayes	EBayes	SAM	MixNor	t-test-BH
KO over-expr.	35	32	1	0	0	6	0
KO under-expr.	25	27	9	0	0	10	8

Table 2: Gene classification for apo AI dataset. The table describes the number of genes declared to be over and under-expressed in knock-out mice out of the 6384 genes. PP indicates the use of posterior probabilities; BH is Benjamini-Hochberg p-value adjustment

analogous to those in Figure 4 and found the assumptions acceptable.

5.1.3 RESULTS

Results of the differential expression analysis are shown in Table 2. WL_2E-PP and WL_2E-BH declared a substantially larger number of genes to be DE than the other methods. WL_2E-EBayes finds 10 DE genes, the t-test with Benjamini & Hochberg's (BH) p-value adjustment detects 8 genes to be under-expressed in knock-out mice, whereas both SAM and EBayes did not detect any (for SAM the complete set of 12870 permutations under the null were used). MixNor, the only competing method finding some genes to be over-expressed in the knock-out mice, detects 16 DE genes. The t-test BH procedure coincided with the findings of Dudoit *et al.* (2003), who claimed significance for the 8 genes with the most extreme values the two-sample t-test statistic. These 8 genes were also found by all our partial mixture approaches.

We now assess the performance of all approaches when analyzing a subset of the data. We randomly select samples 2, 3, 5, 7 and 8 from the KO group and samples 1, 3, 4, 7 and 8 from the control group. Producing a plot analogous to Figure 4(b) revealed a stronger departure from normality than that observed for the full dataset. WL_2E-PP and WL_2E-BH found 100 and 95 genes, respectively, *i.e.* more than for the full dataset, and only about 31% of these genes were found again when analyzing the full dataset. WL_2E-EBayes found 10 genes, 8 of which were confirmed with the full data, and t-test BH found 2, none of which were confirmed with the full data. SAM and EBayes did not declare any genes to be DE. MixNor found 12 DE genes, 33% of which were not found again in the full dataset. These findings suggest that WL_2E-PP, WL_2E-BH and t-test BH can lead to an inflated FDR when the normality assumption is violated, but that WL_2E-EBayes is more robust.

5.2 LEUKEMIA STUDY

5.2.1 DESCRIPTION

Golub *et al.* (1999) compared gene expression levels between acute lymphoblastic leukemia (ALL) cells and acute myeloid leukemia (AML). We used the version of the dataset posted with the original publication at `http://www.broad.mit.edu/cgi-bin/publications/display_pubs.cgi?id=201`. The study used Affymetrix HuGeneFL arrays that measured mRNA expression for 7129 genes, and had 27 ALL and 11 AML samples. The original dataset also contains a variable indicating, for each array, which genes had enough mRNA to be considered to be present. In our analysis we only included the 4763 genes that were present in at least 1 microarray.

5.2.2 NORMALIZATION AND MODEL CHECKING

The data obtained were already normalized per Golub *et al.* (1999). A qq-normal plot revealed serious departure from normality.

The partial mixture assumptions are considered in Figure 5. Panel (a) reveals that the test statistic median decreases slightly as the average intensity (A_i) increases. The spread tends to increase with the average intensity. Without any additional normalization a WL$_2$E fit gives $\hat{\mu} = 0.24$, $\hat{\sigma} = 1.79$ and $\hat{w} = 0.99$, indicating only 1% of the genes are differentially expressed. This result is dubious since 28% of the genes have an observed test statistic exceeding 2 in absolute value, suggesting that the proportion of DE genes is higher than 1%. A SAM analysis estimated $\hat{w} = 0.53$. The very small value of $\hat{w} = 0.99$ by WL$_2$E-EBayes is likely due to a lack of separation between the components in (1.1). Therefore, we used the fixed-component WL$_2$E, which is more robust to overlap since f_θ is specified by theoretical considerations and not estimated by the data. To this end, f_θ is fixed to a t distribution with augmented degrees of freedom estimated as 37. Applying (2.8) we obtain $\hat{w} = 0.63$, which is more in line with SAM and the distribution of t values. The Student's t qq-plot in panel (b) suggests that the t assumption is not unreasonable for 63% of the test statistic values closest to the mean.

5.2.3 RESULTS

WL$_2$E-EBayes called 744 genes as DE, whereas EBayes declared 881, SAM 662, MixNor 1303 and the Wilcoxon test with BH p-value adjustment 610. In terms of concordance between methods, WL$_2$E-EBayes classified 94% of the genes in the same category (equally, over or under-expressed) as did EBayes did. For SAM this percentages was 87%; MixNor 96%; the Wilcoxon test 92%.

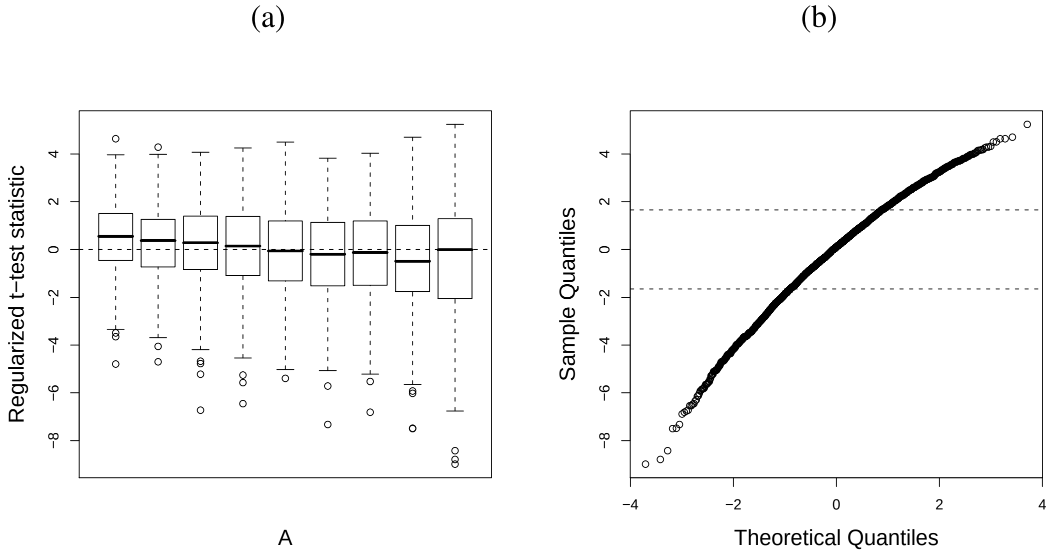

Figure 5: Assessing partial mixture assumptions for Leukemia dataset. (a): mean intensity values (A) are categorized in 10 groups according to the observed quantiles. (b): horizontal lines contain a proportion $\hat{w} = 0.63$ of the data.

Sample Size	WL$_2$E-EBayes	EBayes	SAM	MixNor	Wilcoxon-BH
27 ALL / 11 AML	744	881	662	1303	610
5 ALL / 5 AML	11	0	0	37	0

Table 3: Gene classification by method and sample size for leukemia data set. Entry is number of genes declared to be differentially expressed among 4763 genes.

To assess the performance of the methods with a smaller sample size we randomly selected samples 2, 6, 10, 19 and 27 from the ALL group and samples 3, 6, 7, 9 and 11 from the AML group, and we repeated all the analyses. Table 3 gives the number of DE genes found by the different methods. All methods detect hundreds of genes using the full sample sizes. However, at the reduced sample sizes all methods except WL$_2$E-EBayes and MixNor fail to detect any DE genes.

6 DISCUSSION

We have proposed the use of partial mixture estimation as a semi-parametric approach to differential expression analysis. This framework requires a parametric model for equally expressed genes only and can handle small group sizes. Other methods require a full probability model or larger group sizes for permutation-based null sampling. We have developed an iterative weighted $\mathrm{L_2E}$ criterion that improves upon the performance of the $\mathrm{L_2E}$ criterion originally proposed by Scott (2004a,b). We have also found that fixing some of the parameters in the EE mixture model component can further improve the fit, making it more resistant to outliers and to the extent to which the mixture components are well-separated. The fixed-component approach allows simple closed-form expressions to estimate the proportion of EE genes, which is used by several frequentist and empirical Bayes procedures to control the FDR. The $\mathrm{L_2E}$ methods are computationally efficient; in our experience only a few seconds are required to run an R-implemented program.

Our approach requires making only two assumptions. First, we assume that the test statistic observations used to classify genes are identically distributed realizations from a common distribution. The method does not assume independence. Second, we require that the distribution of equally expressed genes has a known parametric form. The $\mathrm{L_2E}$ and $\mathrm{WL_2E}$ criteria make an implicit third assumption that most genes are EE, even though in simulations we found that this can typically be overcome by carefully setting the initial parameter estimates. By construction, the fixed-component $\mathrm{WL_2E}$ is robust to this assumption. We have illustrated by example how variance stabilizing normalization can render identically distributed observations, and we have demonstrated accompanying techniques for visual validations of both assumptions. The choice of test statistic remains important, since it has an effect on the final results. For example, we expect test statistics that borrow information across genes to perform better than those that are computed separately for each gene.

Simulation studies and real data analyses have been used to compare our approach with EBayes, SAM, MixNor and the t-test with p-value adjustment. The results suggest that partial mixture estimation can provide significant advantages over the other approaches, especially when the sample size is small and most genes are equally expressed. This is an important feature since under high EE mixing probabilities it can be hard to control the FDR and some authors prefer to control the family-wise error rate instead. Our method appears to control the FDR at desired levels, given that the assumptions hold.

Based on our work we find fixed-component $\mathrm{WL_2E}$ to be more useful than regular $\mathrm{WL_2E}$. The former can be less aggressive in identifying observations as outliers (*i.e.* DE genes), which can result in a better FDR control, and it also appears more robust to the underlying parametric assumption of the EE mixture component. In fact, we expect inflated FDR rates to be more common in scenarios where f_0 has thicker tails than the assumed normal or t and the sample size is small. Tail thickness can be assessed graphically via qq or density plots when a small proportion of

DE genes is expected. As sample size grows permutation methods can be used to obtain a sample under the f_0, although many test statistics exhibit an improved normality for larger sample sizes and hence inflated FDR rates should a lesser concern.

In general, SAM performed best among the competitors, particularly in the simulation studies. This supports the findings of Schwender *et al.* (2003), who found that SAM performed better than EBayes in simulations but the latter gave more significant hits in a real data set. MixNor achieved good power but it controlled the FDR poorly, especially when the proportion of DE genes was small. This suggests that, as MixNor models both EE and DE genes, it performs better as the proportions of data arising from both components are more balanced. Our partial-mixture approaches avoid this issue by modeling only f_0.

Although partial mixture estimation does not define a full probability model, we are able to provide some summaries with connections to more formal model-based approaches such as pseudo-posterior probabilities of differential expression. The lack of a full probability model is tempered by the fact that differential expression analysis is most commonly used for data exploration and hypothesis generation and less for definitive inference.

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

M.J. Callow, S. Dudoit, E.L. Gong, T.P. Speed, and E.M. Rubin. Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. *Genome research*, 10:2022–2029, 2000.

K. Do, P. Müller, and F. Tang. A bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society C*, 54:627–664, 2005.

S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the Americal Statistical Association*, 97:77–87, 2002.

S. Dudoit, H.Y. Yang, M.J. Callow, and T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:972–977, 2002.

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

B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23:70–86, 2002.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the Americal Statistical Association*, 96:1151–1160, 2001.

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

C. Fraley and A.E. Raftery. Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *Journal of Classification*, 20:263–286, 2003.

C. Fraley and A.E. Raftery. *MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering*. Department of Statistics, 2006.

C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society B*, 64:499–518, 2002.

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

F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel. *Robust Statistics - The Approach Based on Influence Functions*. Wiley, New York, 1986.

P. Huber. *Robust Statistics*. Wiley, New York, 1981.

C.M. Kendziorski, M.A. Newton, H. Lan, and M.N. Gould. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22:3899–3914, 2003.

Roger Koenker. *quantreg: Quantile Regression*, 2007. R package version 4.10.

M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society B*, 67:555–572, 2005.

P. Müller, G. Parmigiani, C. Robert, and J. Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99:990–1001, 2004.

P. Müller, G. Parmigiani, and K. Rice. *FDR and Bayesian Multiple Comparisons Rules*. Oxford University Press, 2007.

M.A. Newton and C.M. Kendziorski. *Parametric Empirical Bayes Methods for Microarrays*. Springer Verlag, New York, 2003.

M.A. Newton, C.M. Kendziorski, C.S Richmond, F.R. Blattner, and K.W. Tsui. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52, 2001.

M.A. Newton, A. Noueriry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics*, 5:155–176, 2004.

W. Pan, J. Lin, and C.T. Le. *A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data*, pages 117–124. Springer-Verlag GmbH, 2003.

G. Parmigiani, E.S. Garret, R. Anbazhagan, and E. Gabrielson. A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society B*, 64:717–736, 2002.

S. Pounds and S.W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 10:1236–1242, 2003.

H. Schwender, A. Krause, and K. Ickstadt. Comparison of the empirical Bayes and the significance analysis of microarrays. Technical Report 44, 2003.

Holger Schwender. *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches*, 2007. R package version 1.13.2.

D.W. Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43:274–285, 2001.

D. W. Scott. Partial mixture estimation and outlier detection in data and regression. In M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, editors, *Theory and Applications of Recent Robust Methods*, Statistics for Industry and Technology, pages 297–306. Birkhäuser, Basel, Switzerland, 2004.

D.W. Scott. Oulier detection and clustering by partial mixture modelling. Proceedings of COMPTSTAT, Ed. Antoch., 2004.

A. Scott. *Denoising by Wavelet Thresholding Using Multivariate Minimum Distance Partial Density Estimation*. PhD thesis, Rice University, 2006.

G.K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

G.K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

J.D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, 31:2013–2035, 2003.

J.D. Storey. The optimal discovery procedure: A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society B*, 69:347–368, 2007.

V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Science*, 98:5116–5121, 2001.

M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.