# Semi-parametric Differential Expression Analysis via Partial Mixture Estimation

DAVID ROSSELL
Department of Biostatistics
M.D. Anderson Cancer Center, Houston, TX 77030, USA
rosselldavid@gmail.com
RUDY GUERRA
Department of Statistics
Rice University, Houston, TX 77005, USA
CLAYTON SCOTT
Department of Electrical Engineering & Computer Science
and Center for Computational Medicine and Biology
University of Michigan, Ann Arbor, MI 48109, USA

September 7, 2007

**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] and a t-test with p-value adjustment to control the FDR [Benjamini and Hochberg, 1995]. Gene expression analysis; Microarray; Partial Mixture estimation.

# 1 INTRODUCTION

The availability of new biotechnologies such as microarrays has made possible the collection of large amounts of biological data, and brought forth the challenge of developing statistical methods to properly analyze it. One scenario that has attracted considerable interest is gene differential expression analysis, *i.e.*, the comparison of some measure of gene expression between groups defined, for instance, by treatments or biological conditions. For example, the apo-AI experiment discussed in Section 5 compares expression levels between 8 mice with the apo-AI gene knocked out and 8 inbred control mice. The question of biological interest is which genes are differentially expressed (DE) between these two groups and which are equally expressed (EE). The goal is to detect as many DE genes as possible while not having too many false positives.

More formally, suppose that expression levels for $n$ genes are measured and normalized to account for systematic biases (see Dudoit *et al.* [2002b]). Furthermore suppose that in order to discriminate EE and DE genes we compute a test statistic $\mathbf{X}$ for each gene. For example, in the case of Affymetrix oligonucleotide cDNA arrays the measurements can be log red-to-green intensity ratios and the test statistic could be the difference of the mean log intensities between two groups. Other statistics are discussed by Efron *et al.* [2001], Efron and Tibshirani [2002], Tusher *et al.* [2001] and Smyth [2004]. Note that this statistic may be multivariate, *i.e.* take values in some set $S_x \subseteq \Re^p$, as warranted for example by time-course studies.

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 mixture density

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

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 and make probability statements about the genes under consideration. Dudoit *et al.* [2002b] review some approaches based on computing t-tests for each gene and adjusting for multiple comparisons. Tusher *et al.* [2001] introduced the significance analysis of microarrays (SAM), which is based on obtaining p-values through permutations and computing their multiple comparisons equivalent, the q-values. Efron and Tibshirani

[2002] and Efron [2004] proposed a non-parametric empirical Bayes approach and Pan *et al.* [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 *bona fide* 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 albeit it requires estimating some unknown quantities from the data.

All of the methods mentioned above either make full distributional assumptions or rely on resampling methods to sample under $f_0$. In some situations, however, these approaches can be difficult to justify. Models for the DE distribution $f_1$ may be difficult to specify when very few genes are DE, meaning there is very little data available for model fitting or validation. Permutation methods, on the other hand, need at least a certain number of microarrays per group. If we have 2 groups of 3 observations each, then there are only 10 distinct permutations of the microarrays, which provides a coarse representation of the test statistic under the null. 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 can be used even 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].

Our approach requires making only two assumptions. First, we assume some parametric form for $f_0$. Second, we assume that the test statistic is identically distributed for all 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 that DE data. Obtaining identically distributed statistics can be achieved via appropriate data pre-processing or normalization procedures, as we illustrate in real data in Section 5.

In the next section we describe Scott's original $L_2E$ approach to partial mixture estimation and we develop two improved variants which we call *weighted* $L_2E$ ($WL_2E$) and fixed-component $WL_2E$, respectively. Interestingly, the latter variant provides closed-form and interpretable estimates of the proportion of EE genes. In Section 3 we describe the application of these algorithms to differential expression analysis. To adjust for multiple testing, we present two techniques, one frequentist

and one Bayesian, that control the false discovery rate (FDR) at a desired level. In Section 4, we show by analyzing 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 implementing the $L_2E$ and $WL_2E$ methods at *www.stat.rice.edu/~rusi*.

# 2 PARTIAL MIXTURE ESTIMATION

In general, any test statistic that we choose to test for differential expression will be distributed 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, *e.g.*, many statistics are approximately normally distributed as the number of measurements per group increases. Also note that if we expect most of the genes to be EE, we can assess whether the parametric assumption is reasonable by exploring the behavior of the observed statistics. For example, we can assess the normality assumption by obtaining a QQ-normal plot and checking that there is no departure from normality for most of the observed test statistics.

Partial mixture estimation is the problem of estimating $w$ and $f_0$ only, without estimating the remaining components of the mixture. 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% of them corresponding to EE genes that follow a $N(0, 1)$ distribution, 20% under-expressed genes following a $N(-4, 1)$, and 20% over-expressed genes following a $N(4, 1)$. As shown in Figure 1(a) the $L_2E$ fit (explained below) provides a local estimate of the overall distribution around 0, therefore effectively estimating the distribution of the EE genes. The estimate $\hat{w} = 0.68$ indicates that the $L_2E$ fit perceives 32% of the data to be anomalous, *i.e.*, not arising from $f_0$.

We distinguish partial mixture estimation from standard robust estimation [Hampel *et al.*, 1986; Huber, 1981], which seeks to estimate $f_{\boldsymbol{\theta}}$ 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, knowledge of $w$ is useful in controlling FDR both in a frequenstist and Bayesian sense, as discussed in Section 3.

In Section 2.1 we review the original $L_2E$ criterion [Scott, 2004a,b], whereas in Sections 2.2 and 2.3 we develop two new criteria to obtain partial mixture fits.

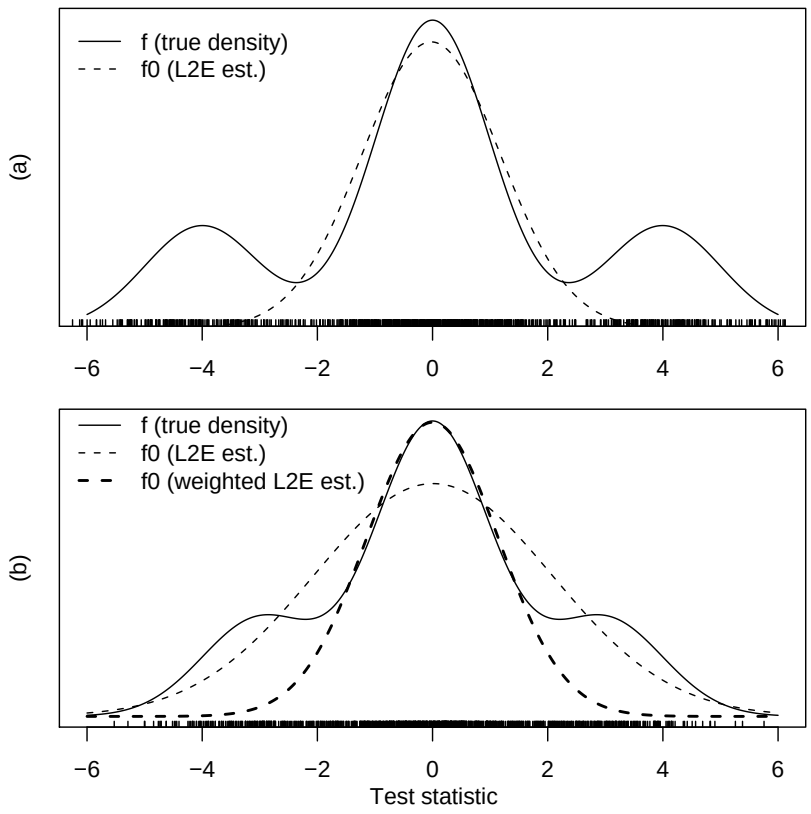


Figure 1: $L_2E$ fit example. (a): $f = .6N(0,1) + .2N(-4,1) + .2N(4,1)$. $L_2E$ estimate: $.68N(0.02, 1.20)$. (b) $f = .6N(0,1) + .2N(-3,1) + .2N(3,1)$. $L_2E$ estimate: $.97N(.03, 1.97)$. $WL_2E$ estimate: $.74N(0.01, 1.25)$. The vertical segments on the x axis indicate the generated test statistic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

## 2.1 $L_2$E CRITERION

To emphasize that as a parametric distribution $f_0$ is indexed by some parameter $\boldsymbol{\theta}$, from now on we will denote it as $f_{\boldsymbol{\theta}}$. Scott [2001] proposed the $L_2$E criterion for parametric density estimation, finding that it is quite efficient asymptotically and that it is robust to departures from the assumed model and to the presence of outliers. Scott [2004a,b] then used the $L_2$E criterion to estimate the local behavior of $f$ with a partial mixture component $wf_{\boldsymbol{\theta}}$. His approach seeks to minimize the integrated squared difference or $L_2$ distance between the true density $f$ and its local approximation $wf_{\boldsymbol{\theta}}$:

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

where $C$ is a constant that does not depend on $w$ or $\boldsymbol{\theta}$ and hence can be ignored for the purpose of minimization. Note that 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 $\boldsymbol{\theta}_0$, the overall minimum in (2.5) is $\hat{w} = 1$, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

By allowing the estimated component to integrate to $w$ instead of 1, this criterion capitalizes on a virtue of minimum-distance methods not shared by likelihood-based and other approaches. The estimate will tend to approximate the largest component of $f$ instead of blurring all components together, to an extent dependent on the separation between the components. The $WL_2$E method developed in Section 2.2 is less dependent on the components being well-separated. In differential expression analysis a common assumption is that most genes are EE, *i.e.* they define the largest component, so $L_2$E should indeed estimate $f_{\boldsymbol{\theta}}$.

The first integral in (2.1) has a closed form for several common distributions, including the multivariate normal and t [Wand and Jones, 1995]. This is convenient for computational speed but not strictly necessary, since numerical approximations can be used. 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 the sample mean computed with respect to the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. 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 observations to be independent. When $f_{\boldsymbol{\theta}}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance $\Sigma$, the criterion simplifies to [Wand and Jones, 1995]

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

To find $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ and $\hat{w}$ minimizing this expression we use the R function `nlmin` for general nonlinear minimization. Note that 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 be the maximum likelihood estimate, although multivariate statistics may require more care.

Finally, note that one is not restricted to assuming normality. For example, one could model $f_{\boldsymbol{\theta}}$ with a general multivariate t distribution. In fact, we explored this possibility but the results were not encouraging, since the heavier tails of the t tended to absorb outliers and hence obscure DE genes. However, we did find the t model useful when holding $\boldsymbol{\theta}$ fixed, as we explain in Section 2.3.

## 2.2 WEIGHTED $L_2$E CRITERION

As seen in the example in 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 $L_2$E can capture the behavior of the EE genes quite well. However, when the separation is not so clear problems can arise, as we will illustrate with another simulated example. 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 it can be seen in Figure 1(b), the $L_2$E estimate tries to approximate the overall density all over its domain, and therefore fails to capture its local behavior around zero. Also, $L_2$E estimates $\hat{w} = 0.97$.

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 approach *weighted $L_2$E for partial mixture estimation* ($WL_2$E).

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 is again approximated by a sample mean. Under the assumption of normality, the $WL_2$E partial

mixture estimate is obtained by minimizing

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

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 $\text{WL}_2\text{E}$ yields $(\hat{\mu}, \hat{\sigma}, \hat{w}) = (0.01, 1.30, 0.76)$, which provides a better local approximation than the initial $\text{L}_2\text{E}$ 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 negligible, *e.g.* smaller than $1\%$ in square norm, and we obtain the final estimate of $(\hat{\mu}, \hat{\sigma}, \hat{w}) = (0.01, 1.25, 0.74)$.

## 2.3 FIXED-COMPONENT $\text{WL}_2\text{E}$

In some situations it is reasonable to assume that $\boldsymbol{\theta}$ is known, and that therefore only $w$ needs to be estimated. Knowledge of $w$ is important in frequentist and Bayesian procedures that adjust for multiple comparisons (see Section 3). 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 $\nu$. For large datasets, a normal component may also be appropriate. Intuitively, this can be advantageous under the presence of a large amount of DE genes, since a full $\text{L}_2\text{E}$ or $\text{WL}_2\text{E}$ may break down and result in inflated estimates of the variance, for instance.

In this section we derive simple closed-form expressions to estimate $w$ via $\text{WL}_2\text{E}$ when fixing $f_{\boldsymbol{\theta}}$, including the important cases of the multivariate normal and multivariate Student's t. We also obtained estimators via non-weighted $\text{L}_2\text{E}$, but they seemed to be slightly outperformed by their $\text{WL}_2\text{E}$ 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 minimum (note that 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 quite better 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 $\Sigma$ and known degrees of freedom $\nu$. 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:

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

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 see that for any choice of null distribution $f_{\boldsymbol{\theta}}$ our estimate takes the form

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

where $\mathrm{E}_g(h(\mathbf{X}))$ denotes the expectation of $h(\mathbf{X})$ when $\mathbf{X}$ is distributed according to the density $g$, and $\hat{f}$ is the empirical distribution of $\mathbf{X}$. Using the $\mathrm{L}_2\mathrm{E}$ criterion corresponds to $q = 1$, whereas the $\mathrm{WL}_2\mathrm{E}$ criterion used in (2.7) and (2.8) corresponds to $q = 2$. We find that having closed-form and interpretable expressions to estimate the proportion of equally expressed genes is attractive, since this topic has received some attention in the literature (see for example Pounds and Morris [2003] or Langaas *et al.* [2005]).

# 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 type I error. 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 the denominator is 0. The Bayesian FDR is defined as the expected value of this proportion, marginalizing with respect to the posterior distribution of the parameters in the model. 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 $\mathbf{x}_i$ 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 $\text{WL}_2\text{E}$ to obtain $\hat{w}$ and $f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})$. Alternatively, treat $\boldsymbol{\theta}$ as known and estimate only $w$ as in Section 2.3.

3a. If Bayesian control of the FDR is desired, estimate the overall density $f(\mathbf{x})$ using any suitable method, *e.g.*, kernel density estimation. Declare DE genes based on the pseudo-posterior probabilities of DE $\hat{v}_i = 1 - \hat{w}\frac{f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_i)}{\hat{f}(\mathbf{x}_i)}$.

3b. If frequentist control of the FDR is desired, compute $p$-values for the observed $\mathbf{x}_i$ using $f_{\hat{\boldsymbol{\theta}}}$ as the null distribution. Declare DE genes based on p-values adjusted by some procedure for controlling FDR.

There are several variants of the algorithm, depending on the choices made at each step. With respect to step 1, the approach can work with a number of sensibly chosen test statistics, the only requirement being that it should be reasonable to assume that for EE genes it follows the parametric form $f_{\boldsymbol{\theta}}$. 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). Step 3 offers a choice between Bayesian and frequentist thinking. For option 3a, one needs to estimate the overall density $f$. We have found kernel density estimators to perform well, as long as the tails of $f_{\boldsymbol{\theta}}$ are not thicker than those of the kernel. If $f_{\boldsymbol{\theta}}$ has too thick tails, genes with extreme test statistic values will have large $f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_i)/\hat{f}(\mathbf{x}_i)$, *i.e.* small probability of DE, which is of course undesirable. In our implementations, 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. In the remainder of this section we elaborate on 3a and 3b.

## 3.1 BAYESIAN CONTROL OF THE FDR

To estimate the FDR we proceed in an empirical Bayes manner. Let $v_i = 1 - w\frac{f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{f(\mathbf{x}_i)}$ be the posterior probability that gene $i$ is DE conditional on $w$, $\boldsymbol{\theta}$ and $f$ (and the data). Both $\text{L}_2\text{E}$ and $\text{WL}_2\text{E}$ 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 $d_i$ indicate the decision of declaring gene $i$ as DE, *i.e.*, $d_i = 1$ means declaring it DE and $d_i = 0$ declaring it EE. We compute the Bayesian FDR,

denoted $\widetilde{\text{FDR}}$:

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

as introduced in Genovese and Wasserman [2002]. The denominator in (3.1) is just the number of genes declared DE. Efron *et al.* [2001] proposed declaring DE those genes with $\hat{v}_i$ greater than a certain threshold, *i.e.*, $d_i = I(\hat{v}_i > t)$, but they leave the choice of $t$ as an arbitrary decision. 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. It is also possible to find the optimal threshold non-parametrically by doing permutations under the null hypothesis, much in the way that Storey [2007] finds the optimal threshold for his ODP test statistic. The validity of such a permutation-based approach becomes questionable when the sample size is very small, which is not unfrequent in microarray studies.

## 3.2 FREQUENTIST CONTROL OF THE FDR

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

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

Multi-dimensional $\mathbf{x}_i$ can often be reduced to a 1-dimensional test statistic through a function $g(\mathbf{x}_i)$. For example, one could choose the Wald-type statistic $g(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu})'V^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$ and the integral could be approximated via an asymptotic $\chi^2$ distribution when the $\mathbf{x}_i$ are normally distributed. More generally, however, the distribution of $g$ will not be known but the integral can be approximated by simulating 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 $\mathbf{x}_i$.

To control for an overall FDR$\leq \alpha$, in our experiments we employ the p-value adjustment of Benjamini and Hochberg [1995], although of course other type-I error controlling methods are also possible. Note that even though our approach makes no assumptions about the dependency between genes, this method for p-value adjustment does assume either independence or some form of positive association.

# 4 SIMULATION STUDY

## 4.1 SETUP

We compare via simulation the power of our partial mixture estimation approach to three popular methods, while controlling the FDR level at $\leq 0.05$. Initially we considered performing the partial mixture fit by $L_2E$, but results showed that it was always outperformed by its $WL_2E$ counterpart, so to be concise we only provide results for $WL_2E$.

We consider 3 variants of our approach. The first fits a partial normal mixture via $WL_2E$ and finds DE genes using pseudo-posterior probabilities (see Section 3.1). We refer to this variant as $WL_2E$-PP. The second variant obtains the same fit but declares DE based on Benjamini & Hochberg p-value adjustment (see Section 3.2), so we refer to it as $WL_2E$-BH. For these two variants we used the difference between the 2 group means as the test statistic.

For the third variant, we used moderated t-test statistics [Smyth, 2004] as implemented in the R functions `lmFit` and `eBayes` in the `limma` package [Smyth, 2005]. Smyth [2004] showed that, when the null hypothesis of equally expressed genes holds and the data is normally distributed, the moderated t-test statistic follows a Student's t distribution with augmented degrees of freedom. We fit a fixed-component partial t mixture via (2.8), and declare DE based on posterior probabilities. Since this is similar to some empirical Bayes methods [Efron and Tibshirani, 2002; Smyth, 2004], we refer to this variant as $WL_2E$-EBayes. The function `eBayes` estimates the augmented degrees of freedom $\hat{\nu}$, though in simulations $\hat{\nu}$ sometimes seemed to take too large a value, *e.g.* $\hat{\nu} = \infty$. This decreased slightly the quality of our fit, so we restricted $\hat{\nu}$ to be $\leq 25$.

The three competing methods are Tusher *et al.* [2001] significance analysis of microarrays (SAM) , Efron and Tibshirani [2002] empirical Bayes (EBayes) and a simple two-sample test with BH p-value adjustment. The two-sample test was a Welch t-test when normal data was being generated (see below) and a Wilcoxon test otherwise. That is, we reproduced what a data analyst that could perfectly assess the normality of the data might do. For computational speed, p-values for both two-sample tests were computed using the asymptotic normal distribution of the test statistic, rather than basing them on permutations. EBayes and SAM were used as implemented in the R library `siggenes`, except that for EBayes the threshold to declare DE was determined as described in Section 3.1 instead of using the default 0.9. Since EBayes requires specifying a test statistic upon which to fit a mixture of the form in (1.1), we used the same one as for the two-sample test (*i.e.* t-test statistic for normal data and Wilcoxon test statistic for non-normal data).

The simulation focuses on the comparison of $n$=5000 genes between 2 groups

| | 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

when we have $m$=3, 5 or 10 microarrays per group, although our approach is applicable in non-balanced situations. First, we set a fixed proportion of genes to be EE, over and under-expressed. We consider three possibilities: (0.8,0.1,0.1), (0.95,0.025,0.025) and (0.95,0.04,0.01). Second, we generate expression values according to one of the 4 scenarios described in Table 1. In the first scenario all distributions are normal, whereas in the second we have normality under EE but uniformity otherwise. We also consider two scenarios that do not satisfy the partial mixture hypothesis of normality under EE. In the third scenario we use uniform distributions and in the fourth we use t-distributions with 5 degrees of freedom and a general location parameter.

The 3 choices of mixing proportions and the 4 distributional scenarios give rise to 12 distinct simulation configurations. Actually, we also analyzed some additional scenarios with smaller amounts of DE. For example, in the normal-normal case we set the mean for over and under-expressed genes to be 1 and -1, respectively. In these circumstances all methods performed very poorly due to the smaller signal-to-noise ratio and are not reported here.

Once the expression data was generated, all the described methods were used to obtain a list of DE genes and we computed the number of correctly and incorrectly classified genes. We estimated the power and FDR of each method by repeatedly generating data, computing the observed power and FDR for each simulated dataset, and then averaging the results. The number of repetitions was large enough to ensure that the width of the 95% confidence interval was $\leq 0.01$. In most scenarios 100 repetitions were enough.

## 4.2 RESULTS

We now present the findings of the simulation set up in Section 4.1. Table 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. Table 3 reports the corresponding estimated FDR. We observe that the $WL_2E$-PP and $WL_2E$-BH variants of our partial mixture algorithm perform very similarly, suggesting that both are equally desirable ways of adjusting for multiple comparisons.

| | 80%/10%/10% | | | 95%/4%/1% | | | 95%/2.5%/2.5% | | |
|---|---|---|---|---|---|---|---|---|---|
| | $m$=3 | 5 | 10 | 3 | 5 | 10 | 3 | 5 | 10 |
| **normal-normal** | | | | | | | | | |
| $WL_2E$-PP | 14 | 65 | 98 | 8 | 40 | 90 | 11 | 46 | 93 |
| $WL_2E$-BH | 12 | 62 | 97 | 11 | 46 | 92 | 11 | 46 | 93 |
| $WL_2E$-EBayes | 19 | 59 | 97 | 5 | 28 | 89 | 07 | 31 | 90 |
| SAM | 0 | 53 | 96 | 0 | 33 | 91 | 0 | 34 | 90 |
| EBayes | 0 | 37 | 94 | 0 | 12 | 80 | 0 | 16 | 84 |
| t-test-BH | 0 | 24 | 93 | 0 | 2 | 79 | 0 | 2 | 78 |
| **normal-uniform** | | | | | | | | | |
| $WL_2E$-PP | 55 | 85 | 99 | 43 | 75 | 98 | 38 | 71 | 97 |
| $WL_2E$-BH | 53 | 84 | 99 | 42 | 75 | 98 | 43 | 75 | 98 |
| $WL_2E$-EBayes | 45 | 79 | 99 | 30 | 65 | 98 | 28 | 62 | 98 |
| SAM | 0 | 22 | 96 | 0 | 49 | 83 | 0 | 49 | 83 |
| EBayes | 0 | 0 | 86 | 0 | 0 | 41 | 0 | 0 | 57 |
| Wilcoxon-BH | 0 | 0 | 91 | 0 | 0 | 72 | 0 | 0 | 72 |
| **uniform-uniform** | | | | | | | | | |
| $WL_2E$-PP | 1 | 30 | 90 | 1 | 20 | 77 | 1 | 15 | 73 |
| $WL_2E$-BH | 0 | 27 | 89 | 1 | 19 | 76 | 1 | 19 | 77 |
| $WL_2E$-EBayes | 4 | 29 | 87 | 1 | 9 | 64 | 1 | 8 | 61 |
| SAM | 0 | 23 | 86 | 0 | 1 | 70 | 0 | 0 | 70 |
| EBayes | 0 | 0 | 72 | 0 | 0 | 17 | 0 | 0 | 31 |
| Wilcoxon-BH | 0 | 0 | 70 | 0 | 0 | 37 | 0 | 0 | 35 |
| **t-t** | | | | | | | | | |
| $WL_2E$-PP | 77 | 96 | 100 | 62 | 90 | 100 | 58 | 89 | 100 |
| $WL_2E$-BH | 74 | 95 | 100 | 61 | 89 | 100 | 61 | 90 | 100 |
| $WL_2E$-EBayes | 67 | 94 | 100 | 34 | 84 | 99 | 29 | 82 | 99 |
| SAM | 0 | 88 | 99 | 0 | 80 | 99 | 0 | 79 | 99 |
| EBayes | 0 | 0 | 99 | 0 | 0 | 98 | 0 | 0 | 98 |
| Wilcoxon-BH | 0 | 0 | 99 | 0 | 0 | 95 | 0 | 0 | 95 |

Table 2: Power for simulation study (in %). $m$: number of observations per group. 80%/10%/10% indicates 80% of the genes were EE, 10% over-expressed and 10% under-expressed. PP is posterior probability and BH is Benjamini-Hochberg.

| | 80%/10%/10% | | | 95%/4%/1% | | | 95%/2.5%/2.5% | | |
|---|---|---|---|---|---|---|---|---|---|
| | $m$=3 | 5 | 10 | 3 | 5 | 10 | 3 | 5 | 10 |
| **normal-normal** | | | | | | | | | |
| $WL_2E$-PP | 1 | 3 | 5 | 4 | 5 | 5 | 4 | 5 | 5 |
| $WL_2E$-BH | 1 | 3 | 4 | 4 | 4 | 5 | 4 | 4 | 5 |
| $WL_2E$ - EBayes | 2 | 2 | 4 | 2 | 2 | 3 | 2 | 1 | 3 |
| SAM | 0 | 2 | 3 | 0 | 3 | 4 | 0 | 3 | 4 |
| EBayes | 0 | 2 | 3 | 1 | 3 | 3 | 0 | 3 | 3 |
| t-test-BH | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 |
| **normal-uniform** | | | | | | | | | |
| $WL_2E$-PP | 3 | 4 | 5 | 4 | 5 | 5 | 5 | 5 | 5 |
| $WL_2E$-BH | 2 | 3 | 4 | 4 | 5 | 5 | 4 | 5 | 5 |
| $WL_2E$-EBayes | 1 | 2 | 5 | 2 | 2 | 6 | 2 | 2 | 6 |
| SAM | 0 | 5 | 5 | 0 | 2 | 5 | 0 | 2 | 4 |
| EBayes | 0 | 0 | 5 | 0 | 0 | 4 | 0 | 0 | 4 |
| Wilcoxon-BH | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 4 |
| **uniform-uniform** | | | | | | | | | |
| $WL_2E$-PP | 0 | 1 | 4 | 0 | 1 | 4 | 0 | 2 | 4 |
| $WL_2E$-BH | 0 | 1 | 3 | 0 | 1 | 4 | 0 | 1 | 4 |
| $WL_2E$-EBayes | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 1 |
| SAM | 0 | 3 | 4 | 0 | 0 | 4 | 0 | 0 | 5 |
| EBayes | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 3 |
| Wilcoxon-BH | 0 | 0 | 4 | 0 | 0 | 5 | 0 | 0 | 4 |
| **t-t** | | | | | | | | | |
| $WL_2E$-PP | 9 | 8 | 8 | 20 | 14 | 11 | 19 | 13 | 10 |
| $WL_2E$-BH | 8 | 7 | 6 | 19 | 13 | 9 | 19 | 13 | 9 |
| $WL_2E$-EBayes | 3 | 5 | 8 | 2 | 4 | 9 | 2 | 4 | 9 |
| SAM | 0 | 2 | 3 | 0 | 3 | 4 | 0 | 3 | 4 |
| EBayes | 0 | 0 | 7 | 0 | 0 | 11 | 0 | 0 | 12 |
| Wilcoxon-BH | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 4 |

Table 3: FDR for simulation study (in %). $m$: number of observations per group. 80%/10%/10% indicates 80% of the genes were EE, 10% over-expressed and 10% under-expressed. PP is posterior probability and BH is Benjamini-Hochberg.

These two $WL_2E$ approaches were the most powerful under almost all conditions in the normal-normal, normal-uniform and uniform-uniform cases, with very significant advantages being observed for sample sizes of 3 and 5. The fixed-component $WL_2E$-EBayes also performed quite well. When the sample size was 3 the power of all competitors was virtually zero in all scenarios, whereas $WL_2E$ achieved a power between 12%-19% in the normal-normal and 45%-55% in the normal-uniform scenario. In all these situations $WL_2E$ controlled the FDR below the desired 5% level, including the uniform-uniform where the normality assumption is violated. The one exception is $WL_2E$-EBayes that presented an FDR=6% in two of the normal-uniform scenarios.

In the t-t scenario $WL_2E$-BH and $WL_2E$-PP presented an FDR well above 5%, especially for smaller sample sizes, while $WL_2E$-EBayes had better FDR levels. 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. It should be noted that in this scenario EBayes also failed to control the FDR.

SAM was the best among the competitors, exceeding their power while controlling the FDR below the desired level. EBayes outperformed the 2-sample t-test in the normal-normal case but it tended to be worse than the 2-sample Wilcoxon test in the normal-uniform and uniform-uniform cases. Note that except in the normal-normal scenario neither method detected any genes for sample sizes of 3 or 5 observations per group.

# 5 CASE STUDIES

We analyze the data from the Apolipoprotein AI (apo AI) experiment presented by Callow *et al.* [2000] and from the leukemia study of Golub *et al.* [1999] using the methods described in the simulation study of Section 4: the three variants of $WL_2E$, SAM, EBayes and t-test BH, all set to control the FDR $\leq 0.05$. When Dudoit *et al.* [2003] analyzed the apo AI dataset with several methods, they found 8 DE genes out of 6384. The analysis of Golub's dataset in the original paper uses a neighbourhood-based analysis that found 1000 DE genes out of 6817. That is, the first dataset represents the case in which few DE genes are expected, whereas in the second the proportion of DE genes is probably relatively large.

## 5.1 APOLIPOPROTEIN EXPERIMENT

### 5.1.1 DESCRIPTION

The apo AI experiment concerned lipid metabolism and artherosclerosis susceptibility in mice. The experiment compared 8 apo AI knock-out mice with 8 inbred control mice in terms of gene expression. cDNA was obtained from mRNA by reverse transcription and it was hybridized to microarrays with 6384 probes. A common reference sample for all hybridizations (knock-out and control) was obtained by pooling cDNA from the 8 control mice. We obtained the dataset from `http://www.stat.berkeley.edu/users/terry/zarray/Data/ApoA1/rg_a1ko_morph.txt`.

For 2 groups and 8 observations per group there are 12,870 possible permutations, which is a large enough number for SAM and EBayes to estimate the null distribution $f_0$ reliably.

### 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 $WL_2E$-PP and $WL_2E$-BH we use the difference between means $x_i$ as a test statistic, *i.e.* we do not divide by its estimated standard error as is done in a t-test. Therefore, we perform a second step to ensure that the variance of the test statistic is constant across genes. Denote the sum of the two group means for the $i^{th}$ gene as $A_i$. In the fashion of the so-called "MA-plots" [Dudoit *et al.*, 2002a], we center the data by obtaining a `lowess` local least squares fit of $x_i$ versus $A_i$ and substract the predicted mean, to obtain the residuals $e_i$, $i = 1 \ldots n$. We then regress $e_i^2$ versus $A_i$ via `lowess`, which gives a gene-specific estimate for the residual variance. The variance-stabilized test statistic is obtained by dividing $e_i$ by the square root of its estimated variance. In both fits the smoothing parameter is chosen by minimizing the mean absolute error by cross-validation.

Figure 2(a) displays the distribution of $x_i$ for different values of $A_i$ after the variance stabilization procedure. The mean and variance of $x_i$ is roughly constant for all $A_i$ values, suggesting that the partial mixture assumption of the test statistic being identically distributed is not unreasonable.

The $WL_2E$ 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 2(b), which presents a QQ-normal plot using the weighted $L_2E$ estimates. The horizontal lines contain $\hat{w}$ =96% of the test statistic values, *i.e.*, observations outside the region delimited by the lines are not considered to

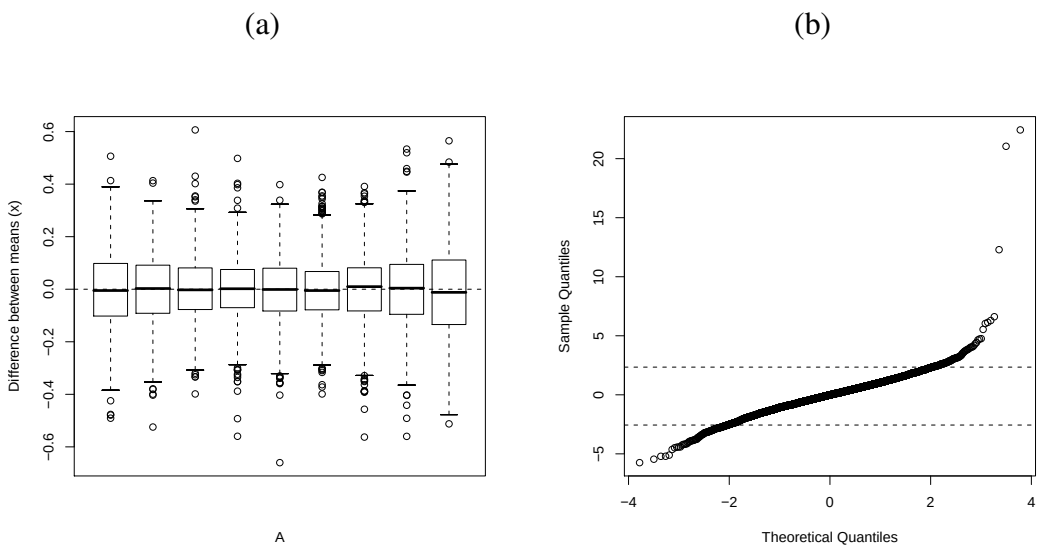


Figure 2: Assessing partial mixture assumptions for apo AI dataset. (a): mean intensity values (A) are categorized in 10 groups according to the observed quantiles. (b): normal quantile plot of the test statistic. Horizontal lines contain a proportion $\hat{w}$ of the data.

arise from equally expressed genes. We consider that the normality assumption is plausible, since departure from normality is observed mainly in the tails.

We also computed the regularized t-test statistics needed for the $WL_2$E-EBayes variant of our algorithm, as described in Section 4.1. The estimated augmented degrees of freedom were 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 regularized t-test statistic we produced plots analogous to those in Figure 2, finding that the assumptions were reasonable.

### 5.1.3 RESULTS

Results of the differential expression analysis are shown in Table 4. $WL_2$E-PP and $WL_2$E-BH declared a substantially larger number of genes to be DE than the other methods. $WL_2$E-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). None of the three competitors finds any gene to be over-expressed in the knock-out mice. The t-test

| | $WL_2$E-PP | $WL_2$E-BH | $WL_2$E-EBayes | EBayes | SAM | t-test-BH |
|---|---|---|---|---|---|---|
| KO over-expr. | 35 | 32 | 1 | 0 | 0 | 0 |
| KO under-expr. | 25 | 27 | 9 | 0 | 0 | 8 |

Table 4: 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

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 2(b) revealed a stronger departure from normality than that observed for the full dataset. $WL_2$E-PP and $WL_2$E-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. $WL_2$E-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. These findings suggest that $WL_2$E-PP, $WL_2$E-BH and t-test BH can lead to an inflated FDR when the normality assumption is violated, but that $WL_2$E-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 dataset is already normalized per Golub *et al.* [1999]. We first considered using the difference between group means as the test statistic to which to apply $WL_2E$-PP and $WL_2E$-BH, in the same way that is described in Section 5.1.2 for the apo AI study. However, a qq-normal plot analogous to that in Figure 2(b) revealed a serious departure of normality. For this reason, we decided to use the regularized t-test statistic that we have been using for the $WL_2E$-EBayes variant of our algorithm also for the two other variants.

The partial mixture assumptions are assessed in Figure 3. Panel (a) reveals that the distributional shape of the test statistic is approximately the same for all values of the mean intensity values $A_i$. A $WL_2E$ fit gives $\hat{\mu} = 0.24$, $\hat{\sigma} = 1.79$ and $\hat{w} = 0.99$. That is, $WL_2E$ regards $1\%$ of the test statistic to be outliers. However, $27.8\%$ of the genes have a test statistic value exceeding 2 in absolute value, indicating that the proportion of DE genes should be higher than $1\%$. Also, a SAM analysis estimated $\hat{w} = 0.53$. To solve this discrepancy we use the fact that, under the null hypothesis, the moderated t-test statistic follows a t distribution with augmented degrees of freedom. The estimated augmented degrees of freedom are 37, slightly higher than the 36 that a classical t-test statistic would have, and applying (2.8) we obtain $\hat{w} = 0.63$, which seems a more reasonable value. The Student's t qq-plot in panel (b) suggests that the t assumption is reasonable for $63\%$ of the test statistic values that are closer to the mean.

The difference in the estimated proportion of DE genes between the two $WL_2E$ fits should not be too surprising: in the simulation example of Section 2 we also observed considerable improvements when fixing $f_{\boldsymbol{\theta}}$. That is, it can be quite benefitial to use theoretical considerations when obtaining a partial mixture fit, rather than using an optimization algorithm blindly. Our interpretation for this particular dataset is that $WL_2E$ breaks down in the presence of a large amount of outliers, but a fixed t component fit is more robust.

### 5.2.3 RESULTS

$WL_2E$-EBayes found 744 genes, whereas EBayes declared 881 genes as DE, SAM 662 and the Wilcoxon test with BH p-value adjustment 610. In terms of concordance between methods, $WL_2E$-EBayes classified about $94\%$ of the genes in the same category that EBayes did (equally, over or under-expressed). For SAM this percentages were around $87\%$, and for Wilcoxon test around $92\%$.

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. With this smaller sample size EBayes, SAM and the Wilcoxon test declared 0 genes as DE,

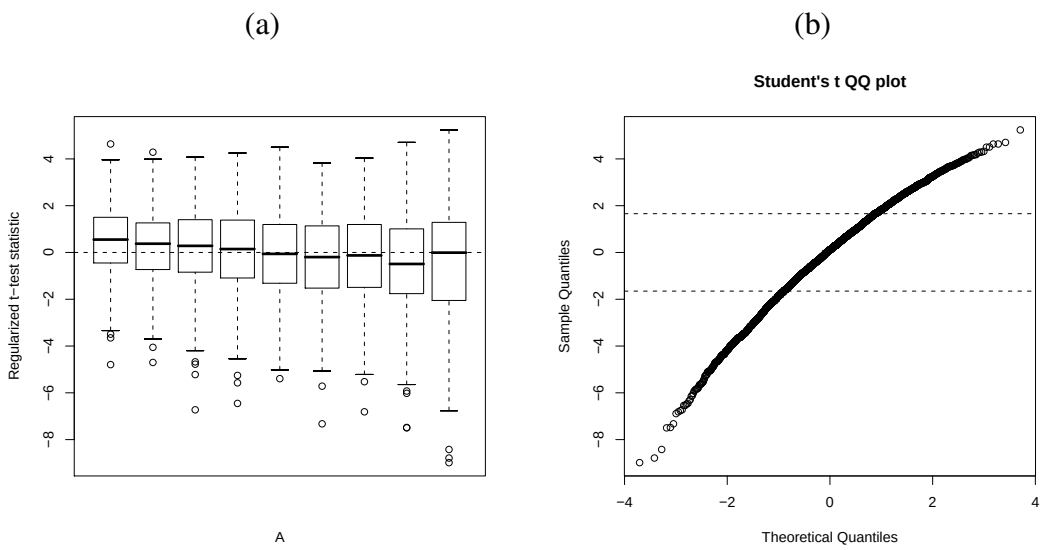


Figure 3: Assessing partial mixture assumptions for Leukemia dataset. (a): mean intensity values (A) are categorized in 10 groups according to the observed quantiles. (b): horizontal lines contain a proportion $\hat{w} = 0.62$ of the data.

whereas $WL_2E$-EBayes found 11 genes, 10 of which had also been declared as DE when analyzing the full dataset.

# 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 on equally expressed genes only, and can handle small group sizes, in contrast to other methods that require a full probability model or larger group sizes for permutation-based null sampling. To perform partial mixture estimation, we have developed an iterative weighted $L_2E$ criterion that improves upon the performance of the $L_2E$ criterion originally proposed by Scott [2004a,b]. Also, we have shown that fixing the parameters of the EE genes distribution can further improve the fit, making it more resistant to outliers and providing simple closed-form expressions to estimate the proportion of EE genes. Both criteria are extremely fast computationally; in our experience, the whole procedure implemented in R takes a few seconds to run, in the worst case scenario.

Our approach requires making only two assumptions. First, we assume that the observed values of the test statistic used to classify genes are identically dis-

tributed 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. 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 can have important effects on the final results. For example, we expect test statistics that borrow information across genes to perform better than those that are computed for each gene separately.

Simulation studies and real data analyses have been used to compare our approach with EBayes, SAM 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. For example, in some simulations with a sample size of 3 per group the power was between 12%-19% for $WL_2E$, and virtually 0% for all other methods. In the apolipoprotein dataset we detected more genes than the competitors, with SAM and EBayes finding no genes. In the leukemia dataset our approach found a number of genes comparable to the competitors. However, when analyzing a subset of only 5 observations per group we found 11 genes while the other approaches did not detect any (10 out of these 11 genes were also found when analyzing the full dataset). Of course, in real datasets it is often not known what genes are DE and therefore it is difficult to be certain of which method is performing best.

In our opinion, $WL_2E$-EBayes is the most attractive variant as a general purpose approach, since it seems to have good detection power while preserving a better control of the FDR when the semi-parametric assumptions do not hold. However, in some simulations we found that $WL_2E$-BH and $WL_2E$-PP were preferrable, since they were less conservative in controlling the FDR. In general SAM seems to perform the best among the competitors, particularly in the simulations. This seems to agree with 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 dataset.

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.

Possibilities for future work include generalization to other parametric forms, such as the F distribution for multi-group or time-course problems. Another direction is the development of alternative methods for partial mixture estimation.

## Acknowledgments

We thank David W. Scott for his useful comments.

## 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. 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.

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.

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.

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.

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 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.