A Laplace Mixture Model for Identification of Differential Expression in Microarray Experiments

Debmani Bhowmick, A. C. Davison*, and Darlene R. Goldstein

Ecole Polytechnique Fédérale de Lausanne
Institute of Mathematics
EPFL-FSB-IMA, Station 8
CH-1015 Lausanne, Switzerland

* corresponding author
email: anthony.davison@epfl.ch
Tel: + 41 21 693 5502
Fax: + 41 21 693 4250

February 18, 2006
Abstract

Microarrays have become an important tool for studying the molecular basis of complex disease traits and fundamental biological processes. A common purpose of microarray experiments is the detection of genes that are differentially expressed under two conditions, such as treatment versus control, or wild-type versus knock-out.

We introduce a Laplace mixture model as a long-tailed alternative to the normal distribution when identifying differentially expressed genes in microarray experiments, and provide an extension to asymmetric over- or under-expression. This model permits greater flexibility than models in current use as it has the potential, at least with sufficient data, to accommodate both whole genome and restricted coverage arrays.

We also propose a REML-type approach to hyperparameter estimation which is equally applicable in the Normal mixture case.

The Laplace model appears to give some improvement in fit to data, although simulation studies show that our method performs similarly to several other statistical approaches to the problem of identification of differential expression.

Keywords: Laplace distribution; Marginal likelihood; Microarray experiment; Mixture model; REML.
1 Introduction

Microarrays have become an important tool for studying the molecular basis of complex disease traits and fundamental biological processes. Two-channel microarrays, such as spotted cDNA or long oligonucleotide arrays, measure relative gene expression in two samples. Once preprocessed, the data from such arrays take the form of normalized base 2 logarithm of the expression ratios. A common purpose of microarray experiments is the detection of genes that are differentially expressed under two conditions, such as treatment versus control, or wild-type versus knock-out. Numerous statistical methods have been proposed for identification of differential expression, with new ones continuing to be introduced.

In early analyses such as Schena et al. (1995, 1996), fold change between conditions exceeding a constant was used to identify differentially expressed genes. This method performs poorly, however, because it ignores the different variability of expression across genes. Such variability can be taken into account by using a \( t \)-test on the average log fold change, but variation in gene expression is poorly estimated with small numbers of replicates, so genes with artificially low variance may be selected even if they are not truly differentially expressed. Numerous refinements have been proposed in order to reduce the numbers of false positive and false negative results. Tusher et al. (2001) and Efron et al. (2000) suggested adding a constant to the \( t \) denominator so that it does not become too small. Lönnstedt and Speed (2002) proposed a normal mixture model for the
gene expression data and defined a log posterior odds statistic, their $B$-statistic, for ranking genes. Their approach was extended to linear models for more general designs by Smyth (2004); these methods are implemented in the R package limma as part of the BioConductor project (www.bioconductor.org). Gottardo et al. (2003) use a similar approach but also include a heuristic, iterative method for estimating the proportion of differentially expressed genes.

The present paper makes three main contributions: we consider robust and asymmetric variants of the mixture modelling approach, we propose a new approach to parameter estimation, and we perform a comparative study intended to elucidate key features of such methods. The methods studied include those listed above, along with our own proposal of a Laplace mixture as a long-tailed alternative to the normal distribution for mean gene expression across replicate arrays proposed by Lönnstedt and Speed (2002).

This paper is organised as follows: §2 describes the model, with estimation of the hyperparameters discussed in §3. Simulation studies outlined in §4 show that this method performs similarly to several other statistical approaches to the problem of identification of differential expression. In §5 we apply our method to a published dataset and compare its results with other methods. The paper ends with a brief discussion.
2 Laplace mixture model

2.1 Symmetric gene expression

Lönnstedt and Speed (2002) propose the use of a mixture of a point mass at zero and a normal distribution for mean gene expression, and of an inverse gamma distribution for the single gene variances. Gaussian variation is the exception rather than the rule in practice, however, so we consider using a Laplace distribution as a potentially more realistic longer-tailed alternative to the normal law. We assume that the data are normalized base 2 logarithms of fold changes, as would be the case for two-channel arrays. It is not difficult to apply our approach to single channel technologies, by modeling the single channel log intensities, and using a contrast matrix to obtain corresponding log ratios.

We suppose initially that it is reasonable to expect approximately symmetric over- and under-expression, as when an array encompasses an entire genome. Arrays with selected genome coverage are considered in §2.2.

We model normalized relative expression measures on each of $G$ genes independently, but in this section we ease the notation by confining our attention to data from a single gene. Although in reality genes do interact with each other, the independence assumption is a useful simplification. The measured relative expression levels for this gene in $n$ replicates, $y = (y_1, \ldots, y_n)$, are taken to be independent and normally distributed with mean relative expression level $\mu$ and
variance $\sigma^2$, i.e. $y_1, \ldots, y_n \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We denote the average and sample variance of the observations $y$ by $\bar{y}$ and $s^2$.

We let $\omega$ represent the probability that this gene is differentially expressed, and assume that the mean expression level $\mu$ has a Laplace distribution with mean zero and variance $2\tau^2$, so that the density is given by

$$f(\mu; \tau) = (2\tau)^{-1} \exp\left[-|\mu|/\tau\right], \quad -\infty < \mu < \infty.$$

If the gene is not differentially expressed, its mean expression level is zero. Thus $\mu$ may be regarded as a random draw from the mixture distribution $\omega \mathcal{L}(0, \tau) + (1 - \omega)\delta_0$, where $\mathcal{L}(0, \tau)$ represents the Laplace distribution for differential expression and $\delta_0$ denotes the distribution which places unit mass at $\mu = 0$. Conditional on $\sigma^2$, we assume that $\tau = \sigma^2 V$, where $V > 0$ represents a type of generalized signal to noise ratio. Thus if the variability $\sigma^2$ of expression levels across replicates is large, the variation of $\mu$ will also be large. This assumption, which simplifies subsequent mathematical developments, is checkable from the data; in the cases we have examined it seems fairly reasonable. Variability of measured expression levels is not constant across genes, and we follow Lönnstedt and Speed (2002) in assuming an inverse gamma distribution for $\sigma^2$, that is $\sigma^2 \sim IG(\gamma, \alpha)$, where $\gamma$ and $\alpha$ are shape and scale parameters, respectively. We use a similar notation for the gamma distribution $\mathcal{G}(\gamma, \alpha)$.

We treat $\omega$, $V$, $\alpha$ and $\gamma$ as hyperparameters, and in §3.1 discuss their estimation.
A straightforward but involved calculation shows that the marginal posterior distribution of \( \mu \), like the prior for \( \mu \), is a mixture of continuous and discrete components. It may be expressed as

\[
f(\mu \mid y) = \frac{\rho}{1 + \rho} h(\mu) + \frac{1}{1 + \rho} \delta(\mu), \quad -\infty < \mu < \infty,
\]

where \( \delta(\mu) \) is a Dirac delta function and

\[
h(\mu) = \frac{\Gamma\left(\frac{\nu+2}{2}\right) a_{\text{sign}(\mu)}^{-(\nu+2)/2}}{d\sqrt{(\nu + 1)\pi \Gamma(\nu + 12)}} \left(1 + \frac{1}{\nu + 1} \left\{\frac{\mu - b_{\text{sign}(\mu)}}{c_{\text{sign}(\mu)}}\right\}^2\right)^{-(\nu+2)/2}.
\]

Here \( \Gamma(u) \) represents the gamma function, \( \nu = n + 2\gamma \),

\[
a_{\pm} = \frac{1}{2} \left\{\frac{2}{\alpha} + (n - 1)s^2 \pm 2\bar{y}/V - 1/(nV^2)\right\},
\]

\[
b_{\pm} = \bar{y} \mp 1/(nV),
\]

\[
c_{\pm} = \mp \left[2n(\nu + 1)^{-1}a_{\pm}\right]^{1/2},
\]

\[
d = a_{-}^{-(\nu+2)/2}c_{-}F_{\nu+1}(b_{-}/c_{-}) + a_{+}^{-(\nu+2)/2}c_{+}F_{\nu+1}(b_{+}/c_{+}),
\]

and \( F_{\nu+1}(t) \) denotes the cumulative distribution function of a Student \( t \) variable with \( \nu + 1 \) degrees of freedom. Thus \( h \), the posterior density of \( \mu \) given that a gene is differentially expressed, comprises two off-centred Student \( t \) densities placed back to back at the origin.

The posterior odds of differential expression may be written

\[
\rho = \frac{\omega}{1 - \omega} \times \frac{\sqrt{(\nu + 1)\pi \Gamma(\frac{\nu+1}{2}) \left(\frac{1}{2} \sum_j y_j^2 + 1/\alpha\right)^{\nu/2}}}{2\Gamma(\frac{\nu}{2})}. \tag{2.3}
\]

The second term on the right of (2.3) is the Bayes factor for differential expression \( \mu \neq 0 \) relative to \( \mu = 0 \).
2.2 Asymmetric gene expression

Most methods for the detection of differential expression assume equal over- and under-expression, but this may be unrealistic when arrays contain genes chosen for their special interest to the investigator, or in other cases where genome coverage is restricted. We therefore consider an asymmetric extension of our mixture model.

The density of an asymmetric Laplace variable, which may be obtained as the difference of two independent exponential variables with different means, may be expressed as (Kotz et al.; 2001, Chapter 3)

\[ f(\mu; \tau, \beta) = (2\tau)^{-1} \exp \left[ \frac{-|\mu|}{\tau B_{\text{sign}(\mu)}} \right], \quad -\infty < \mu < \infty, |\beta| < 1, \]

where \( B_{\pm} = 1 \pm \beta \). Taking \( \beta = 0 \) yields the symmetric Laplace distribution, while \( \beta \to 1 \) gives increasing weight to the right tail of the density, and conversely when \( \beta \to -1 \). We write the corresponding distribution as \( \mathcal{L}(\beta, \tau) \), in terms of which the asymmetric Laplace mixture model for \( \mu \) may be written \( \omega \mathcal{L}(\beta, \tau) + (1 - \omega)\delta_0 \).

A calculation generalising that in the symmetric case shows that the posterior density of \( \mu \) is again given by (2.1) and (2.2), but with

\[
\begin{align*}
a_{\pm} &= \frac{1}{2} \left\{ 2/\alpha + (n - 1)s^2 \pm 2\bar{y}/(B_{\pm}V) - 1/(nB_{\pm}^2V^2) \right\}, \\
b_{\pm} &= \bar{y} \mp 1/(nB_{\pm}V).
\end{align*}
\]

The posterior odds of differential expression are again given by (2.3), and the hyperparameters \( \omega, V, \alpha, \gamma, \) and \( \beta \) may be estimated by an empirical Bayes procedure, as in the symmetric case.
2.3 Bayesian estimation

The discussion above relates to data from a single gene, but in applications many genes will be considered, making it useful to reduce the entire marginal posterior density of $\mu$ to a single value. One way to do this is to consider the posterior median $\tilde{\mu}$ of $\mu$. If the posterior probability that $\mu > 0$ exceeds one-half, then $\tilde{\mu} > 0$, whereas if the posterior probability that $\mu < 0$ exceeds one-half, then $\tilde{\mu} < 0$. If neither of these events occurs, then the point mass at $\mu = 0$ ensures that $\tilde{\mu} = 0$. Thus the use of a posterior median to estimate $\mu$ yields a thresholding rule according to which $\tilde{\mu} = 0$ if the posterior distribution of $\mu$ is too closely concentrated around zero. This method has been used for selection of non-zero components in Bayesian implementations of wavelet smoothing (Johnstone and Silverman; 2005).

3 Estimation procedures

3.1 Marginal likelihood estimation

We now suppose that measurements $y_g = (y_{g1}, \ldots, y_{gn_g})$ are available on gene $g$, for $g = 1, \ldots, G$, and that the genes behave independently. The posterior density of the expression level for gene $g$, $\mu_g$, depends on the data $y$ and on hyperparameters $(\omega, V, \alpha, \gamma)$ in the symmetric case and $(\omega, V, \alpha, \gamma, \beta)$ in the asymmetric case.
Consider the symmetric model under which
\[ y_{g1}, \ldots, y_{gn_g} \mid \mu_g, \sigma_g^2 \sim \mathcal{N}(\mu_g, \sigma_g^2), \quad \mu_g \mid \sigma_g^2 \sim \omega \mathcal{L}(0, V\sigma_g^2) + (1-\omega)\delta_0, \quad \sigma_g^2 \sim IG(\alpha, \gamma), \]
for \( g = 1, \ldots, G \). A simple and in principle efficient approach to estimating the hyperparameters is to maximise the log marginal likelihood
\[
\sum_{g=1}^{G} \log f(y_g; \omega, V, \alpha, \gamma) \quad (3.1)
\]
based on the marginal densities
\[
f(y_g; \omega, V, \alpha, \gamma) = \int f(y_g \mid \mu_g, \sigma_g^2) \pi(\sigma_g^2 \mid \alpha, \gamma) \pi(\mu_g \mid \omega, V) \, d\mu_g d\sigma_g^2, \quad g = 1, \ldots, G.
\]
Unfortunately the resulting estimates can be quite poor, with the estimated proportion of differentially expressed genes \( \omega \) being particularly unstable. Lönnstedt and Speed (2002) skirted this difficulty by fixing this proportion at an arbitrary value and then estimating other parameters using the method of moments, while Gottardo et al. (2003) use a heuristic approach. Below we describe a different likelihood procedure which can yield fairly stable estimates.

The basis of our approach is the elementary fact that under a model with Gaussian errors, the average and sample variance \( \bar{y}_g \) and \( s_g^2 \) are jointly sufficient statistics. That is,
\[
f(y_g \mid \mu_g, \sigma_g^2) \propto f(\bar{y}_g \mid \mu_g, \sigma_g^2) f(s_g^2 \mid \sigma_g^2),
\]
where the terms on the right are normal and chi-squared density functions. As the marginal density of \( s_g^2 \) depends only on \( \sigma_g^2 \), hyperparameters \( \varphi \) of the density
for $\sigma^2_g$ can be estimated by maximizing the marginal likelihood obtained as the product of
\[
\int f(s^2_g | \sigma^2_g) \pi(\sigma^2_g; \varphi) \, d\sigma^2_g, \quad g = 1, \ldots, G.
\] (3.2)
The resulting estimates do not take into account any information about $\varphi$ contained in $\bar{y}_1, \ldots, \bar{y}_G$, but have the useful property that they do not depend on parameters $\theta$ of the prior density for $\mu_g$. Estimates of $\varphi$ obtained from the marginal likelihood based on (3.2) can be substituted into (3.1), which is then maximised with respect to the remaining hyperparameters. The result is a two-stage procedure, which is compared to an overall optimization in the following section.

An important special case is where the density of $\sigma^2_g$ is inverse gamma, that is, $1/\sigma^2_g$ is gamma distributed with shape and scale parameters $\gamma$ and $\alpha$ respectively. Then the marginal distribution of $\gamma \alpha s^2_g$ is $F_{n-1,2\gamma}$, and estimates $\hat{\alpha}, \hat{\gamma}$ are readily obtained by maximum likelihood estimation based on this distribution. Moreover the suitability of the inverse gamma distribution can be indirectly assessed by plotting the ordered $s^2_g$ against quantiles of the fitted $F$ distribution. We have found that with this model, standard routines can be applied without difficulty in both optimization stages.

The above approach to estimation of the hyperparameters associated with the variances $\sigma^2_g$ applies also to designed experiments, provided they are replicated, because it is independent of any superstructure for the $\mu_g$. It also extends to cases where the error structure is equivariant, that is, where we may write $y_{gj} =$
\( \mu_g + \sigma_g \varepsilon_{gj} \), with the \( \varepsilon_{gj} \) having a known density \( w \). Let \( y_{g(1)} < \cdots < y_{g(n)} \) represent the ordered \( y_{gj} \). The differences of order statistics \( y_{g(j)} - y_{g(1)} \) depend only on \( \sigma_g^2 \) and \( w \), and so \( \varphi \) may be estimated using a likelihood based on the marginal densities

\[
\int \sigma_g^{-n} \int \prod_{j=1}^{n} w \left( \frac{y_{g(j)} - y_{g(1)} + u}{\sigma_g} \right) du \pi(\sigma_g^2, \varphi) d\sigma_g^2, \quad g = 1, \ldots, G.
\]

Typically these integrals must be obtained numerically or approximated, unlike when \( w \) is the normal density and the prior density of \( \sigma_g^2 \) is inverse gamma.

### 3.2 Numerical comparison

In this section we compare our two-stage and certain other approaches to hyperparameter estimation using simulated data. The simulation study is based on three scenarios. Initially we generated 100 datasets each with \( G = 3000 \) genes and \( n = 2 \) replicates from the Laplace mixture model with \( \omega = 0.1 \) and \( V = 1.2, 2, 3 \).

For Scenario I, we simulate datasets with parameter values chosen to match the simulations of L"onnstedt and Speed (2002). Variances \( \sigma_g^2 \) are generated from the inverse gamma distribution with parameter values \( \alpha = 0.04 \) and \( \gamma = 2.8 \) estimated from the SR-BI dataset (Callow et al.; 2000) give \( \text{E}(\sigma_g^2) \approx 14 \) and \( \text{var}(\sigma_g^2) \approx 241 \).

For Scenario II, we generate datasets using parameters estimated from a data set on 150 defence related genes on the plant \( \text{Arabidopsis thaliana} \) (Reymond et al.; 2000). In this case the mean and variance of the error variances were lower, being roughly 2 and 8 respectively, with corresponding parameter values \( \alpha = 0.33 \) and
Figure 1: $M$ versus log variance plot of a simulated dataset. Highlighted points are truly differentially expressed genes, with mean $\mu_g$ simulated from a Laplace distribution.

$\gamma = 2.5$. Scenario III uses parameter values intermediate between the other two, taking the mean and variance of the error variances to be 5 and 50, respectively; this yields $\alpha = 0.13$ and $\gamma = 2.5$.

For each simulated dataset, we estimated the parameters by our two-stage marginal likelihood procedure (MML) and by ordinary maximum likelihood (ML). In almost all cases the new procedure converged much faster than maximum likelihood estimation, which failed to converge in about 5% of cases. Table 1
Table 1: Root mean squared error (bias) ($\times 10^2$) of maximum marginal likelihood (MML) and maximum likelihood (ML) estimators of the parameters of the symmetric Laplace mixture model. In all cases the proportion of differentially expressed genes is $\omega = 0.1$ and the number of replicates is $n = 2$.

<table>
<thead>
<tr>
<th>$V$</th>
<th>Scenario</th>
<th>$\gamma$</th>
<th>$\alpha$</th>
<th>$\omega$</th>
<th>$V$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>I</td>
<td>13 (-0.4)</td>
<td>0.2 (-0.01)</td>
<td>1.0 (0.8)</td>
<td>13 (-7.1)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>10 (0.9)</td>
<td>1.0 (-0.1)</td>
<td>1.0 (1.0)</td>
<td>16 (-10)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>7.0 (4.7)</td>
<td>1.0 (-0.7)</td>
<td>1.0 (0.3)</td>
<td>11 (-0.5)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>16 (1.0)</td>
<td>1.0 (-0.01)</td>
<td>3.0 (0.3)</td>
<td>16 (-7.0)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>17 (3.0)</td>
<td>2.0 (0.1)</td>
<td>2.0 (1.0)</td>
<td>25 (14)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>15 (2.9)</td>
<td>1.0 (-0.1)</td>
<td>1.0 (0.7)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>13 (4.7)</td>
<td>17 (1.2)</td>
<td>17 (1.2)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>15 (2.9)</td>
<td>5.0 (0.5)</td>
<td>11 (3.0)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>7.0 (4.7)</td>
<td>17 (1.2)</td>
<td>28 (13)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>16 (1.0)</td>
<td>1.0 (1.0)</td>
<td>1.0 (0.3)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>17 (3.0)</td>
<td>2.0 (0.1)</td>
<td>2.0 (1.0)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>16 (1.0)</td>
<td>2.0 (-3.0)</td>
<td>3.0 (0.3)</td>
<td>16 (-7.0)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>15 (2.9)</td>
<td>5.0 (1.0)</td>
<td>11 (3.0)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>17 (3.0)</td>
<td>51 (14)</td>
<td>37 (20)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>16 (1.0)</td>
<td>2.0 (0.1)</td>
<td>2.0 (1.0)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>17 (3.0)</td>
<td>51 (14)</td>
<td>37 (20)</td>
<td>26 (-15)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>15 (2.9)</td>
<td>11 (3.0)</td>
<td>30 (10)</td>
<td>30 (-5.0)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>17 (3.0)</td>
<td>37 (20)</td>
<td>28 (12)</td>
<td>30 (-5.0)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>16 (1.0)</td>
<td>2.0 (1.0)</td>
<td>2.0 (1.0)</td>
<td>30 (-5.0)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>17 (3.0)</td>
<td>37 (20)</td>
<td>28 (12)</td>
<td>30 (-5.0)</td>
</tr>
<tr>
<td>3</td>
<td>I</td>
<td>32 (-7.0)</td>
<td>2.0 (-1.1)</td>
<td>2.0 (0.2)</td>
<td>47 (31)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>24 (4.1)</td>
<td>8.0 (1.0)</td>
<td>2.0 (0.2)</td>
<td>47 (31)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>13 (-0.3)</td>
<td>10 (2.6)</td>
<td>2.0 (0.2)</td>
<td>47 (31)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>30 (-5.0)</td>
<td>1.0 (1.0)</td>
<td>2.0 (0.2)</td>
<td>47 (31)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>28 (12)</td>
<td>2.0 (0.2)</td>
<td>2.0 (0.2)</td>
<td>47 (31)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>30 (-5.0)</td>
<td>2.0 (1.0)</td>
<td>2.0 (1.0)</td>
<td>47 (31)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>28 (12)</td>
<td>2.0 (0.2)</td>
<td>2.0 (1.0)</td>
<td>47 (31)</td>
</tr>
</tbody>
</table>

Table 1: Root mean squared error (bias) ($\times 10^2$) of maximum marginal likelihood (MML) and maximum likelihood (ML) estimators of the parameters of the symmetric Laplace mixture model. In all cases the proportion of differentially expressed genes is $\omega = 0.1$ and the number of replicates is $n = 2$.

The bias and root mean square error for the three scenarios, based on 100 datasets where both approaches converged. The new procedure generally performs best, except in the high variance scenario with $\alpha = 0.04$ and $\gamma = 2.8$, for which $\text{var}(\sigma^2_g) \gg \text{E}(\sigma^2_g)$, where both methods perform similarly.

In a second study, we generated data from the asymmetric Laplace mixture model. For each of various scenarios we generated 100 datasets each with 2000 genes. As this model is likely to be considered for subsets of genes selected because they are thought to be important, the probability of differential expression $\omega$ was taken to be appreciably higher than the first study ($\omega = .25$ or .75). The
asymmetry parameter $\beta$ was varied ($\beta = 0, 0.5, 0.8$), with the remaining parameters taken to be similar to the values for the symmetric cases described above. We also considered the effect of taking $n = 8$ replicates as well as setting $n = 2$.

Unfortunately, with the low numbers of replicates considered here there is difficulty estimating most parameters under full maximum likelihood, especially for values of $\beta$ substantially different from 0. Even with REML estimation, typically only the parameters $\alpha$ and $\gamma$ are well estimated unless $\beta$ is close to 0. For some of the parameter sets we generated 100 replicates, and in these cases estimation appeared to be improved. Apparently large numbers of replicates are required for acceptable parameter estimates in this more complex mixture model.

4 Comparative study

4.1 Methods and statistics

A number of methods have been proposed for identifying differentially expressed genes, with new ones continuing to be introduced. We compare the performance of several of these after brief descriptions of those we consider.

In order to generalize and compare our work with that of Lönnstedt and Speed (2002), we include the statistics they used, both versions of our statistic, and another recently introduced method. These seven statistics and their justification are:
$L$-stat – Laplace mixture model

$AL$-stat – Asymmetric Laplace mixture model

$B$-stat – Normal mixture model (Lönnstedt and Speed; 2002)

$B1$-stat – Normal mixture (Gottardo et al.; 2003)

$M$-stat – Average log$_2$ fold change

$t$-stat – $t$-statistic based on individual gene standard deviation

Pen-$t$-stat – Penalised $t$-statistic (Efron et al.; 2000).

Note that when, as is the case here, all genes are equally replicated, the $B$-statistic reduces to the moderated-$t$ statistic (Smyth; 2004).

### 4.2 Simulation studies

Simulated datasets were generated under different mixture models. For each scenario we generated 100 datasets with 3000 genes each and $n = 2, 4, \text{ or } 8$ replicates, as in §3.2. Parameter values for the simulations were chosen based on our analysis of a publicly available *Arabidopsis thaliana* dataset, presented in §5.

The parameter corresponding to the proportion of differentially expressed genes ($\omega$ or $p$) is estimated from the data for the $L$- and $AL$-statistics, and set to 0.1 for the rest of the statistics.

Parameters for $L$-stat and $AL$-stat are estimated by the two-step procedure described in §3. The $B$-stat is computed using the eBayes function of the R package limma (Smyth; 2004). This package is available as part of the open source
BioConductor project (Gentleman et al.; 2004). We computed the $B_1$-stat with default values of the function $B1.bayes$ in the package and that works with older versions of R; similar functionality is now available through the BioConductor packages rama and bridge.

We generated data from the Laplace mixture model, as described in §3.2, with parameter values estimated by applying this model to the Arabidopsis dataset.

Table 2 gives the generating parameter values, with summaries of estimates from the $L$ and $AL$ models for sample sizes $n = 2, 4$ and 8.

We also simulated datasets under the Normal mixture model as for the Laplace mixture model. Because the findings are very close to those for data generated under the Laplace model, we do not show detailed results. The $B$-statistic and $L$-statistic perform quite similarly, although with data generated from the Normal mixture model the $B$-statistic performs slightly better.

4.3 Results

For each gene, the statistics are calculated for each of the 100 datasets. For a range of cutoff values for all the methods, the numbers of false positive and false negative results are calculated in each dataset. For each cutoff value of a given statistic, the observed numbers of false positive and false negative genes are then averaged over the hundred datasets. For each method, a sufficient range of cutoffs was chosen to be able to observe the behavior of the ROC curves over a large range. The average
Table 2: Parameter values for Laplace model simulations

The numbers of false positive genes are plotted against the average numbers of false negative genes for the statistics in variants of receiver operating characteristic (ROC) plots. Curves falling more steeply toward the lower left corner correspond to more powerful test procedures.
Figure 2: ROC curves for statistics computed on data simulated under the Laplace model, 3000 genes, \( n = 4, \omega = 0.1 \); (a) \( \alpha = 0.05, \gamma = 44, V = 1.25 \); (b) \( \alpha = 5.89, \gamma = 3.4, V = 3.58 \).

A few of these plots are presented in Figure 2. The \( AL \)-statistic performed practically identically to the \( L \)-statistic, so only the curve corresponding to the \( L \)-statistic is shown, along with those of the other six statistics. We summarize our general findings from examining several such curves.

Several of the methods perform broadly similarly across simulation conditions: \( L \)-stat, \( AL \)-stat, \( B \)-stat, \( B1 \)-stat and Pen-\( t \). In these examples, mean log\(_2\) fold change \( M \) also shows reasonably good performance as the great majority of gene-specific variances are small. Unsurprisingly, the \( t \)-statistic has poor performance for these sample sizes and cannot be recommended.
To reduce computing time in these simulations, we generated arrays consisting of a few thousand genes rather than the tens of thousands of genes more typically encountered in practice. We also examined our scenarios using 20,000 genes, varying \( \omega \) from 0.01 to 0.10, in order to assess the sensitivity method performance in this more realistic situation. Within a generating scenario, relative method performance was broadly similar across values of \( \omega \).

5 Example: *Arabidopsis thaliana* dataset

5.1 Data

We present an example using data publicly available from the Stanford Microarray Database (SMD) [http://genome-www5.stanford.edu/MicroArray/SMD/](http://genome-www5.stanford.edu/MicroArray/SMD/). The experiments were carried out to compare the general effect on disease resistance RNA transcript levels of *Arabidopsis thaliana* infected by rhizobacterium *Pseudomonas thivervalensis* (strain MLG45) to axenic control plants (Cartieaux et al.; 2003). Here we consider the experiment on plant leaves, which has slides containing 16,416 spots for four biological replicates hybridized to a common reference (SMD Experiment ID numbers 27084, 27000, 26995 and 26718). Further details are available from Cartieaux et al. (2003).

From the raw channel intensities, we computed print-tip loess normalized log ratios for each spot (Yang et al.; 2002b). To avoid inflated variance for low in-
tensity spots, we did not carry out any background correction; we also removed control spots. From these normalized log ratios, we estimated the parameters for the Laplace and Asymmetric Laplace models (Table 3). These values were used as a starting point for part of the simulation study of §4.

5.2 Model fit

Here we examine the fit of the Laplace mixture model to the data by comparing gene-specific means and variances of the true and simulated data.

The values of the log ratios seen in the simulated datasets tend to be larger than in the original data, although QQ plots comparing genewise data means to simulated means indicate a reasonably good fit (data not shown).

Under the proposed mixtures models, the marginal distribution of the sample variances follows an $F_{n-1,2\hat{\gamma}}$ distribution, where $n$ is the number of replicates. We can thus check the fit of the model to the dataset variances by plotting the sample variances against the quantiles of the $F_{n-1,2\hat{\gamma}}$ distribution. Figure 3(a) shows that

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\omega}$</th>
<th>SE</th>
<th>$\hat{V}$</th>
<th>SE</th>
<th>$\hat{\alpha}$</th>
<th>SE</th>
<th>$\hat{\gamma}$</th>
<th>SE</th>
<th>$\hat{\beta}$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace</td>
<td>.075</td>
<td>.007</td>
<td>3.58</td>
<td>.15</td>
<td>5.89</td>
<td>.204</td>
<td>3.39</td>
<td>.095</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Asymm. Laplace</td>
<td>.08</td>
<td>.006</td>
<td>3.58</td>
<td>.15</td>
<td>5.89</td>
<td>.204</td>
<td>3.39</td>
<td>.095</td>
<td>-0.03</td>
<td>.06</td>
</tr>
</tbody>
</table>

Table 3: Parameter estimates and standard errors for Laplace and Asymmetric Laplace models for the Arabidopsis thaliana dataset.
Figure 3: Variance fits: (a) QQ plot of sample variances; (b) square of the sample means ($\bar{y}_g^2$) versus log sample variances ($s_g^2$) – solid line for Laplace model, broken line for Normal model.

The agreement seems very good overall, with some deviation for the highest few variances.

To further investigate the fit of the Laplace mixture model, and compare it with that of the Normal mixture model (Lönnstedt and Speed; 2002), we examine the relationship between the sample mean $\bar{y}_g$ and the sample variance $s_g^2$.

Under the Laplace model, we find that

\[ E(\bar{y}_g^2) = \frac{1}{n} \sigma_g^2 + \frac{2\omega V}{n} \sigma_g^4, \quad E(s_g^2) = \sigma_g^2, \quad (5.1) \]
which is quadratic in $\sigma_g^2$, with coefficients that can be determined from $n$ and the maximum likelihood estimates of $\omega$ and $V$.

Under the normal mixture model (Lönnstedt and Speed; 2002), the mean $\mu_g$ for differentially expressed genes is distributed as Normal with mean zero and variance $\tau_g^2$, where $\tau_g^2 = \sigma_g^2 V$. Thus, the sample mean squared, $\bar{y}_g^2$, is linear in the sample variance $(s_g^2)$:

$$E(\bar{y}_g^2) = \frac{1 + \omega V}{n} \sigma_g^2, \quad E(s_g^2) = \sigma_g^2. \quad (5.2)$$

Here the coefficient can be determined by assuming $\omega = .01$, the default value for the B statistic, and with $V$ estimated by the method of moments (Lönnstedt and Speed; 2002).

We further investigate the relationship between the sample variances and the sample means by fitting generalised linear models. The models account for the squared means in terms of linear (Normal) or linear and quadratic (Laplace) functions of the variances, using the gamma family with identity link. The quadratic effect is significant at level around 0.001, suggesting a strong preference for the Laplace model over the Normal one.

6 Discussion

This study makes three main contributions to statistical methodology for analysis of microarray experiments. First, we have introduced the Laplace mixture model.
This model permits greater flexibility than models in current use as it has the potential, at least with sufficient data, to accommodate both whole genome and restricted coverage arrays. In addition, it also appears to provide a somewhat better fit to data. We have also proposed a fundamentally sound approach for estimating the proportion of differentially expressed genes that also appears to perform well in practice. Last, we have further characterized and compared several methods of identifying differential expression.

Our original motivation for proposing the Laplace mixture model was that the Normal mixture model of Lönnstedt and Speed (2002) did not fit our data well. We were initially surprised to find that this lack of fit only marginally affects performance: the Laplace model fits better but the performance of both is similar, so gene rankings from these models are anticipated to be alike across a variety of data sets. Thus, although there may be some room for ‘fine tuning’, we do not argue that the Laplace mixture should supplant the commonly used Normal mixture in practice. Rather, we believe that efforts aimed at further model refinement might more profitably be focused on solving new and more challenging problems.

Our proposed REML procedure for estimating the proportion of differentially expressed genes seems to perform quite well in the simulations. Unlike the heuristic procedure of Gottardo et al. (2003), our approach does not require an arbitrary criterion for calling a gene differentially expressed. This approach could
also be readily incorporated into existing Normal mixture model-based software. However, our approach might in some cases give poor results, in which case the obvious workaround is to simply fix the value of \( \omega \), as suggested by Lönnstedt and Speed (2002).

In the comparison study, several procedures—those which in essence use some type of global shrinkage to modify the denominator—appear to perform quite similarly across simulation conditions. Some forms of local shrinkage give worse performance than global shrinkage (Kooperberg et al.; 2005); we did not consider them here. These results underline the futility of endeavors to create substantially better procedures by further tinkering. The procedure that will be optimal in a given experiment depends on many factors which in practice are typically unknown, such as the form of the data generating distribution and the true parameter values, including the proportion of differentially expressed genes. The key aspect seems to be how the smallest and largest single gene variances are treated.

Finally, we would like to draw attention to the issue of software. Carrying out this type of study and, more importantly, data analyses in general, depends heavily on the availability of readily usable software. Here, we mostly relied upon the open source R-based BioConductor project (Gentleman et al.; 2004), which provides an ideal framework for software distribution as well as a large number of quality software tools for statistical genomics computing. We are in the process of preparing a more user-friendly version of our software as an R package for
BioConductor. We urge those who create software intended for wide dissemination also to consider implementation in R and contribution to BioConductor.

Acknowledgment

This work was funded by the Swiss National Science Foundation through the National Centre for Competence in Research in Plant Survival.
References


**URL**: http://genomebiology.com/2004/5/10/R80


List of Tables

1. Root mean squared error (bias) ($\times 10^2$) of maximum marginal likelihood (MML) and maximum likelihood (ML) estimators of the parameters of the symmetric Laplace mixture model. In all cases the proportion of differentially expressed genes is $\omega = 0.1$ and the number of replicates is $n = 2$. .......................... 14

2. Parameter values for Laplace model simulations .......................... 18

3. Parameter estimates and standard errors for Laplace and Asymmetric Laplace models for the Arabidopsis thaliana dataset. .............. 21

List of Figures

1. $M$ versus log variance plot of a simulated dataset. Highlighted points are truly differentially expressed genes, with mean $\mu_g$ simulated from a Laplace distribution. .......................... 13

2. ROC curves for statistics computed on data simulated under the Laplace model, 3000 genes, $n = 4$, $\omega = 0.1$; (a) $\alpha = 0.05$, $\gamma = 44$, $V = 1.25$; (b) $\alpha = 5.89$, $\gamma = 3.4$, $V = 3.58$. .......................... 19

3. Variance fits: (a) QQ plot of sample variances; (b) square of the sample means ($\bar{y}_g^2$) versus log sample variances ($s_g^2$) – solid line for Laplace model, broken line for Normal model. .......................... 22