Gene expression

Weighted analysis of microarray gene expression using maximum-likelihood

David J. Bakewell1,∗ and Ernst Wit2

1Cancer Research UK Beatson Laboratories, Garscube Estate, Bearsden, Glasgow G61 1BD, UK and 2Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK

Received on April 23, 2004; revised on September 17, 2004; accepted on September 20, 2004
Advance Access publication September 28, 2004

ABSTRACT

Motivation: The numerical values of gene expression measured using microarrays are usually presented to the biological end-user as summary statistics of spot pixel data, such as the spot mean, median and mode. Much of the subsequent data analysis reported in the literature, however, uses only one of these spot statistics. This results in sub-optimal estimates of gene expression levels and a need for improvement in quantitative spot variation surveillance.

Results: This paper develops a maximum-likelihood method for estimating gene expression using spot mean, variance and pixel number values available from typical microarray scanners. It employs a hierarchical model of variation between and within microarray spots. The hierarchical maximum-likelihood estimate (MLE) is shown to be a more efficient estimator of the mean than the ‘conventional’ estimate using solely the spot mean values (i.e. without spot variance data). Furthermore, under the assumptions of our model, the spot mean and spot variance are shown to be sufficient statistics that do not require the use of all pixel data.

The hierarchical MLE method is applied to data from both Monte Carlo (MC) simulations and a two-channel dye-swapped spotted microarray experiment. The MC simulations show that the hierarchical MLE method leads to improved detection of differential gene expression particularly when ‘outlier’ spots are present on the arrays. Compared with the conventional method, the MLE method applied to data from the microarray experiment leads to an increase in the number of differentially expressed genes detected for low cut-off P-values of interest.

Availability: The Matlab code is available at http://www.stats.gla.ac.uk/~microarray/software/

Contact: d.bakewell@beatson.gla.ac.uk

1 INTRODUCTION

Microarray spot summary statistics, such as mean, median and mode are commonly used as gene expression indicators from image processed microarray experiments. Ideally, all pixels of each spot located on each optically scanned microarray should be recorded to yield maximum available information but this would further exasperate existing data storage requirements and/or analysis capabilities. As such, the use of the abovementioned summary statistics is a trade-off between data storage capability and requirements of the end-user, and often one statistic is chosen by the analyst as the basis for quantifying gene expression. For example, the pixel median statistic for each spot is often chosen since it is resilient to a strongly skewed distribution of pixel intensities compared with, e.g. the mean. Recent work (Nuñez-Garcia et al., 2003) has shown that the choice of pixel mode statistic is a more accurate statistic for estimating gene expression for spots exhibiting a donut spot shape spatial pixel intensity distribution. Intuitively, the use of more than one summary statistic may lead to improvements in the estimation of gene expression and elucidation of the sources of data variability. For example, recent work has shown that using the mean, median and variance, combined together, can correct for signal saturation effects (Wit and McClure, 2003).

In this paper, we consider the use of the spot mean and standard deviation (SD) for estimating the mean expression across all spots of each gene under the same condition. These improved estimates will lead, for example, to a more accurate detection of differential expression. The ‘take-away’ message is to evaluate weighted spot means where the weights are roughly inversely proportional to the spot variance.

It is important to note that the technique developed is applicable to all types of microarrays and experimental designs for which a measure of variation is available. Thus, for example, although the microarray study in Section 5.4 uses a dye-swap, by no means is the technique confined to this particular type of experimental design using two channels. In fact, the technique can be applied to any type of experimental measurement (genomic, proteomic, etc.) for which the error variance is known.

2 MODEL AND METHODS

Quantitative evaluation of the expression of a single gene is achieved by optical measurement of spot fluorescence. Typically, for each of the spots the fluorescence of a large number of pixels is recorded. There are many sources of gene expression variation, both biological and technical (as discussed in Section 6), but a key characteristic of spotted microarrays is the physical, spatial separation of spots defined by printing. This motivates representation of gene expression variation on a microarray as a directed acyclic graph with two layers, as shown in Figure 1.

The hierarchical model pertains to a single gene with a mean expression value μ measured by n spots, which contain m independent pixels (i = 1,…,n). The between-spot and within-spot deviations are denoted by quantities σ² and ε, respectively. The top layer represents the variation between spot values. Spots can vary from each other, for example, because they represent cDNA from different biological samples. In this case, the top layer principally describes the biological variation of the expression of a
An example of within-spot pixel variation for two spots with the same log-expression, \( \xi \), but different within-spot pixel variation: \((\xi_{m_1} - \xi_{m_2}) \) poorer quality spot with high within-spot variation; \((\xi_{m_1} - \xi_{m_2}) \) higher quality spot with low within-spot variation: \((\xi_{m_1} + \xi_{m_2}) \) poorer quality spot with high within-spot variation; \((\xi_{m_1} + \xi_{m_2}) \) higher quality spot with low within-spot variation.

The log-mean of the \( i \)-th spot for the hierarchical model has the following distribution:

\[
x_{ij} = \mu + \xi_i + \varepsilon_{ij}
\]

and are shown as the second level of the hierarchy. In Equation (2), \( \varepsilon_{ij} \sim N(0, \sigma^2_b) \) represents the within-spot variation and \( \sigma^2_i \) is the variance within the \( i \)-th spot consisting of \( m_i \) independent pixels.

The pixel values are in principle available, but in practice one tends to prefer to work with summary parameters, such as, the log-pixel sample average and sample variance,

\[
x_i = \bar{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}
\]

and the \( m_i \) values. Note that since we have assumed the pixel intensities are log-normally distributed, \( x_{ij} \) is the logarithm of each pixel value and not the raw intensity. Therefore, currently, these \( x_{ij} \) values are not provided by most scanner imaging software. An estimation method is presented in Section 3.2 and the assumption of the log-normal pixel distribution is addressed in Section 6.

The log-mean of the \( i \)-th spot for the hierarchical model has the following distribution if the pixels are independent,

\[
x_i \sim N(\mu, \sigma^2_i + \sigma^2_b / m_i).
\]

In Section 2.3 and subsequent calculations, the sample variance \( s^2_i \) is used as a plug-in estimate of \( \sigma^2_i \) for pragmatic reasons (otherwise the problem becomes intractable). Thus, \( \sigma^2_i \) is given by Equations (3) and (4).

### 2.3 Maximum-likelihood estimate (MLE) of \( \mu \) and \( \sigma^2_b \) for the hierarchical model

Assuming \( \sigma_i \) and \( m_i \) are known, the log-likelihood \( l(\mu, \sigma_b) \) given \( x_i \) for \( n \) spots, \( i = 1, 2, \ldots, n \), is expressed as follows:

\[
l(\mu, \sigma_b) = \log p(x_1, \ldots, x_n | \mu, \sigma^2_b) = \log \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2_b}} \right) = -\frac{1}{2} \sum_{i=1}^{n} \left( \log(2\pi\Phi_i) + \frac{\sigma^2_i}{\sigma^2_b} \right).
\]
where $\Psi_1 = x_i - \mu$ and $\Phi_1 = \sigma_i^2 + \sigma^2/m_i$, and the last equality results from (5).

Expressions for the maximum log-likelihood are found by setting the partial derivatives of (7) to zero: $\partial_\mu l(\mu, \sigma_b) = \sum_{i=1}^n \Psi_1 / \Phi_1 = 0$ and $\partial_\sigma_b l(\mu, \sigma_b) = \sigma_b \sum_{i=1}^n (\Psi_1^2 / \Phi_1^2) = 0$. This results in

$$\mu_w = \frac{\sum_{i=1}^n x_i w_i^0}{\sum_{i=1}^n w_i^0},$$

(8)

where the coefficient, $w_i^0$, in the weighted sum is

$$w_i^0 = \frac{1}{\Phi_i} \sum_{k=1}^n \frac{1}{\Phi_k},$$

and

$$\sigma_{bw} = 0$$

(9a)

or the interesting case $\sum_{i=1}^n (\Psi_1^2 / \Phi_1^2) = 0$, leading to

$$\sigma_{bw} = \sigma_i^2 = \sigma_i^2 (x_i - \mu_w)^2 - \sigma^2/m_i,$$

(9b)

where the weight coefficient is

$$w_i^0 = \frac{1}{\Phi_i} \sum_{k=1}^n \frac{1}{\Phi_k}.$$

The subscript ‘w’ on the estimates of $\mu_w$ and $\sigma_{bw}^2$ signifies these parameters are weighted sums originating from Equation (7). The solutions depend implicitly on the unknown $\mu$ and $\sigma_b$. Either iteration or an appropriate computational algorithm will yield estimates for $\mu$ and $\sigma_b$. The coefficients weight the spots inversely to their variance, i.e. giving high weight to high-quality spots with low $\sigma_i^2/m_i$, and vice versa.

### 2.4 Naive gene expression estimates

The estimates for $\mu_n$ and $\sigma_{n}^2$ can be compared with estimates from a maximum-likelihood analysis that uses solely the spot mean $x_i$ data for the microarray scanner and does not include the within-spot, $\sigma_i^2/m_i$ data. The approach is called ‘naive’. Standard expressions for MLE unbiased estimates of the mean and variance of $x_i$ are as follows (Dougherty, 1990, p. 342):

$$\mu_n = \frac{1}{n} \sum_{i=1}^n x_i$$

(10)

and

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_n)^2,$$

(11)

where the subscript ‘n’ denotes that the estimates of $\mu$ and $\sigma_b^2$ are naive.

### 2.5 Comparison of hierarchical and naive MLE expressions for mean and variance

Rearranging Equation (9b)

$$\sigma_{bw}^2 = \sum_{i=1}^n w_i^0 \sigma_i^2 / m_i = \sum_{i=1}^n w_i^0 (x_i - \mu_w)^2$$

(12)

shows that the weighted sample variance consists of the between- and within-spot components.

A useful comparison of $\mu$ and $\sigma_b^2$ MLEs between the hierarchical and naive models is achieved by considering the special case for the hierarchical model where the spot variances are all equal in Equations (8) and (12), $\sigma_i^2/m_i = c_i$ for all $i = 1, \ldots, n \Rightarrow w_i^0 = 1/n$ and $w_i^0 = 1/n$ where $c_i$ is a constant. Consequently, the spots have equal weighting and Equations (8) and (12) simplify to

$$\mu_w = \frac{1}{n} \sum_{i=1}^n x_i$$

and

$$\sigma_{bw}^2 + c_i = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_w)^2,$$

where the subscript ‘w’ denotes that the estimates of $\mu$ and $\sigma_b^2$ assume equal weighting for each spot.

Comparing the hierarchical model where the within-spot variances are all equal to $c_i$ with the naive model [given by Equations (10) and (11)] shows that the estimates are the same for the mean, $\mu_n = \mu_w$, and differ by $c_i$ for the variance, $\sigma_n^2 + c_i = \sigma_{bw}^2$. This illustrates that the naive estimate of the spot mean variance $\sigma_n^2$ contains the within-spot and between-spot components. Furthermore, without including the use of the within-spot information $\sigma_i$ and $m_i$ (contained in this case within $c_i$), it would not be possible to estimate the between-spot variance, $\sigma_{bw}^2$.

### 3 ALGORITHM

This section describes the algorithms used to maximize the log-likelihood function, given by Equation (7), followed by a description of algorithms required for log-transforming microarray scanner data. We implement the $t$-test to determine the genes that are differentially, or non-differentially, expressed under a particular treatment (Chen et al., 1997; Yang et al., 2000; Kerr et al., 2000).

#### 3.1 Numerical considerations on finding a global maximum

The bounds of the maximum log-likelihood $l(\hat{\mu}, \hat{\sigma}_b)$ are found by applying Cauchy’s inequality to Equations (8) and (9b),

$$|\hat{\mu}_n| \leq \hat{\mu}_{\max}, \quad \hat{\mu}_{\max} = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$$

(13)

and

$$0 \leq \hat{\sigma}_{bw}^2 \leq \hat{\sigma}_{bw\max}^2.$$

(14)

In Equation (14) the lower and upper bounds arise from (9a) and (9b). The upper bound $\hat{\sigma}_{bw\max}^2$ is found by numerically maximizing the function

$$\sigma^2(\hat{\mu}_w) = \left(\sum_{i=1}^n (x_i - \hat{\mu}_w)^2 - \sigma_b^2/m_i\right)^{1/2}$$

with respect to $\hat{\mu}_w$ over the interval defined in Equation (13),

$$\hat{\sigma}_{bw\max}^2 = \max(\sigma^2(\hat{\mu}_w), |\hat{\mu}_n| \leq \hat{\mu}_{\max}).$$

The maximum-likelihood $l(\hat{\mu}_w, \hat{\sigma}_b)$ is rapidly evaluated by taking advantage of the properties of the double partial derivative with respect to $\mu$ that is negative for finite $\sigma_b^2 \geq 0$ and $\sigma_b^2/m_i > 0$,

$$\partial_{\mu,\sigma} l(\mu, \sigma_b) = -\sum_{i=1}^n \frac{1}{\Phi_i} \leq 0 \quad \forall \Phi_i > 0,$$

where it is understood $\mu \equiv \mu_w$ and $\sigma_b \equiv \sigma_{bw}$. This property ensures $l(\mu, \sigma_b)$ is a maximum with respect to $\mu$ wherever $\partial_{\mu} l(\mu, \sigma_b) = 0$. Substituting (8) into (7) simplifies the maximization of $l(\mu, \sigma_b)$ to
The procedure for finding a global maximum of the log-likelihood, 
\( l(\hat{\mu}_w, \hat{\sigma}_w) \), can be described as a two-step process. The first step is to find the approximate location of \( (\hat{\mu}_w, \hat{\sigma}_w) \) denoted as \( (\tilde{\mu}_w, \tilde{\sigma}_w) \). This is determined by maximizing Equation (15) with respect to \( \sigma_{bw} \) over the interval defined in (14)

\[
    l(\hat{\mu}_w, \hat{\sigma}_w) = \max \left\{ l(\sigma_{bw}), \ 0 \leq \sigma_{bw}^2 \leq \sigma_{bw}^2 \right\}.
\]

The second step for finding the location of the global maximum \( (\hat{\mu}_w, \hat{\sigma}_w) \) is to progressively refine (or ‘polish’) the starting point, \( (\tilde{\mu}_w, \tilde{\sigma}_w) \). However, for starting points arbitrary located (i.e., not \( (\tilde{\mu}_w, \tilde{\sigma}_w) \)) within the bounds given by Equations (13) and (14), the Nelder–Mead was found to be the most robust numerical optimization method.

### 3.2 Transformation of data to log-scale

Current scanners provide spot mean, \( y_i \), and variance, \( s_i^2 \), ‘raw’ summary statistical data that can be written as follows:

\[
    y_i = \frac{1}{m_i} \sum_{j=1}^{m_i} e^{y_{ij}},
\]

\[
    s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (e^{y_{ij}} - y_i)^2.
\]

The aim is to find estimates for the true spot mean and variance of the log pixel values \( \theta_i \) and \( \sigma_i^2 \) for each \( i \)-th spot.

These estimates can found using the Method of Moments (MoMs) (Dougherty, 1990, pp. 345–348). The ‘raw’ data \( y_{ij} = e^{y_{ij}} \) are assumed to have a true mean \( y_i \) and true variance \( \sigma_i^2 \),

\[
    y_i = \mathbb{E}(y_{ij}) = e^{\theta_i + \sigma_i^2/2},
\]

\[
    \sigma_i^2 = \mathbb{V}(y_{ij}) = e^{2\theta_i + \sigma_i^2} (e^{\sigma_i^2} - 1),
\]

where the second equality in both equations follows from standard formulae for log-normal mean and variance (Dougherty, 1990, p. 163). The MoM estimators \( \hat{\theta}_i \) and \( \hat{\sigma}_i^2 \) are defined as solutions of

\[
    y_i = e^{\hat{\theta}_i + \hat{\sigma}_i^2/2} \quad \text{and} \quad s_i^2 = e^{2\hat{\theta}_i + \hat{\sigma}_i^2} (e^{\hat{\sigma}_i^2} - 1).
\]

Hence, it can be easily deduced that

\[
    \hat{\theta}_i = \log_e \left( y_i^2 s_i^2 + \frac{1}{e^{y_i^2}} \right) \quad (16)
\]

and

\[
    \hat{\sigma}_i^2 = \log_e \left( s_i^2 y_i^2 + 1 \right). \quad (17)
\]

Note that in Section 2.2 the sample spot mean of the log-pixel values \( x_i \) is an estimate of \( \theta_i \). In Section 5.4, we set \( x_i = \hat{\theta}_i \) and \( \sigma_i^2 = \hat{\sigma}_i^2 \).

### 3.3 \( t \)-test

Typically the mean expression, \( \mu_t \), for a particular ‘treated’ gene is compared with the same parameter corresponding to a ‘control’ (or wild-type) gene, \( \mu_c \). The \( t \)-test is often used to test the null hypothesis that the means are the same, \( H_0: \mu_t = \mu_c \), or different, \( H_1: \mu_t \neq \mu_c \).

After log-transforming the data and applying the MLE algorithms from Sections 2 and 3.1, the \( t \)-test constitutes the final step for identifying differentially expressed genes. The test subsequently also offers a way of evaluating the performance of the naive and weighted \( \mu \) estimation methods as discussed later in Section 5.3.

The \( t \)-statistic for the difference between the treatment and control means for gene \( g \) estimated using the weighted method can be written as follows:

\[
    T_{w_g} = \frac{\mu_{w_g} - \mu_{c_g}}{\sqrt{\frac{\sum_{i=1}^{n} (\mu_{i_g} - \mu_{w_g})^2}{\sum_{i=1}^{n} \Phi_i^{-1}}}} \quad (18)
\]

where the subscripts ‘g’, ‘c’ and ‘t’ signify ‘gene identity’, ‘control’ and ‘treatment’, respectively. In Equation (18) the spots means are assumed to be independent so the variances can be evaluated,

\[
    \mu_{w_g} = \mu_{c_g} = \mu_{g_t} = \frac{\sum_{i=1}^{n} (\mu_{i_g})^2}{\sum_{i=1}^{n} \Phi_i^{-1}},
\]

\[
    \mu_{c_g} = \frac{\sum_{i=1}^{n} (\mu_{i_g})^2}{\sum_{i=1}^{n} \Phi_i^{-1}}, \quad \Phi_i = \frac{\sigma^2}{m_i} + \frac{\sigma}{m_i}
\]

Similarly, the \( t \)-statistic for the difference between the treatment and control means estimated using the naive method is,

\[
    T_{n_g} = \frac{\mu_{n_g} - \mu_{c_g}}{\sqrt{\frac{\sum_{i=1}^{n} (\mu_{i_g} - \mu_{n_g})^2}{\sum_{i=1}^{n} \Phi_i^{-1}}}} \quad (19)
\]

Each gene can then be classified as non-differentially, or differentially, expressed (NDE or DE) according to acceptance or rejection of the null hypothesis. For example, the two-sided \( t \)-test for the weighted method comprises

\[
    |T_{w_g}| \leq t^{-1}(1 - \alpha/2, v) \Rightarrow H_0 \quad \mu_t = \mu_c,
\]

\[
    |T_{w_g}| > t^{-1}(1 - \alpha/2, v) \Rightarrow H_1 \quad \mu_t \neq \mu_c,
\]

where \( t^{-1} \) denotes the inverse student-\( t \) function, \( \alpha \) is the level of confidence or cut-off \( P \)-value (typically 5%) and \( v = 2(n - 1) \) specifies the degrees-of-freedom.

### 4 IMPLEMENTATION

A Monte Carlo (MC) simulation of MLE was implemented using MATLAB 6© (Math Works, Inc., MA). The parameters specified were \( n, \sigma_b, \sigma, m_i \) as defined previously, and the number of trials. Spot log-mean values \( x_i \) were computed using a pair of normally distributed random number generators serially connected to achieve the hierarchical variation illustrated in Figure 1.

The quasi-Newton method routine in MATLAB 6© used a Davidson–Fletcher–Powell (DFP) algorithm (see documentation) but another option was the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method.
5 RESULTS

There are four results that stem from the MLE model presented in the preceding Sections 2–4. The first result is that the weighted estimate given by (8) is a more efficient estimate of \( \mu \) than (10). An example illustration is given. The second result shows the sufficiency of spot mean and variance statistics for this model. The third and fourth results show application of the weighted and naive methods to data from both MC simulations and a two-channel dye-swapped spotted microarray experiment. Comparisons of these simulations and microarray studies show the advantages of the weighted approach.

5.1 Relative efficiency

The relative efficiency for unbiased estimators can be expressed as the ratio of the variance of the weighted and naive estimates of \( \mu \) (Dougherty, 1990, p. 334; Casella and Berger, 2001, p. 476). Applying additivity properties for the variance of a sum of a Gaussian R.V. given in Equation (5) to \( \hat{\mu}_w \) and \( \hat{\mu}_n \) it can be shown using Cauchy’s inequality (Abramowitz and Stegun, 1965) and \( w_i^\mu \) given in Equation (8), that the relative efficiency is,

\[
\frac{V(\hat{\mu}_w)}{V(\hat{\mu}_n)} = \frac{\sum_{i=1}^{\infty}(w_i^\mu)^2 \phi_i}{\sum_{i=1}^{\infty}(w_i^\mu)^2} \leq 1, \quad (20)
\]

where the third equality results from \( w_i^\mu \) given in Equation (8). Excluding the trivial case where \( \hat{\mu}_w = \hat{\mu}_n \), Equation (20) shows \( \hat{\mu}_w \) is a more efficient estimator than \( \hat{\mu}_n \).

Figure 3 illustrates example distributions for estimates of the mean and SD using the weighted hierarchical MLE method and naive method, Equations (8) and (10). The data were generated using a MC simulation with 1000 trials and parameter values \( \mu = 10 \), \( \sigma_0 = 1 \), \( m_i = 40 \) \( \forall i = 1, \ldots, 10 \) and \( \sigma_i = 2 \), \( 2 \), \( 4 \), \ldots, \( 20 \). The histograms indicate that the means of the distributions are close to the correct value \( \mu = 10 \) and they are the same, \( E(\hat{\mu}_w) \approx E(\hat{\mu}_n) \), whereas the spread of \( \hat{\mu}_w \) is less than \( \hat{\mu}_n \). The ratio of the SD was \( \sqrt{V(\hat{\mu}_w)/V(\hat{\mu}_n)} \approx 0.8 \) (\( \bar{V} \) denotes sample variance).

5.2 Sufficiency

The spot means and variances can be shown to be sufficient, that is, if the spot means and variances are known, then the pixels do not give any additional information for this model. This can be easily shown as follows. The likelihood using pixel values is given by

\[
L_{\psi_i} = p(x_{ij}|\mu, \sigma^2_i) = \prod_{i=1}^{n} p(x_{ij}|\theta_i, \mu_i, \sigma^2_i) = \prod_{i=1}^{n} \left( \frac{\theta_i - \mu}{\sigma_i} \right)^{n_i} \prod_{i=1}^{n} \left( \frac{\theta_i - \theta_i}{\sigma_i} \right)^{\frac{n_i}{2}} \times \theta_i^{n_i} \phi_i \left( \frac{x_{ij} - \theta_i}{\sigma_i} \right) \phi_i \left( \frac{\theta_i - \mu}{\sigma_i} \right), \quad (21)
\]

where the vectors \( x_{ij} = x_{i1}, x_{i2}, \ldots, x_{i12}, x_{i22}, \ldots, x_{i10} \) and \( \theta_i = \theta_1, \ldots, \theta_{10} \).

After some algebra, the second integrand in Equation (21), \( I_{\psi_i} \), can be evaluated and re-expressed

\[
I_{\psi_i} = \int \sqrt{m_i} \phi_i \left( \frac{x_{ij} - \theta_i}{\sigma_i} \right) \phi_i \left( \frac{\theta_i - \mu}{\sigma_i} \right) \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2\sigma_i^2}(x_{ij} - \theta_i)^2} \phi_i \left( \frac{\theta_i - \mu}{\sigma_i} \right)^{n_i} \phi_i \left( \frac{x_{ij} - \theta_i}{\sigma_i} \right) \phi_i \left( \frac{\theta_i - \mu}{\sigma_i} \right),
\]

where

where \( C_i \) is independent of \( \mu, \sigma_0 \) and \( \theta_i \) and can be considered to be a constant. Hence,

\[
L_{\psi_i} = \prod_{i=1}^{n} C_i \int \phi \left( \frac{\theta_i - \mu}{\sigma_0} \right) \phi \left( \frac{x_{ij} - \theta_i}{\sigma_i} \right) \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2\sigma_i^2}(x_{ij} - \theta_i)^2} \phi \left( \frac{\theta_i - \mu}{\sigma_i} \right) d\theta_i = C \prod_{i=1}^{n} C_i \int \phi \left( \frac{x_{ij} - \theta_i}{\sigma_i} \right) \phi \left( \frac{\theta_i - \mu}{\sigma_i} \right) d\theta_i = C \phi \left( \frac{x_{ij} - \theta_i}{\sigma_i} \right) \phi \left( \frac{\theta_i - \mu}{\sigma_i} \right).
\]

where \( L_{\psi_i} \) is the likelihood using the spot means and variance summary statistics, and \( C = \prod_{i=1}^{n} C_i \) simply acts as a constant of proportionality. Hence, it follows \( \log(\phi(\psi_{ij} | \mu, \sigma)) = l_{\psi_i}(\mu, \sigma) + \text{constant} \) and the maximum log-likelihood using pixels leads to the same \( \hat{\mu}_w \) and \( \hat{\sigma}_w^2 \) as the maximum log-likelihood using summary statistics.

5.3 Simulation study

The performance of the MLE weighted and naive methods for estimating gene expression were compared using Monte Carlo simulated gene expression data. A set of 400 genes was split into one-half NDE: with gene identities \( g = 1, \ldots, 200 \) with \( \mu = 1, \mu = 1 \), and second-half DE: \( g = 201, \ldots, 400 \) \( \mu = 3, \mu = 3 \). The between-spot SD was held constant \( \sigma_0 = 1 \) for all genes for both control and treatment categories and the number of microarray spots (for each gene) likewise remained constant at \( n = 5 \).

Two microarray experiments were simulated with the number of independent pixels per spot, \( m_i = 4 \), and within-spot SD \( \sigma_i \) applied to both control and treatment categories as follows:

\[
(1) \text{ Constant variance (CV) with } \sigma_i = 4 \text{ for all spots } i = 1, \ldots, 5 \text{ and gene identities } g = 1, \ldots, 400, \text{ that is } \sigma_{ij=1...5} = 4, 4, 4, 4, 4, \quad g = 1, \ldots, 400.
\]

Fig. 3. Distributions for estimates of the mean using the weighted and naive methods, \( \hat{\mu}_w \) and \( \hat{\mu}_n \). The histograms were generated using a MC simulation with 1000 trials and parameter values \( n = 10 \) spots, gene expression log-mean \( \mu = 10 \) and variance \( \sigma_0 = 1, m_i = 40 \) independent pixels per spot and spot variance \( \sigma_i = 2 \). See text for details.
(1) constant within-spot variance (CV) and (2) constant variance with the exception of outliers with high variance (CV-O). See text for details.

(2) Constant variance with the exception of occasional high variance spots, or outliers (CV-O),

\[
\sigma_{i=1,...,5} = [1, 1, 1, 1, g], \quad g = 1, \ldots, 100.
\]

\[
\sigma_{i=1,...,5} = [1, 1, 1, 1, 20], \quad g = 101, \ldots, 200.
\]

\[
\sigma_{i=1,...,5} = [1, 1, 1, 1, 1], \quad g = 201, \ldots, 300.
\]

\[
\sigma_{i=1,...,5} = [1, 1, 1, 1, 20], \quad g = 301, \ldots, 400.
\]

\[t\text{-values evaluated using (18) and (19) for each gene were re-ordered to generate lists of genes identified as either true positive, or false positive, for a particular cut-off } P\text{-value. Combining the counts of true positive, or false positive, genes for a range of } P\text{-value cut-offs forms a Receiver Operating Characteristic (ROC)—well-known as a performance indicator in digital communications (Van Trees, 1968, pp. 36–46).}

Figure 4 shows two sets of ROCs that enable a direct comparison of the weighted and naive estimation methods for each of the two simulated microarray experiments. ROCs that follow the left-hand and top borders show a more accurate test than those that follow a 45° diagonal. Figure 4 shows that in both microarray simulations the weighted method for estimating gene expression performs better than the naive method, in particular, for the case of constant variance with outliers.

5.4 A microarray study

The MLE weighted and naive methods for estimating gene expression were applied to a two-channel dye-swapped spotted microarray dataset (four separate technical replicates) used for studying skin cancer (Wit and McClure, 2003, 2004). To enable a fair comparison of the estimation method, the dataset excluded genes exhibiting extreme expression levels (very low or very high) indicative of unwanted technical effects, such as, fluorescence saturation.

The spot mean, \(y_{ig}\) (spot \(i\), gene \(g\)), and variance, \(s_{ig}^2\), data for both treatment genes (extracted from cancerous fibroblast cells) and control (normal) genes, were initially log-transformed using MoM estimates provided by the scanner software. Estimates of \(\mu_{uw}\) and \(\sigma_w^2\) for treatment and normal genes, and corresponding \(t\)-statistics were evaluated as described in Sections 2, 3 and 4. Figure 5 shows the proportion of genes (out of 928) where the null hypothesis is rejected (i.e. accept differential expression hypothesis, \(H_1\)) versus the cut-off level over the range of typical interest, \(0 \leq \alpha \leq 0.2\). The plots for this dataset show that the weighted method detects a higher proportion of differentially expressed genes compared with the naive method. Both the weighted and naive procedures can be compared with the plot of proportion of genes expected when differential expression does not occur.

6 DISCUSSION AND CONCLUDING REMARKS

The preceding sections have shown that the MLE hierarchical model leads to a more efficient estimate of gene expression and subsequent improvement in differential detection. There are, however, areas where improvements can be made. One example concerns MoM estimation (prior to use in the MLE model) of the spot mean and variance of the log-pixel values from the corresponding ‘raw’ quantities provided by the scanner software (i.e. \(x_i\) and \(\sigma_i^2\) from \(y_i\) and \(s_{ig}^2\)). MC simulations of differential expression using the MoM estimates indicate that there would be a significant benefit in differential detection if the scanner software provided the spot mean and variance of the log-pixel values thereby avoiding the need for MoM estimates.

We also remark that the pixel distributions are not well represented in the literature compared with spot distributions, and that our use of
a log-normal pixel distribution approximates the normal distribution using a square-root transformation (Glasbey and Ghazal, 2003). Implementing a square-root normal distribution would only require modifications to the MoM relations in Section 3.2.

In general, we note that the hierarchical model (Fig. 1) representing variations between-spots and within-spots may, possibly be generalized to investigate ‘biological’ and ‘technical’ variation. Biological variation arises from a number of factors in the process of gene expression, including, for example, spatial inhomogeneity of cells in the growth medium, poor cell-cycle synchronization, variations in DNA transcription, etc. Technical variation, in contrast, is attributed to uncertainties in the microarray, sample preparation and measurement process. Spotted microarrays, for example, exhibit technical variation arising from the DNA hybridization and labeling of the fluorescent dyes, spot printing effects, spatial variations in coatings of the slides, etc.

Considering all the available information, we can improve the quality of microarray gene expression estimation. Algorithms based on MLE are used to weight the measured mean of each spot inversely to the pixel variance measured on a gene-by-gene basis and we have shown the resulting improvements on the estimation of microarray expression values.

ACKNOWLEDGEMENTS

The authors thank Dr N. Barr for kindly making available the skin cancer data. D.J.B. would like to thank staff at the Bacterial Microarray Group, St George’s Hospital Medical School, London, and Dr K. Vass at the Beatson Laboratories (CR UK), for their guidance and encouragement, and providing office and computing facilities. E.W. would like to thank the Dipartimento di Scienze Statistiche ‘Paolo Fortunati’ of the University of Bologna for its hospitality from January until July 2004. D.J.B. would also like to thank The Wellcome Trust for financial support (Project 062511).

REFERENCES