Gene expression

Segmentation and intensity estimation of microarray images using a gamma-t mixture model

Jangsun Baek\(^1\),*, Young Sook Son\(^1\) and Geoffrey J. McLachlan\(^2\)

\(^1\)Department of Statistics, Chonnam National University, Gwangju 500-757, South Korea and \(^2\)Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia

Received on September 20, 2006; revised on November 30, 2006; accepted on December 6, 2006
Advance Access publication December 12, 2006
Associate Editor: Satoru Miyano

ABSTRACT

Motivation: We present a new approach to the analysis of images for complementary DNA microarray experiments. The image segmentation and intensity estimation are performed simultaneously by adopting a two-component mixture model. One component of this mixture corresponds to the distribution of the background intensity, while the other corresponds to the distribution of the foreground intensity. The intensity measurement is a bivariate vector consisting of red and green intensities. The background intensity component is modeled by the bivariate gamma distribution, whose marginal densities for the red and green intensities are independent three-parameter gamma distributions with different parameters. The foreground intensity component is taken to be the bivariate t distribution, with the constraint that the mean of the foreground is greater than that of the background for each of the two colors. The degrees of freedom of this t distribution are inferred from the data but they could be specified in advance to reduce the computation time. Also, the covariance matrix is not restricted to being diagonal and so it allows for nonzero correlation between R and G foreground intensities. This gamma-t mixture model is fitted by maximum likelihood via the EM algorithm. A final step is executed whereby nonparametric (kernel) smoothing is undertaken of the posterior probabilities of component membership.

The main advantages of this approach are: (1) it enjoys the well-known strengths of a mixture model, namely flexibility and adaptability to the data; (2) it considers the segmentation and intensity simultaneously and not separately as in commonly used existing software, and it also works with the red and green intensities in a bivariate framework as opposed to their separate estimation via univariate methods; (3) the use of the three-parameter gamma distribution for the background red and green intensities provides a much better fit than the normal (log normal) or t distributions; (4) the use of the bivariate t distribution for the foreground intensity provides a model that is less sensitive to extreme observations; (5) as a consequence of the aforementioned properties, it allows segmentation to be undertaken for a wide range of spot shapes, including doughnut, sickle shape and artifacts.

Results: We apply our method for gridding, segmentation and estimation to cDNA microarray real images and artificial data. Our method provides better segmentation results in spot shapes as well as intensity estimation than Spot and spotSegmentation R language softwares. It detected blank spots as well as bright artifact for the real data, and estimated spot intensities with high-accuracy for the synthetic data.

Availability: The algorithms were implemented in Matlab. The Matlab codes implementing both the gridding and segmentation/estimation are available upon request.

Contact: jbaek@chonnam.ac.kr

Supplementary information: Supplementary material is available at Bioinformatics online.

1 INTRODUCTION

Microarray technology, such as complementary DNA (cDNA) arrays provides a means for measuring the expression levels of thousands of genes simultaneously. cDNA arrays are mostly used in 2-channel experiments. In such an experiment two different cDNA probes are prepared independently for the analysis; the first probe is prepared from control mRNA and labeled with Cy3 (green) dye, and the second from mRNA isolated from the treated cells or tissue under investigation is labeled with Cy5 (red) dye. Both probes are hybridized together on the same microarray slide. Then the fluorescent signal from Cy3 and Cy5 labels is scanned to provide a 16-bit gray-scale image, respectively. The relative intensities of the dyes in individual spots measure the relative amounts of specific transcripts in each sample, meaning differential gene expression can be examined.

The image analysis for cDNA arrays consists of three steps: (1) the addressing step aligns the spots on the array; (2) the segmentation step partitions the set of pixels into foreground and background; and (3) the intensity estimation step extracts the red (R) and green (G) intensities and assigns the log ratio (log\(_2\) R/G) after background correction to represent the (log) relative abundance of each spot.

In this paper, we use a simple addressing procedure similar to the one in Li et al. (2005) and mainly focus on segmentation and intensity estimation. There are many segmentation methods for microarrays. The existing methods can be grouped into two categories according to whether a parametric distribution of the pixel intensity is assumed or not: parametric segmentation and nonparametric segmentation (see Yang et al., 2002 for other categorizations.)

Nonparametric segmentation methods do not assume any particular type of distribution on intensities, but use only the descriptive statistics or some information of the histogram, which is a nonparametric estimator for the true distribution of the pixel intensity. Fixed circle segmentation (Eisen, 1999) and adaptive circle segmentation (Axon Instruments Inc., 2003, http://www.axon.com) use a circle of fixed and adaptive radius for fitting foreground, respectively. Jain et al. (2002) use histogram information within a circle.

\(^*\)To whom correspondence should be addressed.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
The seeded region growing method (Adams and Bischof, 1994) selects the pixel whose intensity is nearest to the average of the intensities in the neighboring foreground or background. Chen et al. (1997) used a threshold value based on Mann–Whitney test and GSI Lumonics (1999, http://las.perkinelmer.com/), defined foreground and background as the mean intensities between some predefined percentile values. All these methods do not correctly segment doughnut-shaped spots.

In parametric segmentation, the distribution for intensity is specified up to a vector of unknown parameters. Glassbey and Ghazal (2003) assumed the bivariate normal distribution for the red and green intensities of foreground and background in one of their methods. Steinfath et al. (2001) fitted a scaled bivariate normal distribution to pixel values, and Brändle et al. (2003) used an M-estimator for fitting the normal spot model. Li et al. (2005) described a model-based segmentation method in which the normal distribution for the combined intensity (R + G) is implicitly assumed. Demirkaya et al. (2005) fitted the exponential distribution to both foreground and background intensities and used a Markov random field model for segmentation. Gottardo et al. (2006) also applied a Markov random field approach to segmentation, but assumed an uncorrelated bivariate t distribution as the joint distribution of the red and green intensities for both foreground and background.

The correct assumption on the underlying density of pixel intensities is crucial for segmentation and mean intensity estimation in parametric methods. The distribution of background intensities is usually skewed in various degrees while the foreground intensities are scattered elliptically. Thus it would be more appropriate to assume a general distribution that can be fitted well to as many real microarray images as possible. We propose the bivariate gamma and t as the underlying distribution for background and foreground, respectively, both of which are quite general and reduce to other distributions in special cases.

In Section 2, we describe our image segmentation and intensity estimation method, including automatic gridding, justification of the proposed pixel intensity distributions, and mixture model-based segmentation and intensity estimation. In Section 3, we demonstrate the stable parameter estimation result of the proposed method using synthetic data, and present the segmentation and estimation results from applying our method to two real experimental microarray image datasets. We compare the results with those from other methods. Finally, we summarize the results in Section 4.

2 METHODS

We have implemented the following for microarray image analysis: automatic gridding, model-based clustering of pixels and intensity estimation. The original image data consist of a pair of unsigned 16-bit images, which are stored as TIFF files. One corresponds to the dye Cy5 and another to Cy3. We transformed these original data by taking square roots or logarithms for gridding and segmentation/estimation. The transformed images (R for Cy5 and G for Cy3) have two advantages compared to the original 16-bit images. Very high pixel intensities are damped in the combined image on which gridding is performed based, to prevent very bright pixels from dominating in both gridding and segmentation steps, and the transformed lower-bit images are computationally efficient (Yang et al., 2002).

2.1 Automatic gridding

A microarray slide consists of several blocks containing hundreds of spots. The gridding procedure is to identify blocks and to position rows and columns of spots within each block. Since the blocks are usually separated clearly on a slide, we can take out each block image and analyze them separately. We use a step similar to that of Li et al. (2005). Our algorithm starts by obtaining the combined image and projecting the intensities by summing up across the pixels in each row and each column. We smooth these projections using robust loess (Cleveland and Devlin, 1988) with the bandwidth approximately equal to the width of a typical spot. Then the smoothed projections form a series of peaks and valleys. The grid is defined by drawing a line in each valley. The two figures in Supplementary material show the results from applying the method to a block containing 18 × 24 spots in an example image at CSIRO (http://spot.cmis.csiro.au/spot/demodownload/SpotExSamples.zip). The first figure shows the projections along with the smoothed projections; the valleys correspond to the grid lines.

The second figure shows the resulting grids positioning the spots well.

We propose the bivariate t distribution as a special case, and the intensities are concentrated at low values and gradually become sporadic at high values on both axes for the background. For the foreground, on the other hand, the intensities are scattered elliptically. We fitted the distribution of the proposed mixture model with the estimates obtained by our method to each clustered dye intensity. We modeled the R and G intensities of the background by the gamma distribution, and plotted the fitted p.d.f. with the histogram in (e) and (f). The background distributions are different in two channels. The distribution of the background R looks like exponential which is a special case of the gamma distribution, and the background G follows the gamma distribution as well. The R and G intensities of the foreground were modeled by the t distribution, and plotted in (g) and (h), respectively. The goodness of fit was significant for each distribution at the significance level of 0.05.

We propose the bivariate distribution as the underlying pixel intensity distribution for the background, whose marginal densities for the R and G are independent three parameter gamma p.d.f.’s. The family of gamma distributions is supported on the positive line and provides a rich class of distributions as special cases. It is an exponential distribution if the shape parameter is equal to 1, contains a χ² distribution as a special case, and converges to a normal distribution as the shape parameter gets large (Johnson and Kotz, 1970). For the underlying distribution for the foreground, we adopt the bivariate t distribution. Although the normal distribution is usually assumed for elliptical data, the tails of the normal distribution are often shorter than required. Figure 1d shows the pixel segmentation result.
when the normal distribution is assumed for the foreground pixels in the mixture model. Some of the pixels with high intensities are classified as the background because the tails of the normal distribution are not long enough to contain them as their population elements. Also, the parameter estimates of the model being fitted can be affected by atypical observations. The $t$ distribution provides a more robust approach to the fitting of normal mixture distributions. The mean correlation coefficient between the foreground and background intensities, respectively, than the exponential or normal distributions. The mean correlation coefficient between the exponential or normal distribution for the background intensities (red and green) and for the $t$ distribution very well (the $t$ distribution is better). It could be argued that the superior results for the gamma over the exponential or normal distribution for the foreground intensities (red and green) and for the $t$ distribution over the latter two distributions for the foreground intensities may be due to the adoption of the former distributions to first classify the pixels as background or foreground. We therefore recalculated the average $P$-values for these goodness-of-fit tests after the pixels were segmented according to the method of Li et al. (2005), which adopts the normal distribution for both background and foreground intensities. The $P$-values so obtained are listed in parentheses in Table 1. It can be seen of comparing the mean $P$-values in parentheses that the gamma and $t$ distributions provide a better fit to the background and foreground intensities, respectively, than the exponential or normal distributions. The mean correlation coefficient between $R_b$ and $G_b$, in the background for all spot rectangles is 0.0262, while the mean $P$-value for the null hypothesis that there is no correlation between $R_b$ and $G_b$ is 0.3929. It suggests that it is reasonable to take $R_b$ and $G_b$ to be uncorrelated.

In our mixture model for segmentation, we assume the two parameter gamma distribution with known location parameters $\gamma_1$ and $\gamma_2$ for $y' = (y_1 = R_b, y_2 = G_b)$ in the background, which has the probability density function (p.d.f.)

$$f_1(y; \Theta) = \frac{2^\alpha \Gamma(\alpha)}{\Gamma(\alpha/2)^2} \left( \frac{\mu}{\gamma_1 - y_1} \right)^{\alpha - 1} \exp\left[ -\beta \frac{(\gamma_1 - y_1)}{\mu} \right],$$

where $\Theta = (\alpha, \beta, \gamma_1, \gamma_2, \alpha > 0, \beta > 0, \gamma_1, \gamma_2 > 0, i = 1, 2$. In practice, the location parameter $\gamma_1$ is estimated from the data. We use the estimate $\hat{\gamma}_1 = \min(y_1, \ldots, y_n) - \epsilon$ for small positive number $\epsilon$, $i = 1, 2$. For the foreground pixel intensity distribution, we assume the bivariate $t$ distribution whose p.d.f. is

$$f_2(y; \mu, \Sigma, v) = \frac{1}{\pi \nu (1 + \beta (y - \mu)^\Sigma^{-1} (y - \mu))^{1/2}},$$

where $\beta(y, \mu, \Sigma) = (y - \mu)^\Sigma^{-1} (y - \mu)$, $\mu = (\mu_1, \mu_2)$ and $\mu_1 = \alpha_1 / \beta_1 + \gamma_1 + \phi_1$, $\mu_2 = \alpha_2 / \beta_2 + \gamma_2 + \phi_2$, $\phi_1 \geq 0$, $\phi_2 \geq 0$. Since the mean of the foreground intensity must be larger than that of the background, the location parameter $\mu$ is parameterized as the mean of the background $(\alpha_1 / \beta_1 + \gamma_1)$ plus the nonnegative parameter $\phi_1$. Thus the background corrected spot intensity is measured by $\phi_1$ for each channel.

### 2.3 Image segmentation and intensity estimation

Let $y_1, \ldots, y_n$ denote the bivariate pixel intensities to be segmented. We will assume that the observed intensities are independent and identically

<table>
<thead>
<tr>
<th></th>
<th>$R_b$</th>
<th>$G_b$</th>
<th>$R_f$</th>
<th>$G_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>0.2264</td>
<td>0.0119</td>
<td>0.0232</td>
<td>0.0279</td>
</tr>
<tr>
<td>Normal</td>
<td>0.0012</td>
<td>0.0655</td>
<td>0.6913</td>
<td>0.5888</td>
</tr>
<tr>
<td>Gamma/ $t$</td>
<td>0.2325</td>
<td>0.2144</td>
<td>0.7726</td>
<td>0.7269</td>
</tr>
</tbody>
</table>

$R_b, R_f$: red intensity in background and foreground; $G_b, G_f$: green intensity in background and foreground. Mean $P$-values for the segmented data by the method of Li et al. (2005) are in the parentheses.

Fig. 1. (a) original image; (b) segmented image; (c and d): scatter plot of the bivariate pixel intensities segmented by the gamma-$t$, gamma-normal mixture model; (e and f): histogram and gamma dist. fit of $R, G$ in background; (g and h): histogram and $t$ dist. fit of $R, G$ in foreground.
distributed realizations from the density
\[ f(y; \Psi) = \pi_i f_i(y; \Theta) + \pi_y f_2(y; \mu, \Sigma, \nu), \]
where \( \Psi = \{ \Theta, \mu, \Sigma, \nu, \pi_1 \}, \pi_i (0 \leq \pi_i, \pi_1 \leq 1, \pi_1 + \pi_2 = 1) \) is the probability that a pixel belongs to the \( i \)th component, and \( f_1 \) and \( f_2 \) are the densities defined by Equations (1) and (2), respectively. We will implement the EM algorithm to obtain the maximum likelihood estimates \( (m.l.e.) \) of the parameters in \( \Psi \). In the EM algorithm framework, the observed data are viewed as being incomplete, as the associated component-label vectors \( x_1, \ldots, x_n \) are not available; \( x_i = (z_{ij}, y_{ij}) \) is defined by \( x_i = 1 \) if \( y_i \) belongs to group 1 (foreground), \( z_{ij} = 0 \) otherwise, and \( z_{ij} = 1 \) if \( y_i \) belongs to group 2 (foreground), \( z_{ij} = 0 \) otherwise. The complete-data vector is defined to be \( \mathbf{y}_i = (y_i^1, z_{ij}) \), where \( y_i = (y_{i1}, \ldots, y_{iy}) \). It is appropriate to assume that the \( x_i \) are distributed according to the multinomial distribution with probabilities \( \pi_1 \) and \( \pi_2 \). Then the complete-data likelihood \( L_c(\Psi) = L_c(\Psi) = \prod_{i=1}^{n} \prod_{j=1}^{I} \left\{ \pi_i f_i(y_{ij}) \right\}^{z_{ij}}, \) and its log likelihood is
\[ \log L_c(\Psi) = \sum_{i=1}^{n} \sum_{j=1}^{I} z_{ij} \left\{ \log \pi_i + \log f_i(y_{ij}) \right\}. \]  

**E-step.** The missing data are the \( x_i \) \( (j = 1, \ldots, n) \), which represent the component (segment) labels. Given the \( \Psi^{(k)} \) of \( \Psi \) after the \( k \)th iteration, the E-step on the \( (k+1) \)th iteration of the EM algorithm involves replacing \( x_i \) by \( E_{\Psi^{(k)}}(Z_i | Y_i) \), its conditional expectation given \( Y_i \), using \( \Psi^{(k)} \) for \( \Psi \). It can be expressed as
\[ \tau^{(k)}_{ij} = \pi_i^{(k)} f_i(y_i; \Psi^{(k)}) / f(y_i; \Psi^{(k)}), \] which is the posterior probability that \( y_i \) belongs to the \( i \)th component of the mixture using the current fit \( \Psi^{(k)} \) for \( \Psi \) \( (i = 1, 2; j = 1, \ldots, n) \).

The multivariate \( r \) distribution can be viewed as a weighted average multivariate normal distribution with the weight gamma distribution of random variable \( U \) (McLachlan and Peel, 2000). The complete-data likelihood \( L_c(\Psi) \) can be factored into the product of the marginal densities of the \( Z_i \), the conditional densities of the \( U_i \) given the \( z_i \), and the conditional densities of the \( Y_i \) given the \( u_i \) and the \( x_i \). Then \( Q(\Psi; \Psi^{(k)}) \), the current conditional expectation of the complete-data log likelihood function, \( \log L_c(\Psi) \) plugged with \( \tau^{(k)}_{ij} \), using the current value \( \Psi^{(k)} \) for \( \Psi \) is given by (see Supplementary material for the derivation)
\[ Q(\Psi; \Psi^{(k)}) = Q_1(\pi; \Psi^{(k)}) + Q_2(\nu; \Psi^{(k)}) + Q_3(\xi; \Psi^{(k)}), \]
where \( \xi = \{ \alpha_1, \beta_1, \alpha_2, \beta_2, \mu \} \) and
\[ Q_1(\pi; \Psi^{(k)}) = \sum_{i=1}^{2} \sum_{j=1}^{I} \tau^{(k)}_{ij} \log \pi_i, \]
\[ Q_2(\nu; \Psi^{(k)}) = \sum_{j=1}^{I} \tau^{(k)}_{ij} Q_{1j}(\xi; \Psi^{(k)}), \]
\[ Q_3(\xi; \Psi^{(k)}) = \sum_{j=1}^{I} \tau^{(k)}_{ij} Q_{1j}(\xi; \Psi^{(k)}) + \sum_{j=1}^{I} \tau^{(k)}_{ij} Q_{2j}(\xi; \Psi^{(k)}), \] where, on ignoring terms not involving the \( \nu \),
\[ Q_{1j}(\xi; \Psi^{(k)}) = -\log \Gamma(\nu(2) + 1) + 1/2 \nu \log \Gamma(\nu(2)) \]
\[ + 1/2 \nu \left\{ \log \left( \nu(2) + 1 \right) + \psi(\nu(2) + 1) \right\} - (\nu(2) + 1/2) \log \Gamma(\nu(2)) + \log \Gamma(\nu(2) - 1/2), \]
\[ Q_{2j}(\xi; \Psi^{(k)}) = \alpha_1 \log \beta_1 + \alpha_2 \log \beta_2 + (\alpha_1 - 1) \log (y_{ij} - \gamma_1) + (\alpha_2 - 1) \log (y_{ij} - \gamma_2) - \log \Gamma(\alpha_1) - \log \Gamma(\alpha_2) + \beta_1 \log (y_{ij} - \gamma_1) - \beta_2 \log (y_{ij} - \gamma_2), \]
\[ Q_{3j}(\xi; \Psi^{(k)}) = -\log(2\pi) - 1/2 \log | \Sigma | + \log \nu(2) + \psi(\nu(2) + 1/2) - (1/2) \nu(2) \left( y_{ij} - \mu \right) \Sigma^{-1} \left( y_{ij} - \mu \right), \]
\[ M-step. \] \( Q(\Psi; \Psi^{(k)}) \) is then maximized as a function of \( \Psi \). The estimate for \( \pi_i^{(k)} \) exists in closed form by maximizing Equation (5):
\[ \pi_i^{(k)} = \frac{\sum_{j=1}^{I} \tau_{ij}^{(k)}}{n}, \quad i = 1, 2. \]

To obtain the estimates of \( \alpha \), \( \beta \), \( \alpha_2 \), \( \beta_2 \), \( \mu \) and \( \Sigma \), we need to take the derivative of \( Q(\xi; \Psi^{(k)}) \) with respect to \( \xi \), the \( 1 \)th parameter in \( \xi \), and set it equal to zero. Set \( \gamma = y_i \) in \( Q(\xi; \Psi^{(k)}) \) of Equation (7). From \( \partial Q(\xi; \Psi^{(k)}) / \partial \alpha_i = 0 \) we get
\[ \beta_i = \alpha_i \sum_{j} \tau_{ij}^{(k)} / \sum_{j} \tau_{ij}^{(k)} (y_{ij} - \bar{y}), \]
where \( \bar{y} = \sum \tau_{ij}^{(k)} y_{ij} / \sum \tau_{ij}^{(k)} \).

On plugging \( \beta_i \) of Equation (8) into the equation \( \partial Q(\xi; \Psi^{(k)}) / \partial \beta_i = 0 \), we then get
\[ \log \alpha_i - \psi(\alpha_i) = \log \left( \sum_{j} \tau_{ij}^{(k)} (y_{ij} - \bar{y}) / \sum \tau_{ij}^{(k)} \right) - \sum \tau_{ij}^{(k)} \log (y_{ij} - \bar{y}) / \sum \tau_{ij}^{(k)}, \]
where \( \psi(\alpha) = \partial \log \Gamma(\alpha) / \partial \alpha \) is the digamma function. We solve Equations (8) and (9) for \( \alpha_i \) and \( \beta_i \), until convergence, and denote the estimates as \( \hat{\alpha}^{(k+1)} \) and \( \hat{\beta}^{(k+1)} \).

For the foreground mean, by solving the equation \( \partial Q(\xi; \Psi^{(k)}) / \partial \mu = 0 \), we obtain
\[ \hat{\mu}^{(k+1)}(1) = \left( \hat{\mu}^{(k+1)}(1), \hat{\mu}^{(k+1)}(2) \right) = \sum \tau_{ij}^{(k)} u_{ij} \left( y_{ij} - \hat{\mu}^{(k+1)} \right) / \sum \tau_{ij}^{(k)}(y_{ij} - \hat{\mu}^{(k+1)}), \]
where \( \hat{\mu}^{(k+1)} = \sum \tau_{ij}^{(k)}(u_{ij} \mu) / \sum \tau_{ij}^{(k)} \).

The estimates of the parameters in \( \Psi, \hat{\alpha}, \hat{\beta}, \hat{\alpha_2}, \hat{\beta_2}, \hat{\mu}, \Sigma \) are then inserted into the E-step. These E-step and M-step are alternated until convergence is reached. We refer to this model as the gamma+ mixture model (GTMM); see also Supplementary material for the EM algorithm of the three-parameter gamma+ mixture model where the \( \gamma_i \) are unknown and estimated iteratively.

Let \( \hat{\tau}_{ij} \) be the estimate of the posterior probability obtained after implementing the EM algorithm. The rectangle containing a spot consists of the foreground at its center and the background surrounding the foreground, and the neighboring pixels in each segment must belong to the same class. Therefore, in order to encourage neighboring pixels to have the same class, we obtain the nonparametric kernel estimate \( \hat{\tau}_{ij} \) by smoothing neighboring \( \hat{\tau}_{ij} \)s in Equation (4) of the final step: Define
\[ K_H(u) = \frac{1}{|H|} K(H^{-1}u), \]
where \( K(\cdot) \) is a bivariate nonnegative kernel function. For simplicity we assume that \( K(\cdot) \) is a multiivariate probability density function with \( \int K(u) du = 1 \) and \( \int u K(u) du = 0 \). It is a nonsingular \( 2 \times 2 \) bandwidth matrix, and \( |H| \) denotes its determinant. Then the modified maximum likelihood estimate of the posterior probability with this nonparametric
kernel smoothing is given by

$$
\tilde{r}_{ij} = \frac{1}{n} \sum_{j=1}^{n} w_j \hat{r}_{ij},
$$

where $w_j = K_h(x_j - x_i) / \sum K_h(x_j - x_i)$ is the weight, the kernel $K_h(\cdot)$ is the standard bivariate normal p.d.f., $x_i$ and $x_j$ is the 2D coordinate for the $j$th pixel in the spot rectangle. We used the $2 \times 2$ matrix $H = (h^2, 0, 0, h^2)$ as the bandwidth matrix with a constant $h$. Since the nonparametric kernel smoothing estimator is consistent (Simonoff, 1996) and accounts for the spatial dependency, $\tilde{r}_{ij}$ reflects the posterior probability better than $r_{ij}$ when there exist a few high-intensity pixels in the background region or low-intensity pixels in the foreground region.

The bandwidth $h$ controls the extent to how many neighboring pixels' information is needed to estimate the posterior probability of the pixel to be classified. The proper kernel bandwidth is needed for good estimation in the smoothing step. Though there are many methods to choose the optimal bandwidth (Simonoff, 1996), we select the smoothing parameter by the following simple step. Let $x_1$ and $x_2$ be the coordinates of the pixel to be classified, and let $x_{10}$ and $x_{20}$ be the coordinates of the $j$th neighboring pixel with intensity $y_j$. Let $n_1$ and $n_2$ be the number of pixels in the row and the column of the spot rectangle image, respectively, then the total number of pixels in the spot rectangle is $n_1 \times n_2$. Let $x_i$ and $x_j$ can take the values from $1, \ldots, n_1, i = 1, 2$. As we use the standard Gaussian kernel, we give different weights to the pixels at $x_i$, which are within $x_i \times 1.96h$ approximately according to their closeness to $x_i$ under the 95% confidence level, $i = 1, 2$. Therefore, roughly speaking, if we want to use the information of the nearest pixel $(1, 1)$ on each direction of each axis for smoothing, we need $h = 0.51h$. Also, we need $h \approx 1.02$, $1.53$ if we smooth up to second and third nearest pixel's posterior probabilities on each direction of each axis, respectively. We used $h = 1.02$ for the segmentation of the experimental image, and it worked satisfactorily.

The pixels are segmented according to the values of the smoothed estimates $\tilde{r}_{ij}$; the $j$th pixel is classified as background if $\tilde{r}_{ij} \geq 0.5$, and foreground if $\tilde{r}_{ij} < 0.5$.

Due to system imperfection and microarray image generation process, there are some undesired spots, such as blanks, low-expressed spots and high-expressed artifacts, all of which are common but difficult to segment in typical microarray images. We flag these unusual spots and label them in the output text file. The user can then decide whether to keep or dismiss the spots in the analysis.

We flag blank and low-expressed spots by the Bayesian Information Criterion (BIC). Let $m$ be the number of components in the mixture model, and let $BIC_m$ be the value of BIC with $m$ components in the mixture model. Then the case $m = 1$ corresponds to the situation where there is no foreground, i.e. a blank since background pixels would be the only group in the spot rectangle. Thus, if $BIC_1 < BIC_m$, we treat all pixels as the background, flag the spot as blank, and do not estimate the spot intensity.

The case $m = 2$ would correspond to the typical situation where there exist both foreground and background. When a spot is not flagged as blank but $BIC_2$ is not significantly less than $BIC_1$, it may be uncertain whether there are two groups in the spot rectangle. If $0 \leq BIC_1 - BIC_2 \leq \delta \mid BIC_1 \mid$ for $0 < \delta \ll 1$, i.e. there is not much difference between $BIC_1$ and $BIC_2$ relative to the absolute magnitude of $BIC_1$, the spot is flagged as having low-expression, which denotes the pixels of uncertain classification. We estimate the intensities for these uncertain spots. These intensity estimates should be used with caution for further analysis. We used $\delta = 0.01$ in producing the results in this paper.

We flag a spot as being valid if it is neither blank nor has uncertain classification (low-expression). Finally, the high-intensity artifact region is detected on the foreground in the valid spots. Since the $t$ distribution has long tails, the segmented foreground may contain the very high-expressed pixels, which consists of artifacts. For each channel, we rearrange the intensities of the pixels segmented as foreground in ascending order. Let $Q_{10}$ and $Q_{90}$ be the 3rd quartile of the foreground pixel intensities of $R$ and $G$, respectively, and let $IQR_R$ and $IQR_G$ be the interquartile ranges of the same foreground pixel intensities for each color. Then the pixel in the foreground whose intensities are $R_i$ and $G_i$ is classified as a high-intensity artifact if $R_i > Q_{10} + 3 \times IQR_R$ and $G_i > Q_{90} + 3 \times IQR_G$. These high-intensity artifacts are excluded from the foreground when the foreground intensity is estimated.

Having effected segmentation, we now estimate the spot intensity using the intensity information of the segmented pixels. Since the $\gamma$’s in Equation (3) are now specified, we maximize the log likelihood to estimate the parameters in $\Psi$. That is, we obtain the m.i.e.’s $\hat{\alpha}^*, \hat{\beta}^*, \hat{\mu}^*$ of the parameters with the constraint $\mu_0 \geq \alpha^*/\beta^* + y^*$ using the intensities of the pixels classified into each class. To estimate the background-corrected intensities, we use $\hat{\phi}^* = \hat{\mu}^* - (\hat{\alpha}^*/\hat{\beta}^* + \hat{y}^*)$, where $\hat{y}^*$ is the location parameter estimate from the data, $i = 1, 2$, and estimate the log-ratio for each spot by $\log(\hat{\phi}^*/\hat{\phi}^*)$. Note it is guaranteed that both $\hat{\phi}^*$ and $\hat{\phi}^*$ are nonnegative.

### 3 RESULTS

#### 3.1 Simulated datasets

We generated artificial spot data, which have similar background and foreground locations as in one of the real spots in the CSIRO experimental image. That is, we assume a spot rectangle consisting of $17 \times 17$ pixels. The background distributions are assumed to be gamma($\alpha_i, \beta_i, \gamma_i), i = 1, 2$, where $\alpha_1 = 1, \beta_1 = 0.1, \gamma_1 = 7, \alpha_2 = 2, \beta_2 = 0.1, \gamma_2 = 10$. Therefore, the background mean is $\mu_0 = (\mu_{01}, \mu_{02}) = (17, 30)$. The two independent $t$ distributions with the location parameters $\mu_0 = \mu_{01} + 20$, degrees of freedom $\nu_{10} = 20$ and the dispersion parameters $\sigma^2_{10} = 100$ are assumed for the foreground channels, $i = 1, 2$, respectively. Let $(x_{10}, y_{10})$ be the coordinate for the $(i, j)$th pixel in the spot rectangle located at $i$th column and $j$th row of the rectangle, where $x_{10}, y_{10} = 1, 2, \ldots, 17$. Then the intensity of the $(i, j)$th pixel is randomly generated from the foreground distribution if $(x_{10} - 9)^2 + (y_{10} - 9)^2 \leq 6^2$, and from the background distribution otherwise.

Our method uses the kernel smoothing technique to increase the chances that neighboring pixels are assigned to the same class after implementing the EM algorithm. Figure 2 shows the more effective segmentation according to the smoothed posterior probability $\tilde{r}_{ij}$ compared to that with $\hat{r}_{ij}$.

Figure 2a shows the combined image of the two channel intensity images which were randomly generated according to the above procedure, and Figure 2b shows the scatter plot of the 2D intensities.
indicating their true classes. Some of the foreground and background pixels are mixed together around the area where the tails of the two distributions overlap, and two particular pixels are located deep inside of the other distribution’s domain. Figure 2c and d give the pixel segmentation results according to the non-smoothed posterior probability $\pi_0$. The classification boundary can be recognized along the line where the two p.d.f.’s meet each other, as shown in Figure 2d. This, however, misclassifies some foreground pixels into the background as well as misclassifying some background pixels into foreground, which is shown in Figure 2c. On the other hand, the smoothed posterior probability $\hat{\pi}_0$ produces the desired segmentation of the spot image as shown in Figure 2e. The high-intensity pixel in the foreground and the low-intensity pixel in the foreground are classified correctly as well as many pixels located on the border of two distributions (Fig. 2f). The correct classification of each pixel is due to utilizing the neighboring pixels’ spatial information on their class.

In order to see how well our method works for estimation, we generated artificial spot data from the distributions with different sets of true parameter values. Then we estimated the parameters for each dataset, and checked the compatibility of the parameter estimates to the true values for each set of parameters.

Each spot rectangle consists of 17 × 17 pixels. We assume the same three-parameter gamma distribution as in Figure 2 for the background intensity. Therefore, the background mean is $\mu_b = (\mu_{b1}, \mu_{b2})^T = (17, 30)^T$. The foreground intensities are generated from the $t$ distributions with location parameter $\mu_\nu = (\mu_{\nu1}, \mu_{\nu2})^T$, degrees of freedom $v_\nu = 20$, dispersion parameters $\sigma^2_\nu = 100$, $i = 1, 2$, and no correlation, where $\mu_\nu$ is $\sqrt{5} \Delta$ (Euclidean) distant from $\mu_b$ in an upper right direction. We set $\Delta = 20, 30, 40$ and

arranged $\mu_{\nu1} = \mu_{b1} + \phi_1$ and $\mu_{\nu2} = \mu_{b2} + \phi_2$ to have $\log_2(\phi_1/\phi_2) = -1, 0, 1$ for each $\Delta$. For each parameter $\theta = (\mu_{b1}, \mu_{b2}, \mu_{\nu1}, \mu_{\nu2}, \phi_1, \phi_2)^T$, let $\theta_0$ be the true parameter value for each combination of $\Delta$ and $\log_2(\phi_1/\phi_2)$. We set $\theta = \theta_0$, generated 100 spot rectangles, estimated the parameter $\hat{\theta}$ using the whole procedure for segmentation and estimation with the GTMM algorithm for each spot, and calculated the mean $\hat{\theta}$ of the estimates $\hat{\theta}$. To check the compatibility of the parameter estimates with the true values, we performed a $t$-test for the hypothesis $H_0: E(\hat{\theta}) = \theta_0$ versus the two-tailed alternative with 100 parameter estimates and obtained the $P$-value for each parameter. The results are summarized in Table 2 for the case where $\log_2(\phi_1/\phi_2) = 1$; see Supplementary Table 1 for other cases. The estimate $\hat{\theta}$ is close to the true parameter value $\theta_0$ on average for each set. The $P$-values for the estimates are well above the usual significance level 0.05 except for some estimates in the case where the two class means are relatively close ($\Delta = 20$). The estimates become closer to the true parameter values (larger $P$-value) as the two class central locations grow apart (larger $\Delta$). The G3TMM algorithm also gave similar results.

### 3.2 Real datasets

We compared our proposed method (using GTMM) with the Spot software of CSIRO (http://spot.cmis.csiro.au/spot/) and the spotSegmentation of Li et al. (2005) in its application to two real microarray images. The first image dataset consists of the files spot1cy5.tif and spot1cy3.tif from http://spot.cmis.csiro.au/spot/demodownload/SpotExamples.zip for Spot, the commercial image analysis software of CSIRO Mathematical and Information Sciences. The second dataset is the spotSegTest dataset of spotSegmentation1.4.0.zip from http://www.bioconductor.org/packages/bioc/html/spotSegmentation.html for the spotSegmentation package (Li et al., 2005), which consists of a portion of the first block from the first microarray slide image from van’t Wout et al. (2003).

Figure 3 displays the original top left 24 spot images of the first block in the example microarray of Spot software with the segmentation results of the spots using our approach (Fig. 3b); spotSegmentation (spotSeg) of Li et al. (2005) (Fig. 3c); and Spot of CSIRO (Fig. 3d), respectively. In Figure 3a, the original images contain spots with various shapes as well as blank spots. Our method

### Table 2. The mean of the parameter estimates and the $P$-value for the $t$-test on the parameter estimate, obtained from 100 synthetic spot data

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>$\mu_{b1}$</th>
<th>$\mu_{b2}$</th>
<th>$\mu_{\nu1}$</th>
<th>$\mu_{\nu2}$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\log_2(\phi_1/\phi_2)$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>17.00</td>
<td>30.00</td>
<td>57.00</td>
<td>50.00</td>
<td>40.00</td>
<td>20.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>17.21</td>
<td>30.23</td>
<td>57.07</td>
<td>50.01</td>
<td>39.86</td>
<td>19.78</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.00</td>
<td>0.02</td>
<td>0.31</td>
<td>0.90</td>
<td>0.15</td>
<td>0.09</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>17.00</td>
<td>30.00</td>
<td>77.00</td>
<td>60.00</td>
<td>60.00</td>
<td>30.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>17.00</td>
<td>30.00</td>
<td>77.04</td>
<td>59.98</td>
<td>60.13</td>
<td>29.91</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.19</td>
<td>0.53</td>
<td>0.58</td>
<td>0.75</td>
<td>0.19</td>
<td>0.51</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>17.00</td>
<td>29.99</td>
<td>97.00</td>
<td>70.01</td>
<td>80.00</td>
<td>40.03</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>17.00</td>
<td>29.99</td>
<td>97.00</td>
<td>70.01</td>
<td>80.00</td>
<td>40.03</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.944</td>
<td>0.894</td>
<td>0.97</td>
<td>0.85</td>
<td>0.99</td>
<td>0.83</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>

$\mu_{b1}, \mu_{b2}$: mean red and green intensity in background; $\mu_\nu, \mu_\nu$: mean red and green intensity in foreground; $\phi_1, \phi_2$, true parameter value; $\hat{\theta}$: mean of the parameter estimates; $p$: $P$-value for $H_0: E(\hat{\theta}) = \theta_0$. 

---

Fig. 2. (a and b): true image and scatter plot of the intensities indicating their true classes; (c and d): image and scatter plot of the segmented intensities based on non-smoothed posterior probability $\pi_0$; (e and f): image and scatter plot of the segmented intensities based on $\pi_0$. The circle in the scatter plot corresponds to the foreground, and * to the background.
recognizes them correctly as shown in Figure 3b. The white pixels correspond to the foreground, the black to the background, and the gray to the low-expressed uncertain foreground. The irregular shapes including doughnut are detected in the low-expressed spots. In each scatter plot of the original \((R, G)\) intensity points whose spot is classified as blank, there exist no two clusters but one. The bandwidth of the kernel for calculating the smoothed posterior probability was fixed at \(h = 1.02\). The package spotSegmentation detected blank spots, but generated noisy foreground segmentations in many spot rectangles (Fig. 3c). On the other hand, the Spot software segmented every spot rectangle into foreground and background including the blank image (Fig. 3d).

The spotSegTest dataset consists of 24 spots arranged in \(4 \times 6\) array. Figure 4 shows the segmentation results of four particular spot rectangles obtained by our method (GTMM) and the Spot and spotSegmentation methods. GTMM used the log transformed data. Each row represents a specific spot rectangle, the first two columns show the \(R\) and \(G\) intensities of the corresponding spot rectangle, respectively, and the remaining three columns contain the segmentations obtained by three different methods. The spots shown in consecutive rows of Figure 4 are (1, 3)th, (2, 1)th, (2, 5)th, (3, 6)th spot rectangles from upper left corner in the spotSegTest dataset image. The light gray region in GTMM (3rd row and 3rd column) represents the bright artifact, and the dark gray region (4th row and 3rd column) represents the low-expressed uncertain spot. The dark gray region in spotSegmentation (4th column) indicates the uncertain classification determined by Li et al. (2005). The result of Spot (5th column) contains the original image overlayed by the segmented image.

For the first spot, the three methods produced similar segmentations. As shown in the first and second columns, there is no expression, however, in the red channel, while there is high-expression in the green channel. Thus the background-corrected mean intensity of \(R, \phi_1\), should be 0 and the value of \(\log_2(\phi_1/\phi_2)\) for this spot is \(-\infty\). The estimate \(\phi_1\) by GTMM is 0 since both foreground and background means of \(R\) were estimated as the same value of 5.6 (log scaled intensity), and \(\log_2(\phi_1/\phi_2)\) is \(-\infty\). The estimate of the background \(R\) mean by spotSegmentation is 303.0, while that of the foreground mean is 259.3. This produces a negative value for the estimate of the background-corrected mean, which is the common logical error arising when two means are estimated separately. Spot also generates a negative estimate for the background-corrected \(R\) mean since the estimates of foreground and background \(R\) mean are 256.6 and 347.9, respectively. The third spot (3rd row) contains the bright region on the bottom right corner of both \(R\) and \(G\) foreground, which indicates the existence of high-expression artifact. GTMM detected the artifact, spotSegmentation classified the region into uncertain area, but Spot did not detect it. The fourth spot (4th row) is blank as shown in the \(R\) and \(G\) columns. GTMM and spotSegmentation classified it as an uncertain spot, while Spot obtained the segmentation from the blank spot with the negative background-corrected mean intensity estimates for \(R\) and \(G\). The rest of the spots in the spotSegTest dataset were segmented well with respect to the original images as the second spot (2nd row) in Figure 4 by all three methods.

## 4 DISCUSSION AND CONCLUSIONS

We have described a mixture model approach based on component gamma and \(t\) distributions to spot segmentation and intensity estimation for microarray images. The gamma distribution in the model is suited for the background intensity in the sense that it accounts for the intensities with a positive lower bound and is very flexible in its shape including asymmetric exponential type to symmetric normal type. It is bivariate by taking the \(R\) and \(G\) intensities to be independent in the background. For the underlying distribution for the foreground, we assume the bivariate \(t\) distribution. The \(t\) distribution provides a longer-tail alternative to the normal distribution, and its location parameter estimate is less affected by atypical observations. We implemented the EM algorithm to estimate the pixels’ posterior probabilities, and then applied a nonparametric kernel smoothing technique that utilizes the neighborhood information in forming the posterior probabilities for the final segmentation. Our model constrains the mean intensity for the foreground to be greater than that for the background. Thus we
can estimate the two mean intensities avoiding the common logical
error that estimates of foreground may be less than those of the
 corresponding background.

We demonstrated our method on some simulated and real datasets
in which it showed superior performance to the existing procedures
considered. Our method identified various shapes of spots and blank
spots as well as bright artifacts in the real experimental microarray
images. In the simulations, it was also shown to estimate the back-
ground subtracted mean intensity closely to its true value.

ACKNOWLEDGEMENTS

The authors are grateful for helpful comments from two referees that
greatly improved the paper. The first author appreciates the hospita-
tality of Department of Mathematics, University of Queensland,
Australia for providing the facilities for this research during his
visit. The work of J.B. and Y.S.S. was supported by Grant no.
RTI04-03-03 from the Regional Technology Innovation Program
The work of G.J.M. was supported by the Australian Research
Council.

Conflict of Interest: none declared.

REFERENCES


Chen,Y. et al. (1997) Ratio-based decisions and the quantitative analysis of cDNA


Demirkaya,Ö. et al. (2005) Segmentation of cDNA microarray spots using markov
random field modeling. Bioinformatics, 21, 2994–3000.


Gottardo,R. et al. (2006) Probabilistic segmentation and intensity estimation for


Sons, NY.

Li,Q. et al. (2005) Donuts, scratches and blanks: robust model-based segmentation of
microarray images. Bioinformatics, 21, 2875–2882.

NY.

statistical inference about gene expression changes from microarray data. J. Com-


Steinfath,M. et al. (2001) Automated image analysis for array hybridization experi-

van’t Wout,A.B. et al. (2003) Cellular gene expression upon human immunodeficiency
type 1 infection of CD4(+) T-cell lines. J. Virology, 77, 1392–1402.

Yang,Y.H. et al. (2002) Comparison of methods for image analysis on cDNA microar-