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

Jangsun Baek, Young Sook Son and Geoffrey J. McLachlan
Bioinformatics 2007, Vol. 23 no. 4, pages 458-465

Seminar in bioinformatics

Triinu Kõressaar

TARTU
2007
Simultaneously performed image segmentation and intensity estimation

**Two-component mixture model**
- background intensity
- foreground intensity
  => Intensity measurement is a bivariate vector
    (red and green intensities)

**Background intensity component** – bivariate gamma distribution

**Foreground intensity component** - bivariate t-distribution
Segmentation methods can be grouped for instance whether a parametric distribution of the pixel intensity is assumed or not:

1. **Nonparametric segmentation**

   * no particular type of distribution on intensities is assumed (e.g. fixed circle segmentation, adaptive circle segmentation, seeded region growing method)

2. **Parametric segmentation**

   * distribution for intensity is specified up to a vector of unknown parameters (e.g. bivariate normal distribution, scaled bivariate normal distribution, exponential distribution, uncorrelated bivariate t distribution)
Description of image segmentation and intensity estimation method

Presentation of the stable parameter estimation result of the proposed method using synthetic data

Presentation of the segmentation and estimation results from applying the method to two real experimental microarray image datasets

Comparison of the results with those from other methods

Summary of the method
microarray image analysis:

1. automatic gridding
2. model-based clustering of pixels
3. intensity estimation

Data

a pair of unsigned 16-bit images (.tiff) -> transformation of images (square root or logarithms) for:

- preventing very bright pixels from dominating
- making the work with images computationally more efficient
**Automatic gridding**

Identifying blocks and positioning rows and columns of spots within each block

Using the combined image, projecting the intensities by summing up across the pixels in each row and each column

Smoothing the projections using robust loess (bandwith – width of a typical spot) -> series of peaks and valleys

Grid – drawing a line in each valley
Distribution of pixel intensities

(a)

(b)

(c)

(d)

\[ \text{Graph} \]

\[ R \quad \sigma \]
Kolmogorov-Smirnov goodness-of-fit test:
for checking the validity of the proposed distributions

Table 1. Mean $P$-values for goodness-of-fit test

<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></td>
<td>(0.1405)</td>
<td>(0.0036)</td>
<td>(0.1031)</td>
<td>(0.1276)</td>
</tr>
<tr>
<td>Normal</td>
<td>0.0012</td>
<td>0.0635</td>
<td>0.6913</td>
<td>0.5888</td>
</tr>
<tr>
<td></td>
<td>(0.0014)</td>
<td>(0.0777)</td>
<td>(0.7380)</td>
<td>(0.6847)</td>
</tr>
<tr>
<td>Gamma/t</td>
<td>0.2325</td>
<td>0.2144</td>
<td>0.7726</td>
<td>0.7269</td>
</tr>
<tr>
<td></td>
<td>(0.1501)</td>
<td>(0.1243)</td>
<td>(0.8237)</td>
<td>(0.8046)</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.
**Image segmentation and intensity estimation**

Density function of observed intensities:

$$f(y; \Psi) = \pi_1 f_1(y; \Theta) + \pi_2 f_2(y; \mu, \Sigma, \nu), \text{ where } \Psi = \{ \Theta, \mu, \Sigma, \nu, \pi_i \},$$

where $f_1$ and $f_2$ are the p.d.f. for background and foreground pixel intensity distributions, respectively.

$\pi_i$ – probability that pixel belongs to the $i$th component

**EM algorithm** to obtain the maximum likelihood estimates of the parameters in $\Psi$
$\hat{\tau}_{ij}$ – the estimate of posterior probability that $y_j$ belongs to the $i$th component of the mixture

Rectangle containing a spot – foreground at the center, background surrounding the foreground -> neighboring pixels in each segment must belong to the same class

Final step of EM: nonparametric kernel estimate $\hat{\tau}^*_{ij}$ ($\hat{\tau}_{ij}$ is multiplied with weight), weight is dependent on $h$

$h$ – bandwidth – how many neighboring pixels’ information is needed to estimate the posterior probability of the pixel to be classified
Step for choosing the $h$: different weights to the pixels at $x_{ij}$’s which are within $x_i \pm 1.96h$ according to their closeness to $x_i$ under the 95% confidence level; smoothing up to first nearest pixel $h \approx 0.51$

Pixels are segmented according to $\hat{\tau}^*_{ij}$:

- $j$th pixel is classified as background if $\hat{\tau}^*_{ij} \geq 0.5$
- as foreground if $\hat{\tau}^*_{ij} < 0.5$
Blank and low-expressed spots – they are flagged

**BIC** – Bayesian Information Criterion

**m**  – the number of components in the mixture model

**m=1** -> no foreground, only background

**BIC** < **BIC**₂ -> all pixels are treated as background

- spot is flagged as a blank
- intensity of spot is _not_ estimated

**m=2** -> foreground + background

*if spot is _not_ flagged as blank, **BIC**₂ is _not_ significantly less than **BIC**₁

-> are there two groups in the spot rectangle?

If **0 ≤ BIC**₁ - **BIC**₂ ≤ δ|**BIC**₁| for **0 < δ < 1** -> spot is flagged as having low expression (the pixels of uncertain classification); intensity of spot _is_ estimated
high-intensity artifact regions

Detected on the foreground in the valid spots

For each channel the intensities of the pixels segmented as foreground are rearranged in ascending order

$Q_{3R}$ and $Q_{3G}$ - 3rd quartile of the foreground pixel intensities of R and G

$IQR_R$ and $IQR_G$ - interquartile ranges of the same foreground pixel intensities

Pixel with intensities $R_i$ and $G_i$ is classified as a high-intensity artifact if $R_i > Q_{3R} + 3xIQR_R$ and $G_i > Q_{3G} + 3xIQR_G$.

High intensity artifacts are excluded from the foreground when the foreground intensity is estimated
Intensity estimation

Spot intensity is estimated by maximum likelihood

The mean intensity of foreground $\geq$ the mean intensity of background

Estimation of intensity for each spot - log ratio of background-corrected R and G intensities ($\log_2(\hat{\phi}^*_1 / \hat{\phi}^*_2)$)
Results

1. Simulated data::segmentation

Spot rectangle is 17x17 pixels

Background: dist. - gamma($\alpha_i, \beta_i, \gamma_i$), i=1,2, where $\alpha_1=1$, $\beta_1=0.1$, $\gamma_1=7$, $\alpha_2=1$, $\beta_2=0.1$, $\gamma_2=7$; $\mu_b=(\mu_{b1},\mu_{b2})'=(17,30)$

Foreground: two independent t distributions - location parameters $\mu_{fi} = \mu_{bi} + 20$, DF $\nu_i=20$, dispersion parameters $\sigma^2_i=100$.

$(x_{1(ij)}, x_{2(ij)})$ - coordinate for the $(i, j)$th pixel in the spot rectangle located at $i$th column and $j$th row of the rectangle lattice, where $x_{1(ij)}, x_{2(ij)}=1, 2, \ldots, 17$.

Intensity of the $(i, j)$th pixel is randomly generated from the foreground distribution if $(x_{1(ij)}-9)^2+(x_{2(ij)}-9)^2 \leq 6^2$, and from the background distribution otherwise.
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 \( \hat{\tau}_{ij} \); (e and f): image and scatter plot of the segmented intensities based on \( \hat{\tau}_{ij}^* \). The circle in the scatter plot corresponds to the foreground, and * to the background.
**Results:** 1. Simulated data:: Intensity analysis

**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>$\mu_b_1$</th>
<th>$\mu_b_2$</th>
<th>$\mu_f_1$</th>
<th>$\mu_f_2$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\log_2(\phi_1/\phi_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_0$</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>
</tr>
<tr>
<td>20 $\hat{\theta}$</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>
</tr>
<tr>
<td>$\theta_0$</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>
</tr>
<tr>
<td>$\phi_1$</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>
</tr>
<tr>
<td>$\phi_2$</td>
<td>16.91</td>
<td>30.06</td>
<td>77.04</td>
<td>59.98</td>
<td>60.13</td>
<td>29.91</td>
<td>1.01</td>
</tr>
<tr>
<td>$\phi_1/\phi_2$</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>
</tr>
<tr>
<td>$\theta_0$</td>
<td>17.00</td>
<td>30.00</td>
<td>97.00</td>
<td>70.00</td>
<td>80.00</td>
<td>40.00</td>
<td>1.00</td>
</tr>
<tr>
<td>30 $\hat{\theta}$</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>
</tr>
<tr>
<td>$\theta_0$</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>
</tr>
</tbody>
</table>

$\mu_b_1$, $\mu_b_2$: mean red and green intensity in background; $\mu_f_1$, $\mu_f_2$: mean red and green intensity in foreground; $\phi_1$: $\mu_f_1 - \mu_b_1$; $\phi_2$: $\mu_f_2 - \mu_b_2$; $\theta_0$: true parameter value; $\hat{\theta}$: mean of the parameter estimates; $p$: $P$-value for $H_0: E(\theta) = \theta_0$. 

Results:

Real datasets: GTMM, Spot software and spot-Segmentation

(a) original image

(b) GTMM

(c) spotSeg

(d) Spot
Conclusion

Gamma distribution for the background intensity:
- is very flexible in its shape (asymmetric exponential type to symmetric normal type)
- is bivariate by taking the R and G intensities to be independent in the background

Bivariate t distribution for the foreground intensity:
- provides a longer-tail alternative to the normal distribution
- less affected by atypical observations
Conclusion

EM algorithm to estimate the pixels’ posterior probabilities, a nonparametric kernel smoothing technique that utilizes the neighborhood information in forming the posterior probabilities for the final segmentation.

Model constrains the mean intensity for the foreground to be greater than that for the background.