On the categorization of a beta-uniform distribution and its application to the gene expression problem

Mariana Rodrigues-Motta[1,*] and Aluísio Pinheiro[1]
[1] Department of Statistics
University of Campinas-*UNICAMP*, Brazil

Abstract

Based on a set of cutting points we propose to categorize a beta-uniform (BUM) distribution, and term the categorized distribution as the C-BUM distribution. We study the categorization effects by comparing the performance of the maximum likelihood estimates for both models, for different sets of cutting points. Using the missing information principle we compute the BUM and C-BUM theoretical Fisher information matrices to assess loss of information. Finally, we evaluate the method to a published cancer gene expression data study.

Key Words: *beta-uniform; EM algorithm; microarray data; p-value distribution*

1 Introduction

In microarray experiments investigators are allowed to compare the expression levels of thousands of genes in samples collected from different tissues or at different time points or conditions. An common goal in DNA microarray experiments is the identification of differentially expressed genes. Thousands of p-values are generated when comparisons among levels of expressions under different conditions are performed by hypothesis testing, one per gene,

[*] *E-mail address*: marianar@ime.unicamp.br

resulting in a very complex multiple-testing problem. In general, levels of expression are compared under one specific condition, for example different time points. However, futurely it may be of interest to investigate how other variables, for example treatment factors, affect those genes that were found to be statistically expressed. Additionally, investigators may prefer working with the outcome in a categorical way, and create reference classes to classify future outcomes.

For a set of p-values, those arising from the null hypothesis are distributed uniformly on the interval (0,1) (Casella and Berger, 1990). A reasonable model for the distribution of p-values arising from microarray experiments was proposed by Pounds and Morris (2003), which express the distribution of a set of p-values as a mixture of a uniform (0,1) component and a beta component, named as beta-uniform (BUM) model. In this study, as a method of categorization of p-values, first we propose to approximate the distribution of a set of p-values by the BUM model. Second, categorize those p-values using arbitrary cutting points, leading to categories of p-values that could represent those genes that are highly expressed, mild expressed or not expressed, for example. If a continuous variable has a linear regression with some predictor variables, then the same effect parameters apply to a discrete version with the categorized variable (Agresti, 2002). Although this is not the purpose of our work, this feature makes it possible to compare estimates from studies using different response scales.

This study proposes to categorize a beta-uniform distribution in the context of DNA microarray experiments. First, we introduce the beta-uniform (BUM) model and develops the categorized beta-uniform (C-BUM) model from an arbitrary set of cutting points, as described in Section 2. The published data set of Luberhofer et al. (2002) used to evaluate the method is described in Section 3. Although the focus is on categorizing genes from significance of levels of expressions, the method described in this article is applicable to any data whose distribution is approximated by a BUM model.

2 Methods

2.1 The BUM and C-BUM models

Let $\mathbf{x} = (x_1, ..., x_n)'$ be a vector of n independent random variables, with $x_i \in (0, 1)$ representing, for example, the i^{th} p-value obtained from some statistical test. We assume that x_i follows a beta-uniform mixture distribution given by

$$f(x_i; \lambda, \alpha) \;=\; \lambda + (1-\lambda)\alpha x_i^{\alpha-1}, I_{(0,1)}(x_i), \tag{1}$$

where $\lambda \in (0,1)$ and $\alpha \in (0,1)$ are the mixture and beta distribution parameters, respectively, and I is an indicator function. Now, using arbitrary cutting points represented by the vector $\mathbf{l} = (l_0 = 0, l_1, l_2, ..., l_c = 1)'$, where $c = 1, 2, 3, ...$ is a known constant, the $(0,1)$ beta-uniform interval is discretized into the intervals $(l_0, l_1), (l_1, l_2), ..., (l_{c-1}, l_c)$. We associate the category 1 to the (l_0, l_1) interval, the category 2 to the (l_1, l_2) interval, and so on. Therefore, if a certain x_i falls into the (l_{j-1}, l_j) interval, the category j, $j = 1, ..., c$, is associated to it. Following this scheme, a new random vector is created, say $\mathbf{y} = (y_1, ..., y_n)'$, with $y_i = 1, ..., c$, and the probability distribution of y_i is given by

$$\begin{aligned} p(y_i; \alpha, \lambda) &= P(Y_i = y_i) \\ &= P(l_{y_{i-1}} \leq X_i \leq l_{y_i}) \\ &= \int_{l_{y_{i-1}}}^{l_{y_i}} [\lambda + (1-\lambda)\alpha x_i^{\alpha-1}] dx_i \\ &= \lambda(l_{y_i} - l_{y_{i-1}}) + (1-\lambda)(l_{y_i}^{\alpha} - l_{y_{i-1}}^{\alpha}). \end{aligned} \tag{2}$$

Therefore, the log-likelihood function is given by

$$l(\lambda, \alpha; \mathbf{y}) \;=\; \sum_{i=1}^{n} \ln[\lambda(l_{y_i} - l_{y_{i-1}}) + (1-\lambda)(l_{y_i}^{\alpha} - l_{y_{i-1}}^{\alpha})], \tag{3}$$

where ln is the natural logarithm function. In particular, the maximization of $l(\lambda, \alpha; \mathbf{y})$ with respect to λ and α is not straightforward and require iterative procedures. The estimation of the parameters of a mixture can be handled by a variety of techniques (see Titterington et al. (1985) for an exhaustive review of those methods), and in this study we used the numerical approach proposed by Byrd et al (1995).

2.2 The Fisher information matrix for the BUM and C-BUM models

Using the missing information principle introduced by Woodburry (1971), we compute the BUM and C-BUM information matrices. Suppose we know which observations came from the first component and which came from the second; that is, suppose we could observe $Z_i = 1$ when Y_i is from the first component and $Z_i = 0$ when from the second component. Then, the C-BUM associated log-likelihood with the complete data $(\mathbf{y}, \mathbf{z})$ is given by

$$\begin{aligned} l(\lambda, \alpha; \mathbf{y}, \mathbf{z}) &= \sum_{i=1}^{n} \ln\{[\lambda(l_{y_i} - l_{y_i-1})]^{z_i}[(1-\lambda)(l_{y_i}^{\alpha} - l_{y_i-1}^{\alpha})]^{1-z_i}\} \\ &= \ln\lambda \sum_{i=1}^{n} z_i + \sum_{i=1}^{n} z_i \ln(l_{y_i} - l_{y_i-1}) \\ &+ \ln(1-\lambda)\sum_{i=1}^{n}(1-z_i) + \sum_{i=1}^{n}(1-z_i)\ln(l_{y_i}^{\alpha} - l_{y_i-1}^{\alpha}). \quad (4) \end{aligned}$$

Let $\boldsymbol{\theta} = (\lambda, \alpha)^{'}$. Louis (1982) demonstrated that

$$\frac{d^2 l_c(\lambda, \alpha; \mathbf{y}, \mathbf{z})}{d\boldsymbol{\theta} d\boldsymbol{\theta}^{'}} = E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left[\frac{d^2 l_c(\lambda, \alpha; \mathbf{y}, \mathbf{z})}{d\boldsymbol{\theta} d\boldsymbol{\theta}^{'}}\right] + Var_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left[\frac{d l_c(\lambda, \alpha; \mathbf{y}, \mathbf{z})}{d\boldsymbol{\theta}}\right],$$

which, on multiplying by -1, yields

$$\mathbf{I}(\boldsymbol{\theta}|\mathbf{y}) = \mathbf{I}_c(\boldsymbol{\theta}|\mathbf{y}) - \mathbf{I}_m(\boldsymbol{\theta}|\mathbf{y}),$$

where $\mathbf{I}(\boldsymbol{\theta}|\mathbf{y})$, $\mathbf{I}_c(\boldsymbol{\theta}|\mathbf{y})$ and $\mathbf{I}_m(\boldsymbol{\theta}|\mathbf{y})$ can be interpreted as the observed, complete and missing information matrices. Here, $E_{\mathbf{z}|\mathbf{y},\theta}\left[-\frac{\partial^2 l_c(\lambda,\alpha;\mathbf{y},\mathbf{z})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{'}}\right]$ and $Var_{\mathbf{z}|\mathbf{y},\theta}\left[\frac{\partial l_c(\lambda,\alpha;\mathbf{y},\mathbf{z})}{\partial\boldsymbol{\theta}}\right]$ are linear in Z_i, so we just replace z_i by

$$\begin{aligned} E_{\mathbf{z}|\mathbf{y};\lambda,\alpha}(Z_i) &= P(Z_i = 1|\mathbf{y}; \lambda, \alpha) \quad (5) \\ &= \left[\frac{(1-\lambda)}{\lambda}\left(\frac{l_{y_i}^{a} - l_{y_i-1}^{a}}{l_{y_i} - l_{y_i-1}}\right) + 1\right]^{-1} \end{aligned}$$

or

$$\begin{aligned} Var_{\mathbf{z}|\mathbf{y};\lambda,\alpha}(Z_i) &= E_{\mathbf{z}|\mathbf{y};\lambda,\alpha}(Z_i)[1 - E_{\mathbf{z}|\mathbf{y};\lambda,\alpha}(Z_i)] \\ &= (9)[1-(9)], \quad (6) \end{aligned}$$

when appropriate.

Now, the log-likelihood of the BUM model is given by

$$L(\lambda, \alpha; \mathbf{x}) = \prod_{i=1}^{n} \ln[\lambda + (1-\lambda)\alpha x_i^{\alpha-1}], \quad (7)$$

and suppose we could observe $V_i = 1$ when X_i is from an uniform distribution in the $(0, 1)$ interval, and $V_i = 0$ when from the $Beta(\alpha, 1)$ distribution. The log-likelihood with the complete data $(\mathbf{x}, \mathbf{v})$ is

$$\begin{aligned} L_c(\lambda, \alpha; \mathbf{x}, \mathbf{v}) &= \ln(\lambda) \sum_{i=1}^{n} v_i + \ln(1-\lambda) \sum_{i=1}^{n} (1 - v_i) \qquad (8) \\ &+ \ln(\alpha) \sum_{i=1}^{n} (1 - v_i) + (\alpha - 1) \sum_{i=1}^{n} (1 - v_i) \ln(x_i), \end{aligned}$$

and $E_{\mathbf{v}|\mathbf{x},\theta} \left[-\frac{\partial^2 L_c(\lambda,\alpha;\mathbf{x},\mathbf{v})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]$ and $Var_{\mathbf{v}|\mathbf{x},\theta} \left[\frac{\partial L_c(\lambda,\alpha;\mathbf{x},\mathbf{v})}{\partial \boldsymbol{\theta}} \right]$ are linear in V_i, so we just replace v_i by

$$\begin{aligned} E_{\mathbf{v}|\mathbf{x};\lambda,\alpha}(V_i) &= P(V_i = 1 | \mathbf{y}; \lambda, \alpha) \\ &= \left\{ \frac{(1 - \lambda^{(k)})}{\lambda^{(k)}} \left[\alpha x_i^{\alpha - 1} \right] + 1 \right\}^{-1} \qquad (9) \end{aligned}$$

or

$$\begin{aligned} Var_{\mathbf{v}|\mathbf{x};\lambda,\alpha}(V_i) &= E_{\mathbf{v}|\mathbf{x};\lambda,\alpha}(V_i)[1 - E_{\mathbf{v}|\mathbf{x};\lambda,\alpha}(V_i)] \\ &= (9)[1 - (9)], \qquad (10) \end{aligned}$$

when appropriate.

Since the complete-data are not fully observed, we take the conditional expectation of $\mathbf{I}(\boldsymbol{\theta}|\mathbf{y})$ and $\mathbf{I}(\boldsymbol{\theta}|\mathbf{x})$ over Y and X, obtaining the Fisher information matrices for the C-BUM and BUM models, respectively.

3 An example

3.1 Description of the experiment

The Luberhofer et al. (2002) data set is used to illustrate the method presented here. This data set comes from a study of gene expression of a hormone-responsive breast cancer cell line (MCF-7) treated with a mitogenic dose of estrogen in the absence of confounding growth factors found in serum. For each replication, cell samples were treated either with ethanol (control) or with estrogen, and gene expression changes were monitored 1, 4, 12, 24, 36, and 48 hours after estrogen stimulation so that RNA levels at critical times throughout cell cycle progression could be monitored. The cDNA analysis were conducted on two replicate cultures. In the first biological replicate the quantity of total RNA of 35 μg for Cy3-labelled probes and 75 μg for Cy5 was used. A total of 30 μg was used for either label during the second replicate. Each RNA was hybridized to four arrays, two of each dye orientation. Because each time point was also biologically replicated a total of eight arrays for each time point was analyzed. This study produced gene expression data for 1901 genes in $n = 48$ arrays. The microarray data are expressed as a ratio of the level of expression of estrogen-treated *vs* control cells.

4 Results

To assess the effect of fluorophore incorporation process in the microarray experiment, which may generate discordant expression values, a Kruskal-Wallis (Kruskal and Wallis, 1952) test was used across replicates hybridizations in the database. Using a significance level of 5%, 671 genes were excluded from the list of estrogen-regulated genes. The remaining 1230 genes were tested as estrogen-regulated inside each of the 6 time points, using the Wilcox (Hollander, M. and Wolfe, D. A., 1973) test. For each time point, the p-value distribution is approximated to a BUM distribution, and afterwards p-values are categorized in c categories, according to the cutting points given by the vector $\mathbf{l}$. We intend to discuss the results by comparing the ML estimates of the parameters α and λ for the BUM and C-BUM models, for different $\mathbf{l}$ vectors. We assess the loss of information by means of comparing the Fisher information matrices for the two models, for the different $\mathbf{l}$ vectors.

References

[1] Agresti, A. (2002) *Categorical Data Analysis.* Wiley Series in Probability and Statistics.

[2] Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, **16**, 1190-1208.

[3] Casella, G. and Berger, R. (1990). *Statistical Inference.* Wadsworth and Brooks/Cole.

[4] Hollander, M. and Wolfe, D. A. (1973) *Nonparametric Statistical Methods.* John Wiley and Sons.

[5] Kruskal, W. and Wallis, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistics Association*, **53**, 814–861.

[6] Lobenhofer, E. K., Bennett, L., Cable, P. A., Li, L., Bushel,P. R. and Afshari, C. A. (2002). Regulation of DNA Replication Fork Genes by 17A-Estradiol. *Molecular Endocrinology*, **16**, 1215-1229.

[7] Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B*, **44**, 226–233.

[8] Pounds, S. and Morris, S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–42.

[9] Titterington, D., Smith, A. and Makov, U. (1985). *Statistical analysis of finite mixture distributions.* Wiley.

[10] Woodburry, M. A. (1971). Discussion of paper by Hartley and Hocking. *Biometrics*, **27**, 808–817.