BAYESIAN MIXTURE MODELS WITH UNKNOWN NUMBER OF COMPONENTS: APPLICATION TO POWER CALCULATION IN MICROARRAY EXPERIMENTS

R. Rekaya

Department of Animal and Dairy Science
University of Georgia, Athens, GA 30602, USA

INTRODUCTION

Research groups form an increasingly diverse range of fields, are becoming involved in the task of designing, gathering and analyzing gene expression data produced by microarray experiments. Although large data sets were generated in the last years, little attention has been given to the statistical requirements of the analysis of such data. Most research done with gene expression data has focused on the development of visualization tools, and standard statistical methods such as cluster analysis and principal components have been applied. These techniques have been useful to summarize information, to identify clusters or groups of genes based on similarity or dissimilarity, and to predict biochemical and physiological pathways for some uncharacterized genes. However, important issues such as experimental design, number of replicates and the power of detecting change of expression have not received much attention if any. Comparison of gene expression patterns of tissues or cells under several conditions provides important information to answer several biological questions. Using the simple fold changes in expression based on the ratio of intensities in the red and green channels, as it has been done in the earlier days, is unreliable and inefficient (Pan *et al*., 2001). Gene expression data is a noisy one and the challenge now is to design and develop methods that allow the detection of the genuine changes.

In this study a mixture normal model with unknown number of components was implemented using Birth-death algorithm (Stephens, 2000). The objective was to calculate the power of detecting specified fold change using gene expression data.

MATERIAL AND METHODS

Data used in this study consisted on the expression levels of 8150 cDNA of individuals with and without cutaneous malignant melanoma. Although the original data set consisted of 38 arrays (31 melanoma and 7 controls), for demonstration purpose, we used only eight arrays (4 melanoma and 4 controls). The data is publicly available on line at www.ncbi.nlm.nih.gov. For a full description of the original data see Bittner *et al*. (2000). A sort of global normalization was applied to the raw data. The observed gene expression levels were transformed to the logarithmic scale and for each microarray, the transformed expression levels were standardized by subtracting their median.

Model. Let $\mathbf{X}_i = (X_{1i},...,X_{4i})$ and $\mathbf{Y}_i = (Y_{1i},...,Y_{4i})$ are expression levels for gene $i = (1,....,8150)$ in the four melanoma samples and the four control samples, respectively.

The following model was assumed:

$X_{ji} = \mu_{1i} + e_{ij}$ and $Y_{ki} = \mu_{2i} + \varepsilon_{ki}$

Where μ_{1i} and μ_{2i} are the mean expression levels for gene i for the two groups of individuals, respectively, and

$E(e_{ji}) = E(\varepsilon_{ki}) = 0$

$Var(e_{ji}) = \sigma_{1i}^2$ $Var(\varepsilon_{ki}) = \sigma_{2i}^2$

Having as objective to calculate the power and to detect all genes with $\mu_{1i} \neq \mu_{2i}$. Following the methodology proposed by Pan *et al.* (2001), two scores were constructed as:

$$z_i = \frac{\mathbf{X}_i \mathbf{a}_i / 4}{\sigma_{1i}^2} + \frac{\mathbf{Y}_i \mathbf{b}_i / 4}{\sigma_{2i}^2} \text{ and } Z_i = \frac{\bar{X}_i}{\sigma_{1i}^2} - \frac{\bar{Y}_i}{\sigma_{2i} 2}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are two column vectors of random permutation of 2 1's and 2 –1's and $\bar{X}_i$ and $\bar{Y}_i$ are the sample means.

Suppose that f_0 and f_1 are the probability density function for $z_i's$ and $Z_i's$, respectively. Obviously those two probability density functions are unknown but they can estimated based on $z_i's$ and $Z_i's$. It will be possible to detect genes with altered expression simple by comparing f_0 and f_1.

Mixture model to estimate f_0 and f_1. A finite normal mixture model was used:

$$f_0(z) = \sum_{i=1}^{k} \pi_i \phi(\eta_i, I\lambda_i)$$

where $\phi(\eta_i, \lambda_i)$ denotes the normal density with mean η_i and variance λ_i, π_i are the mixing proportions subject to $\sum_{i=1}^{k} \pi_i = 1$ and k is the number of components in the mixture (unknown).

Let $\boldsymbol{\theta}_i = (\eta_i, \lambda_i)'$ and $s = \{(\pi_1, \boldsymbol{\theta}_1), ..., (\pi_k, \boldsymbol{\theta}_k)\} \in \Omega_k$ to represent all parameters in the model. Applying Bayes theorem,

$$p(\boldsymbol{\theta}, k, \boldsymbol{\pi} \mid \mathbf{z}) \propto p(\mathbf{z} \mid k, \boldsymbol{\pi}, \boldsymbol{\theta}) p(k, \boldsymbol{\pi}, \boldsymbol{\theta} \mid \mathbf{w})$$
$$\propto p(\mathbf{z} \mid k, \boldsymbol{\pi}, \boldsymbol{\theta}) p(k \mid \mathbf{w}) p(\boldsymbol{\theta}_1 \mid \mathbf{w}) .... p(\boldsymbol{\theta}_k \mid \mathbf{w}) \qquad 0 \leq \pi_i \leq 1 \ \ (i = 1, ..., k)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, ..., \boldsymbol{\theta}'_k)'$, $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_k)'$ and $\mathbf{w}$ is a vector of known hyper-parameters.

Before deriving the needed conditional distributions, we will describe shortly the birth–death algorithm used to sample the number of component in the mixture (for a detailed information, see Stephens, 2000).

Birth-Death algorithm. If at a time t the process is at $s \in \Omega_k$ and a birth is said to occur at point $(\pi_b, \boldsymbol{\theta}_b)$, then the process jumps to:

$s \cup (\pi_b, \boldsymbol{\theta}_b) = \{(\pi_1(1-\pi_b), \boldsymbol{\theta}_1), ..., (\pi_k(1-\pi_b), \boldsymbol{\theta}_k), (\pi_b, \boldsymbol{\theta}_b)\} \in \Omega_{k+1}$

If at a time t the process is at $s \in \Omega_k$ and a death is said to occur at point $(\pi_i, \boldsymbol{\theta}_i)$ (one of the existing component in the mixture), then the process jumps to:

$$s \setminus (\pi_b, \boldsymbol{\theta}_b) = \{(\frac{\pi_1}{(1-\pi_i)}, \boldsymbol{\theta}_1), ..., (\frac{\pi_{i-1}}{(1-\pi_i)}, \boldsymbol{\theta}_{i-1}), (\frac{\pi_{i+1}}{(1-\pi_i)}, \boldsymbol{\theta}_{i+1}), ..., (\frac{\pi_k}{(1-\pi_i)}, \boldsymbol{\theta}_k)\} \in \Omega_{k-1}$$

Given the definition stated above, a birth (death) will increase (decrease) the number of components in the mixture by one.

Assume that $\beta(s)$ is the over all rate of birth and that a birth at a point $(\pi_b, \boldsymbol{\theta}_b)$ occurs according to a density $b(s;(\pi_b, \boldsymbol{\theta}_b)) = k(1-\pi_b)^{k-1} p(\boldsymbol{\theta}_b \mid \mathbf{w})$. Similarly, we assume that a death at each point $(\pi_i, \boldsymbol{\theta}_i)$ occurs with a rate given by: $\delta_i(s) = \beta(s) \frac{L(s \setminus (\pi_i, \theta_i))}{L(s)} \frac{p(k-1 \mid \mathbf{w})}{kp(k \mid \mathbf{w})}$ such that the over all rate of death is $\delta(s) = \sum_i \delta_i(s)$. $L(s)$ is the likelihood function evaluated at the current values of the parameters on *s*.

Having the over all rates of birth $\beta(s)$ *and death* $\delta(s)$*, the next jump of the process will be a birth with probability,* $\Pr(birth) = \frac{\beta(s)}{\beta(s)+\delta(s)}$ *or a death with probability* $\Pr(death) = \frac{\delta(s)}{\beta(s)+\delta(s)}$.

If the jump was a birth, the point $(\pi_b, \boldsymbol{\theta}_b)$ at which the birth takes place will be simulated from the density $b(s;(\pi_b, \boldsymbol{\theta}_b)) = k(1-\pi_b)^{k-1} p(\boldsymbol{\theta}_b \mid \mathbf{w})$ (we simulate π_b and θ_b). However, if the next jump is a death, one component will be eliminated with probability equal to $\frac{\delta_i(s)}{\delta(s)}$.

Returning to our model, the following priors were used:

$p(k) \propto \frac{\alpha^k}{k!}$ with ($k = (1,2,...,k_{\max} = 10)$;

$\beta(s) \propto b = 3$; $p(\boldsymbol{\eta} \mid w) \sim N(\eta_0, \Sigma_0)$

$p(\lambda \mid w) \sim \chi^{-2}(2, s_0)$; $p(\pi \mid t) \sim Dir(t)$

where $\eta_0, \Sigma_0, \alpha, s_0$ and t are known hyper-parameters. *Dir(t)* denotes the Dirichlet distribution with parameter *t*.

For computational convenience, we made used of data augmentation technique where the missing data was $\mathbf{u} = (u_1, u_2, ..., u_n)$, such that u_j is an indicator of the mixture component from which the observation z_j was generated. Hence, $\Pr(u_j = i \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \pi_i$ with $(j = 1,2,...,n; i = 1,2,...,k)$

where $n = \sum_{i=1}^{k} n_i$ is the total number of observations and n_i is the number of observations in the component *i* of the mixture. Finally, the algorithm proceeds by sampling successively from the following conditional distributions:

$$p(u_j = i \mid \boldsymbol{\pi}, k, \boldsymbol{\theta}, \mathbf{z}) \propto \pi_i N(\eta_i, \lambda_i)$$

$$p(\eta_i \mid \boldsymbol{\pi}, k, \mathbf{u}, \boldsymbol{\eta}_{-i}, \lambda_i, \mathbf{z}) \sim N[(n_i\lambda_i^{-1} + \lambda_0^{-1})(n_i\lambda_i^{-1}\, z_i + \bar{\lambda}_0^{-1}\eta_{0i}), (n_i\lambda_i^{-1} + \lambda_0^{-1})]$$

$$p(\lambda_i \mid \boldsymbol{\pi}, k, \mathbf{u}, \boldsymbol{\eta}, \mathbf{z}) \sim \chi^{-2}(2 + n_i, [2s_0 + \sum_{j:u_j=i} (z_j - \eta_i)^2)])$$

$$p(\boldsymbol{\pi} \mid k, \mathbf{u}, \boldsymbol{\theta}, \mathbf{z}) \sim Dir(t + n_1, ..., t + n_k)$$

To the above four steps, an extra step using the Birth-death algorithm is used to simulate k, the number of components in the mixture. After estimating both f_0 and f_1, for each gene we check whether its score Z_i falls within the rejection region of f_0 (change of expression) or not (no change of expression). To do so, the rejection region of f_0 was determined for a given false positive rate α as: $\alpha = \int_{-\infty}^{-C_\alpha} f_0(z)dz + \int_{C_\alpha}^{\infty} f_0(z)dz = \sum_{i=1}^{k} \pi_i[\Phi_{\eta_i,\lambda_i}(-C_\alpha) + 1 - \Phi_{\eta_i,\lambda_i}(C_\alpha)]$

where $\Phi_{\eta_i,\lambda_i}(.)$ is the CDF for a normal density with parameters η_i and λ_i. The bisection method (Press *et al*., 1992) was used to obtain C_α. For a specified magnitude of expression change (d) and a false positive rate (α), the power was calculated as:

$$power(d, \alpha) = \sum_{i=1}^{k} \pi_i[\Phi_{\eta_i+d,\lambda_i}(-C_\alpha) + 1 - \Phi_{\eta_i+d,\lambda_i}(C_\alpha)]$$

RESULTS AND DISCUSSION

Posterior means of number of components in the mixture for f_0 and f_1 were 1.47 and 1.63, respectively. Those values suggest that there is no strong evidence against the parametric Gaussian model. However we expect some changes on the number of genes differentially expressed using a mixture model. For a false positive rate of 0.1%, the power was 0.16, 0.51 and 0.79 for 2, 3 and 4 fold change in expression, respectively. The following study has to be extend to the situation were several replicates are available not only to calculate the power of detecting a specified magnitude of change, but more importunately to estimate the number of replicates needed for precise inferences.

REFERENCES

Stephens, M. (2000) *Ann. Stat.* **28** : 40-74.

Pan, W. *et al*. (2001) Report 2001-012 .Division of Biostatistics, University of Minnesota.

Bittner, M. *et al*. (2000) *Nature* **406** : 536-540.

Press, W.H., Vetterling, W.T, Teukolsky, S.A. and Flannery, B.P. (1992) "Numerical Recipes". 2nd ed. Cambridge, New York.