Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 28 - 29, 2006, pp. 39-42

Bayesian Curve Clustering in Microarray

Kyeong Eun Lee[1], *Bani K. Mallick*[2]

Abstract

We propose a Bayesian model-based approach using a mixture of Dirichlet processes model with discrete wavelet transform, for curve clustering in the microarray data with time-course gene expressions.

Keywords : Curve clustering, Dirichlet mixture processes, Discrete wavelet transform, Microarray data.

1. Introduction

In this paper, we propose a mixture of Dirichlet processes model using discrete wavelet transform for curve clustering in microarray data as a fully Bayesian approach. In order to characterize these time-course gene expressions, we consider them as trajectory functions of time and gene specific parameters and obtain their wavelet coefficients by discrete wavelet transform. We then build cluster curves based using a mixture of Dirichlet processes prior. Each iteration of MCMC algorithm generates the cluster structure of these coefficients as a by-product (Escobar and West. 1998). Subsequently, the proposed models are applied to a yeast cell cycle microarray data set: Cho et al. (1998).

2. Wavelet Based Dirichlet Process Model

The proposed method looks for relevant clusters in the observed curves by the posterior sampling of the wavelet coefficients in Dirichlet process mixtures $DP(\alpha, G_0)$. The prior of covariance Σ is modified as in an example of normal structure in Escobar and West (1998) and assume the following hierarchical structure :

[1]Full Time Instructor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail : artlee@knu.ac.kr

[2]Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA. E-mail : bmallick@stat.tamu.edu

$$[\mathbf{Y}_i|\beta_i,\sigma^2] \sim N(\mathbf{X}\beta_i,\sigma^2 I),$$

$$[\beta_i|\mu,\Sigma] \sim DP(\alpha, MN(\mu,\sigma^2\Sigma)),$$

$$[\sigma^2] \sim \mathrm{IG}\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right),[\alpha] \sim \mathrm{G(a,b)}$$

where IG is Inverse Gamma distribution and G is Gamma distribution. $\Sigma = diag(\{v_{jk}\}, 0\le k\le 2^{j-1}, j_0\le j\le J)$ and is intended for shrinkage with

$$[v_{jk}]\sim IG\left(\frac{s_{jk}}{2},\frac{r_{jk}}{2}\right),\ 0\le k\le 2^{j-1}, j_0\le j\le J$$

where $r_{j.}$ and $s_{j.}$ are specified levelwise to maintain a mean of roughly $n2^{-cj}$ for some constant c. Here, the constant c models the decay in the average size of wavelet coefficients and, thus, the mean of the inverse gamma prior, $\mathrm{E}(v_{jk}) = s_{jk}/(r_{jk}-2), r_{jk}>2$ is specified to match this decay. For all k, fixing $r_{j.} = c+2$, we get $s_{j.} = cn2^{-cj}$.

The posterior distributions are as follows:

$$\begin{aligned}[\beta_i|\ Y,\ \beta_k, k\neq i,\sigma^2,\Sigma] &\propto \exp\left\{-\frac{\sigma^2}{2}\sum_{i=1}^{I}(\mathbf{Y}_i-\mathbf{X}\beta_i)'(\mathbf{Y}_i-\mathbf{X}\beta_i)\right\}\\ &\times\left(\frac{\alpha}{\alpha+I-1}MN(\mu,\sigma^2\Sigma)+\frac{1}{\alpha+I-1}\sum_{j\neq i}\delta(\beta_i|\beta_j)\right)\\ &\propto q_0 MN(\mu_i,\sigma^2\ \mathbf{V})+\sum_{j\neq I}q_j\delta(\beta_i|\beta_j)\end{aligned}$$

where $V=(\Sigma^{-1}+I)^{-1}, \mu_i = V(\Sigma^{-1}\mu+\ \mathbf{X}'\ \mathbf{Y}_i)$ and the weights q_j are defined as

$$q_0 \propto \alpha\phi(\mathbf{Y}_i|\ \mathbf{X}\mu,\sigma^2(I+\ \mathbf{X}'\Sigma\ \mathbf{X}))$$

$$q_k \propto \phi(\mathbf{Y}_i|\ \mathbf{X}\beta_k,\sigma^2 I)$$

subject to $\sum_{j\neq i}q_j = 1$, where $\phi(y|\theta,\gamma)$ is the multinormal density function of mean θ and covariance γ. Since the conditional probability of sampling a new β is proportional to q_0, if it is small relative to the sum of other q_j's, the number of distinct β_i's is also small and samples of β's change much. Let superscript * denote distinct values. Escobar and West (1998) used a "remixing algorithm" in order to avoid this problem by resampling β_j^* at each iteration, and to, additionally, improve the convergence.

$$[\beta_j^*|\ \mathbf{Y},\sigma^2,\Sigma]\propto MN(\mu_j^*,\sigma^2\ \mathbf{V}_j^*)\ \text{ for each } j=1,\cdots,I^*$$

where $V_j^* = (\Sigma^{-1} + |J(j)|I)^{-1}$, $\mu_j^* = V_j^*(\Sigma^{-1}\mu + \sum_{j \in J(j)} \mathrm{X}' \mathrm{Y}_j)$ and $J(j)$ is the index set of jth cluster.

Since

$$[\sigma^2, \beta | \mathrm{Y}] \propto \left(\frac{1}{2\sigma^2}\right)^{I \cdot T/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}(\mathrm{Y}_i - \mathrm{X}\beta_i)'(\mathrm{Y}_i - \mathrm{X}\beta_i)\right\} \times \left(\frac{1}{2\sigma^2}\right)^{I \cdot T/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\beta_i - \mu)'\Sigma^{-1}(\beta_i - \mu)\right\} \times \mathrm{IG}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right),$$

the full conditional distribution of σ^2 after integrating out β is

$$[\sigma^2 | \cdot] \sim \mathrm{IG}\left(\frac{\nu_1 + \mathrm{N}}{2}, \frac{\mathrm{S}}{2}\right)$$

where $\mathrm{S} = \sum_{i=1}^{I}(\mu'\Sigma^{-1}\mu + \mathrm{Y}_i' \mathrm{Y}_i - \mu_i' \mathrm{V}^{-1}\mu_i) + \nu_2$ and $N = I \cdot T$. In addition, with $(\beta_i | \Sigma, \sigma^2) \sim \mathrm{N}(\mu, \sigma^2\Sigma)$, the posterior distribution of scaling parameters v_{jk} are drawn as

$$(v_{jk} | \beta_i, \sigma^2) \sim \mathrm{IG}\left(\frac{\mathrm{s}_{jk}^*}{2}, \frac{\mathrm{r}_{jk}^*}{2}\right)$$

where $s_{jk}^* = I + s_{jk}$ and $r_{jk}^* = (\sigma^2)^{-1}\sum_{i=1}^{I}(\beta_{ik} - u_k)^2 + r_{jk}$. The precision parameter α in the Dirichlet process plays an important role in determining the number of clusters. Assuming a continuous prior density for $p(\alpha)$, Escobar and West (1995) provided a distribution of number of components through Antoniak (1974)'s results

$$p(I^* | \alpha, I) = c_I(I^*) I! \alpha^{I^*} \Gamma(\alpha)/\Gamma(\alpha + I), \quad I^* = 1, \cdots, I.$$

where $c_I = p(I^* | \alpha = 1, I)$ and $\Gamma(\cdot)$ is the Gamma function. According to the relationship between the Gamma function and the Beta function,

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha + I)} = \frac{(\alpha + I)\beta(\alpha + 1, I)}{\alpha\Gamma(I)},$$

where $\beta(\cdot, \cdot)$ is the Beta function, the $p(\alpha | I^*)$ can be written as follows:

$$p(\alpha | I^*) \propto p(I^* | \alpha)p(\alpha) \propto p(\alpha)\alpha^{I^*-1}(\alpha + I)\int_0^1 \eta^{\alpha}(1 - \eta)^{I-1} d\eta$$

and it can be considered as the marginal distribution (of α) from a joint distribution for α and a latent variable η such that

$$p(\alpha, \eta | I^*) \propto p(\alpha) \alpha^{I^*-1} (\alpha + I) \eta^{\alpha} (1-\eta)^{I-1}.$$

Therefore choosing $p(\alpha)$ to be $G(a, b)$, leads to

$$p(\alpha | I^*, \eta) \sim \pi_\eta G(a + I^*, b - \log(\eta)) + (1 - \pi_\eta) G(a + I^* - 1, b - \log(\eta)),$$

where $\frac{\pi_\eta}{1-\pi_\eta} = \frac{a + I^* - 1}{I(b - \log(\eta))}$. Next, η is updated as

$$p(\eta | \alpha, I^*) \propto \eta^{\alpha} (1-\eta)^{I-1} = B(\alpha + 1, I).$$

We apply our proposed hierarchical model to two yeast cell cycle data and check the model adequacy using the Bayesian Information Criterion, BIC,