Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 3, pp. 575~584

Curve Clustering in Microarray

Kyeong Eun Lee[1)]

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

After a novel technology, microarray, was introduced, many statistical issues have been raising including clustering. When gene expressions are observed by time or temperature, we want to cluster these gene expressions. In the setting of time-course gene expressions, clustering methods can be divided into two categories: non-model- based methods and model-based methods. Non-model-based methods are such as hierarchical clustering (Eisen et al., 1998), clustering using correlation (Chu et al., 1998), self-organizing maps (Tamayo et al., 1999). Mixture-effects model with B-splines (Luan et al., 2003) and hidden Markov model (Schliep et al., 2003) are considered methods in model-based methods. We are motivated by Wakefield et al. (2003) who modeled the trajectory as a function of time and gene specific parameters and clustered these curves based on gene specific parameter using a reversible jump MCMC. Wakefield et al. (2003) used a first-order random walk model for gene-based parameters in a sporulation data (Chu et al., 1998) and a mixture of periodic function model for the cell-cycle data (Spellman et al., 1998).

In this paper, we propose a mixture of Dirichlet processes model using discrete wavelet transform for curve clustering as a fully Bayesian approach. In order to characterize these time-course gene expressions, we consider them as trajectory

1) 1969 N. Hicks #102, Palatine, IL 60074, U.S.A.
E-mail : artlee1971@yahoo.com

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. Bayesian Hierarchical Model

2.1. Mixture of Dirichlet Processes

Ferguson (1974) defined a Dirichlet process (DP) for a Bayesian nonparametric approach as follows: Let μ be a finite non-null measure on (X, Ω) where X is a space and Ω is a σ-field of subsets. If for any $k \in \{1, 2, \cdots,\}$ and any measurable partition $(B_1, \cdots, B_k)$ of X,

$$(P(B_1), \cdots, P(B_k)) \sim \text{Dirichlet}(\mu(\text{B}_1), \cdots, \mu(\text{B}_\text{k}))$$

then a stochastic process P is defined as a Dirichlet process $DP(\mu)$ on (X, Ω) with parameter μ. Especially, when $\mu = \alpha G_0$ with G_0 is distribution, $E(P(B)) = G_0(B)$ and $Var(P(B)) = \frac{G_0(B)(1 - G_0(B))}{\alpha + 1}$ for any $B \in \Omega$. In this case, G_0 is called the base measure, α is called the precision parameter, and DP is denoted by $DP(\alpha, G_0)$ to acknowledge the dependence μ through α and G_0. Escobar and West (1998) explored Dirichlet process $DP(\alpha, G_0)$ in order to model the "uncertainty" of the prior distribution, while referring to G_0 as the "location" distribution of Dirichlet process prior. We are interested in the following property of Dirichlet process: given a set of $\beta = \{\beta_1, \cdots, \beta_I\}$ from a random distribution G following a Dirichlet process $DP(\alpha, G_0)$, the conditional distribution

$$\beta_i | \beta_{-i} \sim \frac{\alpha}{\alpha + I - 1} G_0 + \frac{1}{\alpha + I - 1} \sum_{j \neq i} \delta(\beta_i | \beta_j), \ \beta_{-i} = \{\beta_j \in \beta : j \neq i\}$$

where $\delta(a | b)$ is a point mass giving probability 1 if $a = b$, follows a mixture of Dirichlet processes. Formal definition of a mixture of Dirichlet Processes can be found in Antoniak (1974).

An important property in the MDP model is that with positive probability some of the β_i have the same value because of the discreetness of random measure of MDP (MacEachern et al., 1998) and clustered property of data. Escobar and West (1998) point out the Polya urn representation for the joint posterior distribution of

$[\beta_i|\ \cdot\ ]$

$$[d\beta_i|\ \boldsymbol{Y}, \beta_{-i},\ \cdot\ ] \propto \prod_{i=1}^{n} f(\boldsymbol{Y}_i|\ \beta_i,\ \cdot\ ) \frac{\alpha G_0(d\beta_i|\ \sigma^2,\ \boldsymbol{V}) + \sum_{k \neq i} \delta(\beta_i|\ \beta_k)}{\alpha + i - 1}$$

2.2. Wavelet Regression

For the ith gene, Y_{it} is the normalized log-ratio of mRNA gene expression level relative to the gene expression of the reference cell at time t, where $i \in \{1, \cdots, I\}$ and $t \in \{1, \cdots, T\}$; I is the number of genes and T is total number of equally spaced time points. We assume that $\boldsymbol{Y}_i = (Y_i, \cdots, Y_{iT})'$ is the vector of observations of the ith trajectory function with additive white noise

$$Y_{it} = f(\theta_i, t) + \epsilon_{it}, \epsilon_{it} \sim N(0, \sigma^2)$$

where $f(\theta_i, t)$ is a trajectory function of a gene-specific set of parameter, θ_i and time t (Wakefield et al., 2003). The trajectory function can be represented in terms of shifted and dilated scale functions $\{\psi(t)\}$ and wavelet functions $\{\phi(t)\}$ as follows:

$$f(t) \approx \sum_{k=0}^{2^{j_0}-1} s_k \phi_{j_0 k}(t) + \sum_{j=j_0}^{J} \sum_{k=0}^{2^j-1} w_{jk} \psi_{jk}(t)$$

where $J = \log_2 T$, $j \geq j_0 \geq 0$, $u_k = < f, \phi_{j_0 k} >$ and $w_{jk} = < f, \psi_{jk} >$ (Daubechies, 1992) or equivalently we may write the model as

$$f = \mathbf{X}\beta$$

where β is the wavelet representation of the true function f and $\mathbf{X}'$ is the $T \times T$ orthogonal wavelet transformation. In this paper, we only assume orthogonal wavelet bases and avoid more general representations for questions of stability and bias introduced in estimation.

2.3. Generic 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 :

$$[\mathbf{Y}_i|\ \beta_i, \sigma^2] \sim \mathrm{N}(\mathbf{X}\beta_i, \sigma^2 I),$$
$$[\beta_i|\ ]\mu, \Sigma] \sim DP(\alpha, \mathrm{MN}(\mu, \sigma^2\Sigma)),$$
$$[\sigma^2] \sim \mathrm{IG}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right),$$
$$[\alpha] \sim \mathrm{G}(a, b)$$

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, $\mathrm{E}(v_{jk}) = s_{jk}/(r_{jk}-2), r_{jk} > 2$, $r_{jk} > 2$ 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:

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

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

$$q_0 \propto \alpha\phi(\mathbf{Y_i}|\ \mathbf{X}\mu, \sigma^2(\mathrm{I} + \mathbf{X}'\Sigma\mathbf{X}))$$
$$q_k \propto \phi(\mathbf{Y_i}|\ \mathbf{X}\beta_k, \sigma^2\mathrm{I})$$

subject to $\sum_{j\neq i} q_j = 1$, where $\phi(y|\ \theta, \Upsilon)$ is the multinormal density function of mean θ and covariance Υ. 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.

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

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

Since

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

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 $\mathbf{S} = \sum_{i=1}^{I}(\mu'\Sigma^{-1}\mu + \mathbf{Y}_i'\mathbf{Y}_i - \mu_i'\mathbf{V}^{-1}\mu_i) + \nu_2$ 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

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

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

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

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:

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

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

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

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

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

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

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

3. Application to cDNA Microarray Data

We apply our proposed hierarchical model to two yeast cell cycle data and check the model adequacy using the Bayesian Information Criterion, BIC, introduced by Schwarz (1978),

$$\mathrm{BIC} = -2 \cdot \log \hat{L} + \log(n) \cdot p$$

where $\hat{L}$ is the maximized likelihood and p is the number of parameter in the model. If the ratio of BIC of two models is considered, it is known to provide an approximation of $2\log(\text{Bayes Factor})$ for sufficiently large sample n (Schwarz, 1978).

3.1. The Yeast Cell Cycle Data

Cho *et al.* (1998) obtained time courses of more than 6000 genes over nearly two full cell cycles (17 time points) and found that 416 among them have periodic- fluctuations in the gene expression due to their cell cycle-dependency. Genes were categorized into five-phases, early G1, late G1, S, G2 and M by their different peak points. Cho *et al.* (1998) identified that 33 of 416 genes have peak points in two different phases. So we used 384 genes for application to our proposed model and compare our clustering results with their identified phases. The data was log transformed and standardized across the cell cycle.

We used the last 16 time points for the computational convenience in the

Discrete Wavelet Transform. We compared our clustering result with Cho et al. (1998) using the adjusted Rand index (Hubert and Arabie, 1985), which evaluates the measure of agreement of two different partitions of one data set and overcomes the problem of non-constant expected value of Rand index (Rand, 1971), the fraction of agreement. Especially Milligan and Cooper (1986) recommended the adjusted Rand index as an external measure after extensive comparisons. If the cluster is random, the expected value of adjusted Rand index is 0. Its maximum value is 1 which indicates the perfect match between the clustering result and the external standard. The more details on the adjusted Rand index is referred to Yeung and Ruzzo (2000).

Table 1 shows the comparison of two partitions. The adjusted Rand index based on Table 1 is 0.4563 and we can see how sharply they are clustered. BIC of the model (=2294.8) is much lower than the BIC of Cho et al.'s clustering (=3659.7). Cho et al. (1998) classified these genes based on their peak time but our proposed model is based on the trajectory pattern. So it may be the reason of relative low adjusted Rand index and the lower BIC of our model supports that our proposed model clusters these curves sharply. In addition, the distribution of I^*do not change much in the same analysis with various priors of α; $G(0.0001, 0.0001), G(1,1), G(2,1)$ and so on.

Table 1. Two Partitions of Yeast Cell Cycle Data (C: Clusters by Cho *et al.* (1998) and D: Clusters by Our Proposed Model)

Class	$C1$	$C2$	$C4$	$C5$	$C3$	Sums
$D1$	40	14	0	0	1	55
$D2$	7	117	0	0	38	162
$D5$	0	1	41	3	35	80
$D4$	3	0	7	37	1	48
$D3$	17	3	4	15	0	39
Sums	67	135	52	55	75	

Figure 1 shows five curve clusters of 384 genes by Cho et al. (1998) and Figure 2 shows those generated by our proposed model. Compared with Figure 1, our model shows clearer classification schemes $(C1 - D3, C2 - D1, C3 - D5, C4 - D2, C5 - D4)$.

Figure 1. Five Clusters of Expression Time Courses in Yeast Data (Cho *et al.*, 1998)

Figure 2. Five Clusters of Expression Time Courses by Our Proposed Model in Yeast Data

4. Discussion

We have proposed a Bayesian model for curve clustering and identified genes which has similar trajectory function over time. We have used a Bayesian hierarchical model and Dirichlet process prior of discrete wavelet coefficients. And by product of it, we obtained clustering result in each iteration and we used the marginal posterior mode of the clustering membership of genes. Additionally, we easily estimated the missing data using the conditional distribution. Finally, we finish this paper by pointing out two potential shortcomings of the proposed procedure. First, it should be remarked that the procedure of this paper can not be applied to classification problem, known the number of clusters. Second, if the time is not equally spaced, we could not use the discrete wavelet transform directly, however, the latter can be handled by a lifting technique.

References

1. Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Annals of Statistics,* 2, 1152-1174.
2. Chu, R. J., Campbell, M. J. *et al.* (1998). A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell,* 2. 65-73.
3. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. and Herskowitz, I. (1998). The Transcriptional Program of Sporulation in Budding Yeast. *Science,* 282, 699-705.
4. Daubechies, I. (1992). Ten Lectures on Wavelets. Philadelphia: *SIAM*.
5. Eisen, M.B. *et al*. (1998). Cluster Analysis and Display of Genome-wide Expression Patterns. *Proceedings of the National Academy of science USA,* 95, 14863-14868.
6. Escobar, M.D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association,* 90, 577-588.
7. Escobar, M.D. and West, M. (1998). Computing Nonparametric Hierarchical Models. P. Muller D.D. Dey and D. Sinha, editors, In Practical Nonparametric and Semiparametric Bayesian Statistics, Lecture Notes in Statistics no. 133, New York: *Springer-Verlag*.
8. Ferguson, T.S. (1974). Prior Distributions on Spaces of Probability Measures. *Annals of Statistics,* 2, 615-629.
9. Hubert, L., and Arabie, P. (1985). Comparing Partitions. J*ournal of Classification,* 2, 193-218.

10. Luan, Y. and Li, H. (2003). Clustering of Time-Course Gene Expression Data Using a Mixed-Effects Model with B-Splines. *Bioinformatics,* 19, 474-482.
11. MacEachern, S. N., and Muller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics,* 7, 223-238.
12. Milligan, G. W. and Cooper, M. C. (1986). A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behavioral Research,* 21, 441-458.
13. Rand, W.M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association,* 66, 846-850.
14. Schliep, A., Schonhuth, A. and Steinhoff, C. (2003). Using Hidden Markov Models to Analyze Gene Expression Time Course Data. *Bioinformatics,* 19, 255-263.
15. Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* 6, 461-464.
16. Spellman, P. T., Sherlock, G., Zhang, M. Q. *et al.* (1998). Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization. *Mol. Biol. Cell.*, 9, 3273-3297.
17. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999). Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation. *Proceedings of the National Academy of science USA,* 96, 2907-2912.
18. Wakefield, J., Zhou, C. and Self, S. (2003). Modelling Gene Expression over Time: Curve Clustering with Informative Prior Distributions. *Bayesian Statistics 7*, Proceedings of the Seventh Valencia International Meeting, Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M. and West, M. (editors), Oxford: Oxford University Press.
19. Yeung, K. and Ruzzo, W. (2001). Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics,* 17(9), 763-774.

[received date : Jun. 2004, accepted date : Aug. 2004]