Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 231897, 4 pages
doi:10.1155/2008/231897

Research Article

Bayesian Functional Data Clustering for Temporal Microarray Data

Ping Ma,[1] Wenxuan Zhong,[1] Yang Feng,[1] and Jun S. Liu[2]

[1] *Department of Statistics, University of Illinois, Champaign, IL 61820, USA*
[2] *Department of Statistics, Harvard University, Cambridge, MA 02138, USA*

Correspondence should be addressed to Ping Ma, pingma@uiuc.edu and Jun S. Liu, jliu@stat.harvard.edu

Received 24 July 2007; Accepted 18 December 2007

Recommended by Nengjun Yi

We propose a Bayesian procedure to cluster temporal gene expression microarray profiles, based on a mixed-effect smoothing-spline model, and design a Gibbs sampler to sample from the desired posterior distribution. Our method can determine the cluster number automatically based on the Bayesian information criterion, and handle missing data easily. When applied to a microarray dataset on the budding yeast, our clustering algorithm provides biologically meaningful gene clusters according to a functional enrichment analysis.

1. INTRODUCTION

Microarray technology enables the scientist to measure the mRNA expression levels of thousands of genes simultaneously. For a particular species of interest, one can make microarray measurements under many different conditions and for different types of cells (if it is a multicellular organism). Genes' expression profiles under these conditions often give the scientist some clues on biological roles of these genes. A group of genes with similar profiles are often "coregulated" or participants of the same biological functions.

When a series of microarray experiments are conducted sequentially during a biological process, we call the resulting dataset a "temporal" microarray dataset, which can provide insights on the underlying biology and help decipher the dynamic gene regulatory network. Clustering genes with similar temporal profiles is a crucial first step to reveal potential relationships among the genes.

Conventional clustering methods, such as the K-means and hierarchical clustering, do not take into consideration the correlation in the gene expression levels over time. Although it is possible to use a general multivariate Gaussian model to account for the correlation structure, such a model ignores the time order of the gene expressions. As evidenced in our example, the time factor is important in interpreting the results of gene expression clustering in temporal data. It is also possible to use an autoregression model to describe the gene expression time series, but such a model often requires stationarity, which is unlikely to hold in most temporal microarray data.

Recently, nonparametric analysis of data in the form of curves, that is, functional data, is subject to active research, see [1, 2] for a comprehensive treatment of functional data analysis; and curve-based functional clustering methods have emerged [3–7], but a rigorous assessment of the estimation precision is still lacking.

In this paper, we propose a Bayesian clustering method, which optimally combines the available information and provides a proper uncertainty measure for all estimated quantities. Our method is based on a mixture of mixed-effect smoothing splines models. For each cluster, we model its mean profile as a smoothing spline function and describe its individual gene's variation by a parametric random effect. Based on the theory of reproducing-kernel Hilbert spaces [8], we represent the mean expression curve as a linear combination of certain basis functions, which enables us to derive the full posterior distribution up to a normalizing constant. All the conditional distributions needed by a Gibbs sampler are also easy to compute and to sample from. Our method automatically takes care of the missing data and infers the number of clusters in the data. Using the method,

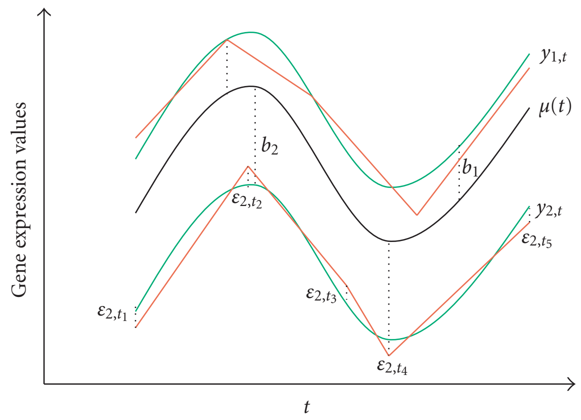

Figure 1: A smoothing-spline mixed effect model for temporal gene expression.

we analyzed a microarray dataset of budding yeast, we found that the majority of the clusters we had obtained are enriched for known and expected biological functions.

Our method is not restricted to temporal microarray data, and can be applied to all curve clustering problems, especially for sparsely and irregularly sampled temporal data.

2. MATERIAL AND METHODS

2.1. Mixed-effect representation of gene expression profile

Let the expression value of the ith gene at time t be y_{it}. To accommodate missing data that occasionally occurs in microarray experiment, we denote $\mathbf{t}_i = (t_1, \dots, t_{n_i})$ and $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^T$, where n_i is the number of measurements of ith gene. Our mixed-effect smoothing spline model [9] for genes in one cluster is

$$\mathbf{y}_i = \boldsymbol{\mu}(\mathbf{t}_i) + Z_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{\mu}(\mathbf{t}_i) = (\mu(t_1), \dots \mu(t_{n_i}))^T$ is the cluster's mean profile, $\mathbf{b}_i \sim N(0, B)$ is the random effect to capture the intragene correlation, Z_i is the known design matrix for the random effect, and $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2 I)$ is the random error independent of $\mathbf{b}$ and of each other.

By taking different $\mathbf{b}$ vectors, we can accommodate different nonrandom effects. For example, when $\mathbf{b}_i = b_i$ and $Z_i = \mathbf{1}$, the expression profile of the ith gene is parallel to the mean profile $\boldsymbol{\mu}$ (Figure 1). If $\mathbf{b}_i = (b_{i1}, b_{i2})^T$ and $Z_i = (\mathbf{1}, \mathbf{t}_i)$, the difference between the ith gene profile and the mean profile is a linear function in time. More complicated structures such as periodicity can be modeled by letting the Z_i be basis of a certain functional space.

By considering μ in a reproducing kernel Hilbert space $\mathcal{H} \subseteq \{\mu : M(\mu) < \infty\}$ in which $M(\mu)$ is a square seminorm, we can represent μ as

$$\mu(t) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{i=1}^{q} c_j R_M(s_j, t), \qquad t \in [0, a], \tag{2}$$

where $\{s_j\}$ is a set consisting of all distinct $\{t_i\}$, q is the number of $\{s_j\}$, and R_M is the kernel of $\mathcal{H}$. The choice of $M(\mu) = \int_0^a (d^2\mu/dt^2)^2 dt$ yields the cubic smoothing spline with

$$\phi_1(t) = 1, \qquad \phi_2(t) = t, \tag{3}$$

$$R_M(t_1, t_2) = \int_0^a (t_1 - u)_+ (t_2 - u)_+ du, \tag{4}$$

where $(\cdot)_+ = \max(\cdot, 0)$ [10].

Writing (2) in a vector-matrix form, we have

$$\boldsymbol{\mu}(\mathbf{t}_i) = S_i\mathbf{d}_i + R_i\mathbf{c}_i, \tag{5}$$

where S_i is $n_i \times m$ with the (i, ν)th entry $\phi_\nu(t_i)$ and R is $n_i \times q$ with the (i, j)th entry $R_M(t_i, s_j)$. Substituting (5) into (1), we have

$$\mathbf{y}_i = S_i\mathbf{d}_i + R_i\mathbf{c}_i + Z_i\mathbf{b}_i + \boldsymbol{\epsilon}_i. \tag{6}$$

Denoting $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T$ and S, R, Z, $\boldsymbol{\epsilon}$ similarly, we have the matrix representation

$$\mathbf{y} = S\mathbf{d} + R\mathbf{c} + Z\mathbf{b} + \boldsymbol{\epsilon}, \tag{7}$$

where $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_n^T)^T \sim N(0, \text{diag}(B, \dots, B))$.

The prior distributions are specified as follows:

$$\begin{aligned} \mathbf{d} &\sim N(0, \text{diag}(\delta_1, \dots, \delta_m)), \\ \mathbf{c} &\sim N(0, \tau^2 I), \\ \sigma^2 &\sim \text{IG}(\alpha_{\sigma^2}, \beta_{\sigma^2}), \\ \tau^2 &\sim \text{IG}(\alpha_{\tau^2}, \beta_{\tau^2}), \\ B &\sim \text{IW}(\nu_0, S_0^{-1}), \end{aligned} \tag{8}$$

where IG and IW are inverse-Gamma and inverse-Wishart distributions, respectively.

These priors lead to the following full conditional posteriors, which are used in our Gibbs sampler:

$$\begin{aligned} [\mathbf{d} \mid \mathbf{b}, \mathbf{c}, \sigma^2, \boldsymbol{\delta}, \mathbf{y}] &\sim N(V_d S^T(\mathbf{y} - R\mathbf{c} - Z\mathbf{b})/\sigma^2, V_d), \\ [\mathbf{c} \mid \mathbf{d}, \mathbf{b}, \sigma^2, \tau^2, \mathbf{y}] &\sim N(V_c R^T(\mathbf{y} - S\mathbf{d} - Z\mathbf{b})/\sigma^2, V_c), \\ [\mathbf{b} \mid \mathbf{d}, \mathbf{c}, \sigma^2, B, \mathbf{y}] &\sim N(V_b Z^T(\mathbf{y} - S\mathbf{d} - R\mathbf{c})/\sigma^2, V_b), \\ [B \mid \mathbf{b}] &\sim \text{IW}\Big(\nu_0 + n, \Big(S_0 + \sum_{i=1}^{n} \mathbf{b}_i\mathbf{b}_i^T\Big)^{-1}\Big), \\ [\tau^2 \mid \mathbf{c}] &\sim \text{IG}(\alpha_{\tau^2} + (q - m)/2, \beta_{\tau^2} + \mathbf{c}^T\mathbf{c}/2), \\ [\sigma^2 \mid \mathbf{d}, \mathbf{b}, \mathbf{c}, \mathbf{y}] &\sim \text{IG}(\alpha_{\sigma^2} + n/2, \beta_{\sigma^2} + \text{SSR}), \end{aligned} \tag{9}$$

where $V_d = (S^T S/\sigma^2 + \text{diag}(\delta_1^{-1}, \dots, \delta_m^{-1}))^{-1}$, $V_b = (Z^T Z/\sigma^2 + \text{diag}(B^{-1}, \dots, B^{-1}))^{-1}$, $V_c = (R^T R/\sigma^2 + 1/\tau^2 I)^{-1}$, and SSR $= (\mathbf{y} - S\mathbf{d} - R\mathbf{c} - Z\mathbf{b})^T(\mathbf{y} - S\mathbf{d} - R\mathbf{c} - Z\mathbf{b})$.

2.2. The mixture model with unknown number of components

When more than one cluster is considered, we assume that the expression of the ith gene has a Gaussian mixture distribution:

$$\mathbf{y}_i \sim p_1 N(\boldsymbol{\mu}_1, \Sigma_1) + \cdots + p_K N(\boldsymbol{\mu}_K, \Sigma_K), \tag{10}$$

where $\boldsymbol{\mu}_k$ and $\Sigma_k = ZB_kZ^T + \sigma^2 I$ are the mean and covariance matrix for the kth component, as given by (7); p_k is the fraction of kth component, and K is the number of Gaussian components.

2.3. Class labels and cluster numbers

To ease the computation, we introduce a "latent" membership labeling variable J_i for the ith gene so that

$$\mathbf{y}_i \mid J_i = j \sim N(\boldsymbol{\mu}_j, \Sigma_j). \tag{11}$$

When the number of Gaussian components K is known, we can get the joint posterior probability as

$$P(\mathbf{J}, \boldsymbol{\mu}, \Sigma \mid \mathbf{y}) = \pi(\boldsymbol{\mu}, \Sigma) \prod_{i=1}^{n} p_{j_i} N\left(\mathbf{y}_i \mid \boldsymbol{\mu}_{j_i}, \Sigma_{j_i}\right), \tag{12}$$

where $\mathbf{J} = (j_1, \ldots, j_n)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $\Sigma = (\Sigma_1, \ldots, \Sigma_K)$, and $\pi(\boldsymbol{\mu}, \Sigma)$ is the joint prior distribution.

Since K is unknown, we used the following Bayesian information criterion (BIC):

$$\text{BIC} = -2 \log p(\mathbf{y} \mid M_K, \widehat{\boldsymbol{\theta}}_K) + l_K \log n, \tag{13}$$

where M_K is the current model with parameters $\boldsymbol{\theta}_K$, $\widehat{\boldsymbol{\theta}}_K$ is the estimate, and l_K is the total number of parameters in our model. A small BIC score indicates the adequacy of the corresponding model. An alternative to our current approach (i.e., each clustering configuration is equally likely given the number of clusters K, and K is determined by BIC) is to use a Polya Urn prior (also called the "Chinese restaurant" process), which postulates that when a new member comes in, its a priori probability for joining an existing cluster of size m_i is $(m_i + c)/(m + c)$, and for forming a new cluster of its own is $c/(m + c)$, where m is the total number of existing members. This prior, however, favors unbalanced cluster configurations (e.g., very large and very small clusters) and may not be appropriate in our applications.

2.3.1. Gibbs Sampling from the Posterior

To complete our Bayesian analysis, we employ the Dirichlet prior Di $(\alpha_1, \ldots, \alpha_K)$ for $(p_1, \ldots, p_K)$, the cluster proportions. Thus, given the cluster indicator $\mathbf{J}$, the posterior distribution of the p's is again a Dirichlet distribution.

Given $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, B_1, \ldots, B_K, \sigma^2$, we have the conditional distribution of J_i:

$$\begin{aligned} p(J_i = j \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, B_1, \ldots, B_K, \sigma^2, \mathbf{y}) \\ = \frac{p_j \mathbf{N}(\mathbf{y}_i \mid \boldsymbol{\mu}_j, ZB_jZ^T + \sigma^2 I)}{\sum_{k=1}^{K} p_k \mathbf{N}(\mathbf{y}_i \mid \boldsymbol{\mu}_k, ZB_kZ^T + \sigma^2 I)}. \end{aligned} \tag{14}$$

With an initial value of $\mathbf{J}$, which gives rise to a partition of $\mathbf{y} : (\mathbf{y}_1^J, \ldots, \mathbf{y}_K^J)$, and the initial values of $\mathbf{d}_k$, $\mathbf{b}_k$, $\mathbf{c}_k$, B_k, where $k = 1, \ldots, K$, as well as σ^2, we iterate the following iterative conditional sampling steps:

(i) for $i = 1, \ldots, n$, draw a new j_i from the conditional distribution from (14) to replace the old one;

(ii) conditional on $\mathbf{J}$, sequentially

   (a) update $\mathbf{d}_k$ by a draw from $[\mathbf{d}_k \mid \mathbf{b}_k, \mathbf{c}_k, \sigma^2, \boldsymbol{\delta}, \mathbf{y}_k^J]$, where $k = 1, \ldots, K$,

   (b) update $\mathbf{b}_k$ from $[\mathbf{b}_k \mid \mathbf{d}_k, \mathbf{c}_k, \sigma^2, B_k, \mathbf{y}_k^J]$, where $k = 1, \ldots, K$,

   (c) update $\mathbf{c}_k$ from $[\mathbf{c}_k \mid \mathbf{d}_k, \mathbf{b}_k, \sigma^2, \tau_k^2, \mathbf{y}_k^J]$, where $k = 1, \ldots, K$,

   (d) update $B_k \sim [B_k \mid \mathbf{b}_k]$, and $\tau_k^2 \sim [\tau_k^2 \mid \mathbf{c}_k]$, where $k = 1, \ldots, K$,

   (e) update $\sigma^2 \sim [\sigma^2 \mid \mathbf{d}, \mathbf{b}, \mathbf{c}, \mathbf{y}]$,

   (f) update $(p_1, \ldots, p_K) \sim \text{Di}(n_1 + \alpha_1, \ldots, n_K + \alpha_K)$, where n_j is the number of genes in the jth cluster.

3. RESULTS AND DISCUSSION

To study oxygen-responsive gene network, Lai et al. [11] used cDNA microarray to monitor the gene expression changes of wild-type budding yeast (*Saccharomyces cerevisiae*) under aerobic condition in galactose medium. Under aerobic condition, the oxygen concentration was lowered gradually until oxygen was exhausted during a period of ten minutes. Microarray experiments were conducted at 14 time points under aerobic condition. A reference sample was obtained from a pooled RNA collected from all time points for hybridization.

For the analysis, Lai et al. [11] normalized gene expression after time 0 to gene expression of time 0 to set the common starting point. They identified 2388 genes whose expression is differentially expressed at one or more time points. Using our method, 2388 genes was clustered to 31 clusters, which yields the smallest BIC. FunSpec [12] was used for gene annotation and biological function enrichment analysis, where the Bonferroni-corrected functional enrichment P-values based on hypergeometric distributions are reported. We found 23 clusters out of 31 clusters discovered have biological functions over-represented. Among them, estimated mean gene expression profiles of three clusters are given in Figure 2.

In cluster A, which consists of 40 genes, the estimated mean expression goes up progressively as oxygen level goes down, which suggests that the genes in this cluster were transiently upregulated in response to aerobisis. Accordingly, genes involved in stress response (function enrichment P-value $= 10^{-4}$) as well as cell rescue and defense are over-represented in this cluster (function enrichment P-value $= 10^{-4}$). Furthermore, genes involved in molecular functions of oxidoreductase and coproporphyrinogen oxidase are also presented, which explains the upregulation of the gene expression levels.

We have 92 genes in cluster B, where the estimated mean gene expression drops down at the beginning rapidly and then goes up gradually. In this cluster, 34 genes are involved in protein synthesis (function enrichment P-value $\leq 10^{-14}$). Moreover, ribosome biogenesis are also over-represented (function enrichment P-value $\leq 10^{-14}$). These processes were affected by oxygen level initially, but were

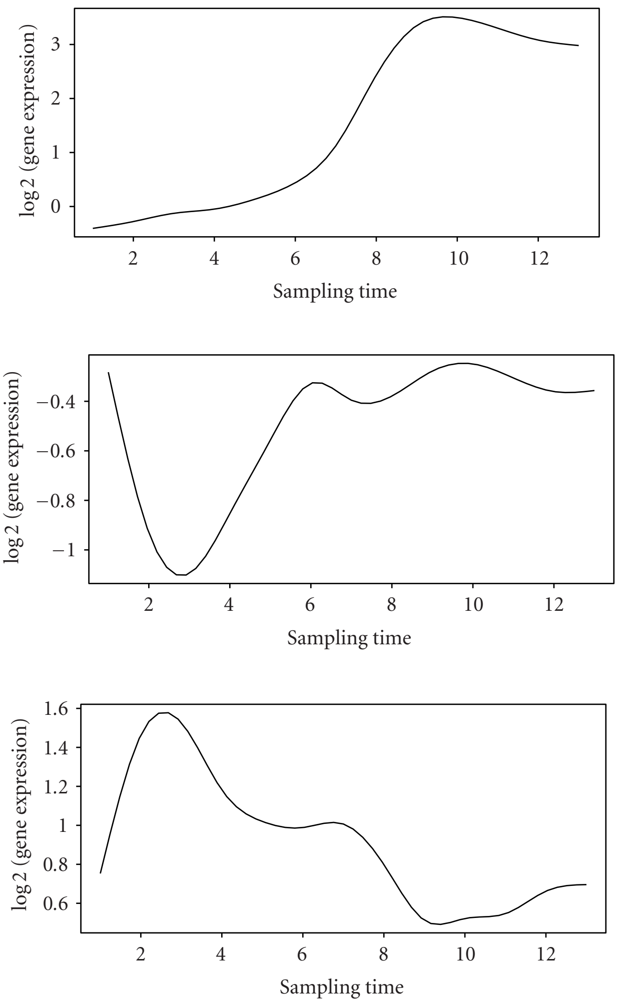

Figure 2: Estimated mean expression curves for cluster A, B, and C (from top to bottom) discovered in the yeast aerobic expression data.

quickly adjusted to high expression levels to maintain living of yeast.

Contrast to cluster B, cluster C (68 genes) consists of genes involved in galactose fermentation (function enrichment P-value $= 10^{-4}$), carbon utilization (functional enrichment P-value $= 10^{-2}$), and carbohydrate metabolism (function enrichment P-value $\leq 10^{-10}$). The initial upregulation of gene expression under aerobic condition can be partly explained by the fact that the cell increases the energy uptaking through the carbon utilization as oxygen level goes down; but as the oxygen level continues to drop down, these processes are replaced by the more energy-efficient processes, which drives the expression levels of genes to be downregulated.

4. CONCLUSIONS

Conventional clustering methods do not take into consideration the correlation in the gene expression levels over time. Multivariate Gaussian models and time series analysis cannot model the time factor and correlation properly. These limitations can be readily overcome by the full Bayesian approach developed here. Although certain prior distributions and the related hyperparameters need to be input by the user, we found the clustering results rather robust to variations in such inputs. Moreover, our Bayesian clustering algorithm serves as a platform to incorporate more biological knowledge. Open source R code is available at www.stat.uiuc.edu/~pingma/BayesianFDAClust.htm.

ACKNOWLEDGMENTS

The authors thank C. I. Castillo-Davis for his help in designing Figure 1 and for the many constructive suggestions and discussion in the early stage of this research. The authors thank Ji Young Kim for designing the software website. The authors also thank Kurt Kwast for providing the yeast microarray data.

REFERENCES

[1] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Springer, New York, NY, USA, 2005.

[2] J. O. Ramsay and B. W. Silverman, *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York, NY, USA, 2002.

[3] G. M. James and C. A. Sugar, "Clustering for sparsely sampled functional data," Journal of the American Statistical Association , vol. 98, no. 462, pp. 397–408, 2003.

[4] Y. Luan and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with B-splines," Bioinformatics , vol. 19, no. 4, pp. 474–482, 2003.

[5] Y. Luan and H. Li, "Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data," Bioinformatics , vol. 20, no. 3, pp. 332–339, 2004.

[6] N. A. Heard, C. C. Holmes, and D. A. Stephens, "A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves," Journal of the American Statistical Association , vol. 101, no. 473, pp. 18–29, 2006.

[7] P. Ma, C. I. Castillo-Davis, W. Zhong, and J. S. Liu, "A data-driven clustering method for time course gene expression data," Nucleic Acids Research , vol. 34, no. 4, pp. 1261–1269, 2006.

[8] G. Wahba, *Spline Models for Observational Data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, Pa, USA, 1990.

[9] C. Gu and P. Ma, "Optimal smoothing in nonparametric mixed-effect models," Annals of Statistics , vol. 33, no. 3, pp. 1357–1379, 2005.

[10] C. Gu, *Smoothing Spline ANOVA Models*, Springer, New York, NY, USA, 2002.

[11] L.-C. Lai, A. L. Kosorukoff, P. V. Burke, and K. E. Kwast, "Metabolic-state-dependent remodeling of the transcriptome in response to anoxia and subsequent reoxygenation in Saccharomyces cerevisiae," Eukaryotic Cell , vol. 5, no. 9, pp. 1468–1489, 2006.

[12] M. D. Robinson, J. Grigull, N. Mohammad, and T. R. Hughes, "FunSpec: a web-based cluster interpreter for yeast," BMC Bioinformatics , vol. 3, pp. 3–35, 2002.