CLUSTERING SAMPLES CHARACTERIZED BY TIME COURSE GENE EXPRESSION PROFILES USING THE MIXTURE OF STATE SPACE MODELS

OSAMU HIROSE[1] ochamu@ims.u-tokyo.ac.jp
RYO YOSHIDA[2] yoshidar@ism.ac.jp
RUI YAMAGUCHI[1] ruiy@ims.u-tokyo.ac.jp
SEIYA IMOTO[1] imoto@ims.u-tokyo.ac.jp
TOMOYUKI HIGUCHI[2] higuchi@ism.ac.jp
SATORU MIYANO[1] miyano@ims.u-tokyo.ac.jp

[1] *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan*
[2] *Institute of Statistical Mathematics, Research Organization of Information and Systems, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan*

We propose a novel method to classify samples where each sample is characterized by a time course gene expression profile. By exploiting the mixture of state space model, the proposed method addresses the following tasks: (1) clustering samples according to temporal patterns of gene expressions, (2) automatic detection of genes that discriminate identified clusters, (3) estimation of a restricted autoregressive coefficient for each cluster. We demonstrate the proposed method along with the cluster analysis of 53 multiple sclerosis patients under recombinant interferon β therapy with the longitudinal time course expression profiles.

Keywords: clustering; mixture of state space models; multiple sclerosis; time course gene expression profile.

1. Introduction

Time course gene expression profiles enable us to understand the temporal structure of gene regulations. A number of researchers have studied the time course gene expression profiles [1, 4, 5]. One major difficulty of time course gene expression analysis comes from the fact that length of time course is usually much smaller than the dimension of data. In order to overcome the difficulties related to imbalance between dimensionality and length of data, we developed a state space model in our previous works [8, 9]. The state space models provide an approach to avoid over-parameterization of the vector autoregressive model by exploring potential sets of the co-expressed genes.

Recently, novel kinds of time series expression profiles that existing methods may fail to analyze have been appeared. Baranzini *et al.* [2] investigated the longitudinal gene expression change of multiple sclerosis (MS) patients with treatments of recombinant interferon β (rIFN β). In this data set, each MS patient is characterized by a gene expression matrix whose column vectors represent gene expression vectors

for corresponding observed time points. They aimed at classifying 53 MS patients, composed of 33 good responders and 20 poor responders for the therapy of rIFN β. Hence, the problem is to classify samples, where each sample is characterized by matrix data.

It is possible to use classical clustering methods such as the hierarchical clustering and the k-means clustering, for example, by vectorizing gene expression matrices and measuring Euclid distances of all pairs of the vectorized matrices. These methods, however, often fail since they do not make use of time series of data effectively and in addition, the direct product between time and gene expands feature space and lead to increase of the imbalance between dimensionality and length of time series. Furthermore, these models lack explicit feature extraction procedure, i.e. lack of interpretablity.

In this paper, we propose a novel clustering method based on a mixture model that make use of time series of data effectively. State space models are used as component models of the mixture in order to handle high dimensional time series and to avoid the over-parameterization by considering dimension compression. The proposed method addresses the following tasks: (1) clustering samples according to temporal patterns of gene expressions, (2) automatic detection of genes that discriminate identified clusters, (3) estimation of a restricted autoregressive coefficient for each cluster. We will demonstrate the proposed method along with the cluster analysis of MS patients.

2. Methods

2.1. *Mixture of State Space Models*

Suppose that $y_n^{(l)} \in \mathbb{R}^p$ denotes a gene expression vector observed at the time point n corresponding to the lth sample among m samples, where the number of genes is p. Let us denote the set of observed time points and the time course gene expression profile of the lth sample by $\mathcal{N}_{\text{obs}}^{(l)} \subseteq \mathcal{N} = \{1, \ldots, N\}$ and $Y_N^{(l)} = \{y_n^{(l)} : n \in \mathcal{N}_{\text{obs}}^{(l)}\}$. Also, we denote an unobserved k dimensional hidden state vector by $x_n^{(l)} \in \mathbb{R}^k$. In order to avoid over-parameterization, we assume that k is much less than p. Our objective is to classify m samples into G clusters according to their time course gene expression profiles $Y_N^{(1)}, \ldots, Y_N^{(m)}$. Here, we assume that $Y_N^{(l)}$ is generated by one of state space models g:

$$y_n^{(l)} = H_g x_n^{(l)} + w_n^{(l)}, \quad n \in \mathcal{N}_{\text{obs}}^{(l)}, \tag{1}$$

$$x_n^{(l)} = F_g x_{n-1}^{(l)} + v_n^{(l)}, \quad n \in \mathcal{N}, \tag{2}$$

with probability α_g among G component models ($g = 1, \ldots, G$), starting with the initial state vector $x_0^{(l)} \sim N(\mu_{0g}, \Sigma_{0g})$. That is, $Y_N^{(l)}$ is generated by G-components mixture of the state space model. $F_g \in \mathbb{R}^k \times \mathbb{R}^k$ and $H_g \in \mathbb{R}^p \times \mathbb{R}^k$ are coefficient matrices which are often referred to as the system matrix and the observation matrix. w_n and v_n are the noise vectors which follow the normal distributions with mean zero and covariance matrices R_g and I, respectively. The lack of identifiability

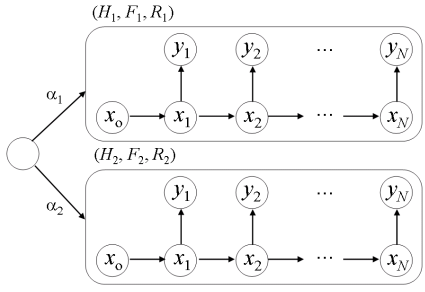

Fig. 1. A schematic expression of the two component mixture of the state space models ($G = 2$).

is a weakness of the state space model. In order to keep the uniqueness of the model, we impose parameter constraints as $H_g^T R_g^{-1} H_g = \Lambda_g$ for all g, where Λ_g is a $k \times k$ diagonal matrix. The detail of the constraints is described by our previous work [8].

2.2. *Clustering*

Let $c_g^{(l)} \in \{0, 1\}$ $(g = 1, \ldots, G)$ be an unknown class label of the lth sample which takes value one, i.e. $c_g^{(l)} = 1$, if the lth sample belongs to class g, otherwise $c_g^{(l)} = 0$. Suppose $\eta_g = (H_g, F_g, R_g, \mu_{0g})$ is the parameter set of the component state space model g and $\theta = (\alpha_1, \ldots, \alpha_G, \eta_1, \ldots, \eta_G)$ is the whole parameter set of the mixture state space model. Also, we suppose that $f_g(Y_N^{(l)}; \eta_g)$ is the gth component distribution given by

$$f_g(Y_N^{(l)}; \eta_g) = \prod_{n \in \mathcal{N}_{\text{obs}}^{(l)}} \phi(y_n^{(l)}; H_g x_{g,n|n-1}^{(l)}, H_g V_{g,n|n-1}^{(l)} H_g^T + R_g),$$

where $\phi(z; \mu, \Sigma)$ is the probability density function of a random vector z followed by the multivariate normal distribution with mean vector μ and covariance matrix Σ. $x_{g,n|n-1}^{(l)}$ and $V_{g,n|n-1}^{(l)}$ are defined by

$$x_{g,n|n-1}^{(l)} = E[x_n^{(l)} | Y_{n-1}^{(l)}, c_g^{(l)} = 1; \eta_g],$$
$$V_{g,n|n-1}^{(l)} = E[(x_n^{(l)} - x_{g,n|n-1}^{(l)})(x_n^{(l)} - x_{g,n|n-1}^{(l)})^T | Y_{n-1}^{(l)}, c_g^{(l)} = 1; \eta_g].$$

Samples can be clustered by applying the Bayes rule. That is, the sample l is assigned to gth group if the posterior probability of $c_g^{(l)} = 1$,

$$p(c_g^{(l)} = 1 | Y_N^{(l)}; \theta) = \alpha_g f_g(Y_N^{(l)}; \eta_g) / \sum_{g'=1}^{G} \alpha_{g'} f_{g'}(Y_N^{(l)}; \eta_{g'})$$

is maximum. More formally, the clustering rule is represented as follows:

$$\hat{c}_g^{(l)} = \begin{cases} 1 & \text{if } \ g = \text{argmax}_{g'} \, p(c_{g'}^{(l)} = 1 | Y_N^{(l)}; \theta), \\ 0 & \text{otherwise.} \end{cases}$$

Since the parameter set θ is unknown, θ will be replaced by the maximum likelihood estimator $\hat{\theta}$. The procedure of maximum likelihood estimation will be discussed in the next subsection.

2.3. *Parameter Estimation*

We use the EM algorithm for the maximum likelihood estimation. Complete data log-likelihood of the mixture state space model L_c is described as follows:

$$L_c(\theta) = \sum_{l=1}^{m}\sum_{g=1}^{G} c_g^{(l)}\big\{\ln \alpha_g + \ln p(Y_N^{(l)}, X_N^{(l)}|c_g^{(l)} = 1; \eta_g)\big\},$$

where $X_N^{(l)} = \{x_n^{(l)} : n \in \{0\} \cup \mathcal{N}\}$. Let $\rho_g^{(l)} = p(c_g^{(l)} = 1|Y_N^{(l)}; \theta)$. Maximizing $E[L_c(\theta)|Y_N^{(1)}, \ldots, Y_N^{(m)}; \bar{\theta}]$ given an arbitrary initial parameter $\bar{\theta}$, we obtain following recursive formula for calculating the maximum likelihood estimator:

$$\begin{aligned}
\alpha_g^{\text{new}} &= \sum_{l=1}^{m} \rho_g^{(l)}/m,\\
\mu_{0g}^{\text{new}} &= \sum_{l=1}^{m} \rho_g^{(l)} E[x_0^{(l)}|Y_N^{(l)}, c_g^{(l)} = 1; \bar{\eta}_g]/m,\\
H_g^{\text{new}} &= T_{yx}(g)T_{xx}(g)^{-1},\\
F_g^{\text{new}} &= T_{xx'}(g)T_{x'x'}(g)^{-1},\\
R_g^{\text{new}} &= \Big\{\sum_{l=1}^{m} \rho_g^{(l)}|\mathcal{N}_{\text{obs}}^{(l)}|\Big\}^{-1}\big\{T_{yy}(g) - T_{yx}(g)T_{xx}(g)^{-1}T_{yx}(g)^T\big\},
\end{aligned}$$

where

$$\begin{aligned}
T_{yy}(g) &= \sum_{l=1}^{m} \rho_g^{(l)} \sum_{n\in\mathcal{N}_{\text{obs}}^{(l)}} E[y_n^{(l)} y_n^{(l)T}|Y_N^{(l)}, c_g = 1; \bar{\eta_g}],\\
T_{yx}(g) &= \sum_{l=1}^{m} \rho_g^{(l)} \sum_{n\in\mathcal{N}_{\text{obs}}^{(l)}} E[y_n^{(l)} x_n^{(l)T}|Y_N^{(l)}, c_g = 1; \bar{\eta_g}],\\
T_{xx}(g) &= \sum_{l=1}^{m} \rho_g^{(l)} \sum_{n\in\mathcal{N}_{\text{obs}}^{(l)}} E[x_n^{(l)} y_n^{(l)T}|Y_N^{(l)}, c_g = 1; \bar{\eta_g}],\\
T_{xx'}(g) &= \sum_{l=1}^{m} \rho_g^{(l)} \sum_{n=1}^{N} E[x_n^{(l)} x_{n-1}^{(l)T}|Y_N^{(l)}, c_g = 1; \bar{\eta_g}],\\
T_{x'x'}(g) &= \sum_{l=1}^{m} \rho_g^{(l)} \sum_{n=1}^{N} E[x_{n-1}^{(l)} x_{n-1}^{(l)T}|Y_N^{(l)}, c_g = 1; \bar{\eta_g}].
\end{aligned}$$

Expectation terms are calculated efficiently with Kalman filtering and smoothing procedure. For example, see [6, 7]. We obtain (local) maximum likelihood estimate $\hat{\theta}$ repeating the parameter updating process until a suitable convergence criterion is satisfied. Some implementation issues are described in Appendix.

2.4. *Feature Selection and Estimation of Restricted Autoregressive Coefficients*

In cluster analysis of gene expression profiles, it is of interest to detect genes that characterize identified clusters. By transforming observation equation Eq. 1 under the constraints $H_g^T R_g^{-1} H_g = \Lambda_g$, $g = 1, \ldots, G$, gene expression vectors y_n can be mapped onto the state space $\mathbb{R}^k$ with the feature extraction matrix $D_g \in \mathbb{R}^k \times \mathbb{R}^k$ as follows:

$$x_n^{(l)} = D_g R_g^{-1/2} (y_n^{(l)} - w_n^{(l)}), \tag{3}$$

where $D_g = \Lambda_g H_g^T R_g^{-1/2}$ [8]. If the dimension of state k is taken to be lower than p, the dimensionality of the noise-removed gene expression vectors $R_g^{-1/2}(y_n - w_n)$ is reduced by the semi-orthogonal matrix D_g in which (i, j)th element of D_g represents the contribution of the jth gene to the ith coordinate of the state space. In practice, it is useful to extract a number of significant genes for the each state variable.
Furthermore, substituting Eq. 3 into the state space model Eq. 1 and Eq. 2, the following first-order vector autoregressive representation are obtained:

$$R_g^{-1/2}(y_n^{(l)} - w_n^{(l)}) = \Psi_g R_g^{-1/2}(y_{n-1}^{(l)} - w_{n-1}^{(l)}) + R^{-1/2} H v_n^{(l)} \tag{4}$$

where $\Psi_g = D_g^T \Lambda_g F_g D_g$ [9]. The (i, j)th element of the autoregressive coefficient matrix Ψ_g represents the influence of the jth gene on the ith gene for the component model g. After estimating parameters, we obtain the estimated Ψ_g which captures the temporal structure of the gene expressions for the identified gth cluster. Comparisons of these estimated coefficient matrices $\Psi_1, \ldots, \Psi_G$ gives us an insight to understand temporal features of G groups.
Note that the degree of freedom in the autoregressive coefficient matrix Ψ_g is of order $O(p) = p(k+1) + k^2 - k(k-1)/2$. From this point of view, the state space model is considered as a parsimonious parameterization of the vector autoregressive model and provides an approach to control the model complexity by choosing dimension of state vector k.

3. Experiments

3.1. *Gene Expression Profiles of MS Patients*

The disease MS is characterized by myelin destruction and oligodendrocyte death and causes relapsing-remitting neurological disorders such as visual disorder and movement disorder. rIFN-β is routinely used for suppressing the relapse of the symptom. However, almost half of MS patients are not benefited by the therapy. Baranzini *et al.* [2] investigated long-term effects of rIFN-β on disease progression with the time course gene expression profiles of 76 genes which are mainly related to the immune system. Expression levels are measured by conducting reverse-transcription PCR at the beginning of the therapy and after 3, 6, 9, 12, 18, and 24 months. This dataset includes profiles of 53 MS patients categorized into 33 "good" responders

and 20 "poor" responders according to their response levels for rIFN-β administration. A group of patients were categorized into poor responders if they suffered two or more relapses or experienced an increase of one point in the expanded disability status scale score (EDSS), a measure of progression for the MS disease, until two years after the initiation of the therapy. Good responders were defined as patients that experienced a total suppression of relapses and no increase in the EDSS.

3.2. *Results*

We applied the proposed method to the expression profiles after converting the real time set {0, 3, 6, 9, 12, 18, 24} months into $\mathcal{N}_{\text{obs}} = \{1, 2, 3, 4, 5, 7, 9\}$ where the entire time points are defined by $\mathcal{N} = \{1, 2, \ldots, 7, 8, 9\}$ and $\mathcal{N}_{\text{obs}}$ is the union of $\mathcal{N}_{\text{obs}}^{(l)}$, $l = 1, \ldots, m$. We preset the number of clusters $G = 2$ and the dimension of states $k = 2$. Among 600 estimated parameters of the EM algorithm, which were computed by the different initial parameters, we chose the best parameter corresponding to the highest local maxima of likelihoods. Missing values included in the dataset were imputed by the EM algorithm, that is, we considered missing values as latent variables as well as the state vectors and we calculated their conditional expectations at the E-step.

The left panel of Fig. 2 shows the identified clusters of 53 MS patients. The identified two groups composed of 28 (cluster A) and 25 (cluster B) patients, respectively. The cluster A included 27 good responders and one poor responder while the cluster B included 19 poor responders and 6 good responders. The resulting two clusters A and B are likely to reflect the diagnostic categories. If we assume the cluster A corresponds to good responders and the cluster B poor responders, the total prediction accuracy is 86.8 $(= 100 \times (27 + 19)/53)\%$.

We investigated the performance of the proposed method by comparing with a result of a hierarchical clustering. The result of a hierarchical clustering is illustrated in the right panel of Fig. 2. We used the complete-linkage-based hierarchical clustering by vectorizing the time course gene expression matrix for each patient, where the distance between two patients was measured based on the Pearson correlation. Patients in the good and poor responder groups are drawn by grey and white squares, respectively. Since all the patients except two belong one of clusters in case we divide patients into two clusters, the threshold value of the correlation was set to 0.80 which seems to be the best value for the classification of patients according to responder groups. As a result, the patients were divided into eight clusters except for three patients. Even if we assign the clusters to two responder groups by the majority decision, 13 patients are belonged in incorrect clusters, and thus, the proposed method seems to outperform the hierarchical clustering in performing the cluster analysis of MS patients from the dataset.

Next, we focus on the significant genes which separate the identified two clusters. Significant genes are extracted by feature extraction matrices D_1 and D_2 for the two clusters. In order to discover genes that characterize a difference between two

Fig. 2. The clustering result of the proposed method (left) and hierarchical clustering (right).

clusters, we computed $S = ||D_1| - |D_2||$ and selected some genes which achieved the highest score of the S. For the first coordinate of the state space, genes with the largest five score of the first row of S were TRAIL, p53, CD5, CD22 and JAK2, while CD22, CD5, TRAIL, JAK2 and p53 for the second coordinate. It was reported that TRAIL was a potential response marker for the rIFN-β treatment in MS [10]. Furthermore, Wosik *et al.* [11] reported that oligodendrocytes were protected from p53-induced cell death by blocking signaling through TRAIL receptors.
Fig. 3 display heatmaps of the estimated autoregressive coefficient matrices Ψ_1 (left) and Ψ_2 (right). Positive and negative influences are color-coded by blue and green, respectively. The estimated coefficient matrices capture clear differences in the longitudinal effect between the two clusters. For example, TRAIL has strongly negative longitudinal effects on many genes for the first cluster while does not for the second cluster. Also, p53 has large positive longitudinal effects on many genes for the first cluster, while does not for the second cluster.

4. Concluding remarks

We proposed a novel method that perform cluster analysis of samples, where each sample is characterized by a time course gene expression profile using the mixture of state space models. The proposed method were applied to the time course gene expression profiles of 53 MS patients under the therapy of rIFN-β. We succeeded in the classification of the MS patients according to the response level for rIFN-β. The proposed method exploits the mixture of the state space models which enables us to understand differences in the temporal structure of gene expressions between the identified clusters. Here, we point out a limitation in the model selection. In the context of the mixture state space models, one of the important tasks is to determinate the number of clusters and the dimension of states. For example, when we used the mixture state space model under $k = 2$ and $G = 3$, in the cluster analysis of MS patients, many of the estimated models got degenerated, that is, at least one of mixing proportions $\alpha_1, \ldots, \alpha_G$ became zero. Probably, such an unstable estimation occurred due to the imbalance between the model complexity and amount of data, i.e. over-parameterization. This indicates that applicability of the proposed method is limited in terms of the number of clusters and the dimension of state. In order to

overcome the problem of over-parameterization, one possible solution is to perform Bayesian estimations, which provide an approach to control the model complexity by constructing prior distributions of parameters appropriately. It is challenging to design the plausible prior distributions and we are now investigating this problem.

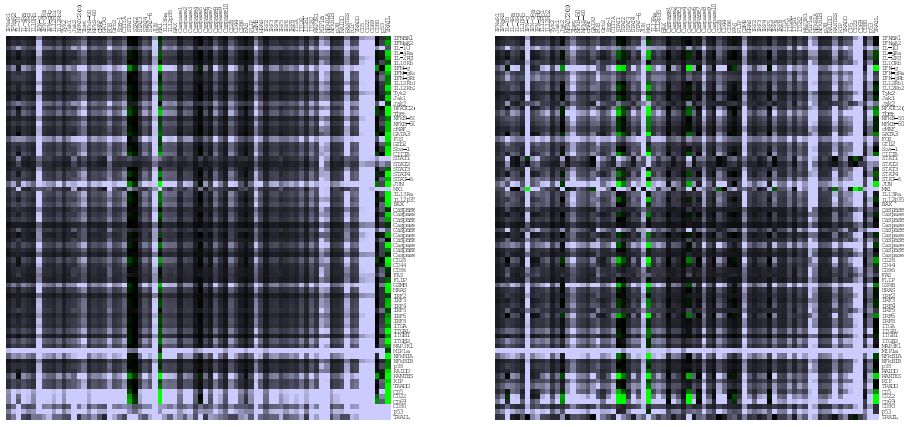

Fig. 3. Estimated autoregressive coefficient matrices Ψ_1 (left) and Ψ_2 (right).

References

[1] Akutsu, T., Kuhara, S., Maruyama, O., and Miyano, S., Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions, *Proc. 9th ACM-SIAM Symp. Discrete Algorithms*, 695–702, 1998.

[2] Baranzini, S.E., Mousavi, P., Rio, J., Caillier, S.J., Stillman, A., Villoslada, P., Wyatt, M.M., Comabella, M., Greller, L.D., Somogyi, R., Montalban, X., and Ksenberg J.R., Transcription-based prediction of response to IFNb using supervised computational methods. *PLoS Biology*, 3(1):166–176, 2005.

[3] Hsiung, K.L., Kim, S.J., and Boyd, S., *Tractable approximate robust geometric programming.* Tech. Rep., Department of Electrical Engineering, Stanford University, California, 2006 2006.

[4] Imoto, S., Tamada, Y., Araki, H., Yasuda, K., Print, C.G., Charnock-Jones, S.D., Sanders, D., Savoie, C.J., Tashiro, K., Kuhara S., and Miyano, S., Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles. *Pacific Symposium on Biocomputing*, 11:559–571, 2006.

[5] Kim, S., Imoto, S., and Miyano, S., Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, 75:(1-3), 57–65, 2004.

[6] Kitagawa, G. and Gersch, W., *Smoothness priors analysis of time series*. New York, Springer-Verlag, 1996.

[7] Shumway, R.H. and Stoffer, D.S., Dynamic linear models with switching. *J. American Statistical Association*, 86:763–769, 1991.

[8] Yamaguchi, R., Yoshida, R., Imoto, S., Higuchi, T., and Miyano, S., Finding module-based gene networks with state space models -Mining high-dimensional and short time-course gene expression data, *IEEE Signal Processing Magazine*, 24(1):37–46, 2007.

[9] Yoshida, R., Imoto, S., and Higuchi, T., Estimating time-dependent gene networks from time series microarray data by dynamic linear models with Markov switching. *Proc. 4th Computational Systems Bioinformatics (CSB2005)*, 289–298, 2005.
[10] Wandinger, K.P., Lunemann, J.D., Wengert, O., Bellmann-Strobl, J., Aktas, O., Weber, A., Grundstrom, E., Ehrlich, S., Wernecke, K.D., Volk, H.D., Zipp, F., TNF-related apoptosis inducing ligand (TRAIL) as a potential response marker for interferon-beta treatment in multiple sclerosis. *Lancet.* 361(9374):2036–2043, 2003.
[11] Wosik, K., Antel, J., Kuhlmann, T., Bruck, W., Massie, B., and Nalbantoglu, J., Oligodendrocyte injury in multiple sclerosis: a role for p53. *Journal of Neurochemistry*, 85:635–644, 2003.

Appendix A. Implementation Issues

We explain issues on implementations. In order to check the convergence of the EM algorithm, we need to calculate the log-likelihood of proposed model:

$$\sum_l \ln \sum_g \alpha_g f_g(Y_N^{(l)}; \bar{\eta}_g) = \sum_l \ln \sum_g \alpha_g \exp\bigl\{\ln f_g(Y_N^{(l)}; \bar{\eta}_g)\bigr\} \tag{A.1}$$

for each step. However, the component log-likelihood $\ln f_g(Y_N^{(l)}; \bar{\eta}_g)$ calculated by Kalman filtering sometimes becomes an extremely large value because of the high dimensinality of time course gene expression profiles. In such situation, $f_g(Y_N^{(l)}; \bar{\eta}_g)$ easily gets infinity from hardware limitations of floating points. Calculating formula such as (A.1) is known for log-sum-exp problem. A few methods have been proposed for approximating log-sum-exp including the method that approximates it through geometrical optimizations [3]. However, log-sum-exp formula is calculated as follows:

$$\sum_l \left\{ \ln f_{g_M^{(l)}}(Y_N^{(l)}; \bar{\eta}_g) + \ln \sum_g \alpha_g \exp\left(\ln f_g(Y_N^{(l)}; \bar{\eta}_g) - \ln f_{g_M^{(l)}}(Y_N^{(l)}; \bar{\eta}_g) \right) \right\},$$

where $g_M^{(l)} = \operatorname{argmax}_g f_g(Y_N^{(l)})$ for $l = 1, \ldots, m$, since the argument of exponential is assured to be at most zero. Note that the similar calculation is needed for the posterior probabilities $p(c_g^{(l)} = 1 | Y_N^{(l)}; \bar{\theta})$.