2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Modeling Gene Expression Dynamics Based on a Linear Dynamical System Model

Naoto Yukinawa[†], Jun-ichiro Yoshimoto[‡†], Shigeyuki Oba[†] and Shin Ishii[†]

†Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara, Japan
‡Initial Research Project, Okinawa Institute of Science and Technology
12-75 Suzaki, Gushikawa, Okinawa, 904-2234, Japan
Email: naoto-yu@is.naist.jp, jun-y@irp.oist.jp, shige-o@is.naist.jp, ishii@is.naist.jp

Abstract—In this study, we propose a linear dynamical system model for analyzing gene expression time-series data. We also provide a variational Bayes inference for the model. The inference algorithm is implemented as maximization of the variational free energy, which can be used as a Bayesian criterion for the model selection. We apply our method to a published yeast cell-cycle gene expression data set. The result shows that our approach can obtain robust detection of time course. The dimensionality estimated based on the free energy seems to be plausible.

1. Introduction

In order to understand biological activities of cells at molecular level, it is required to know various aspects of gene expression; the concentration, timing, conditions, and localization in subcellular organelle, in detail. Gene expression in living organisms is controlled within a complicated gene regulatory system consisting of many intermolecular interactions of various components such as nucleic acids, proteins or other small molecules. Even if we have individual concentration information of each biological component in cells, it is still difficult to understand its dynamics directly. At the present day, some analysis methods have been developed for this problem, which are based on statistics and/or information technology for gene expression profiles – comprehensive measurement of expression levels of thousands of genes in cells using mRNA quantification technology [10].

In this study, we propose a linear dynamical system (LDS) approach to modeling gene regulatory systems aiming to analyze dynamic behaviors of regulatory factors. We also expect that it helps extracting character of each gene from time-series data of gene expression. The LDS model belongs to the Gaussian process model family, and assumes observation time-series are generated from unobservable time-series of the internal state variable plus Gaussian noise. In order to relate this model to the actual gene regulatory system, we assume that observation time-series represent alteration of gene expression levels, and that the internal state variables represent higher level systems in cells or external environments, i.e., regulatory factors for the genes. These factors, for instance, are assumed to be some transcriptional regulatory proteins or some environmental conditions such as pH or temperature and so on.

The parameter estimation of the LDS model, which is one of exponential family models, can be done as maximum likelihood (ML) estimation by using expectation maximization algorithm (EM algorithm) [6, 7].

According to an ML-based method, however, it is difficult to identify the dimensionality of internal state of the model, because the ML estimation is likely to prefer complicated models. For this problem, variational Bayes (VB) algorithm [5] is expected to provide efficient estimation. We formerly devised a VB algorithm for an LDS model [4, 2] in order to estimate parameters and to determine the internal state dimensionality which represents complexity of the internal system. In this study, we introduce this method to a system identification problem of a gene regulatory network. We also examine the estimated observation matrix each of whose row vector is regarded as a feature representation of a gene, and compare it with a known biological feature of gene.

Our LDS-based approach is applied to a published gene expression data of yeast *Saccharomyces cerevisiae* measured along cell cycles [8]. As a result, our approach provides good model selection criterion and extracts plausible features of individual genes.

2. Linear Dynamical System Model

2.1. Gene Expression Profiles

A gene expression profile represents measured expression levels of many genes in a biological sample at a specific experimental condition. Log expression ratio between a target sample and a control sample is typically used for a measurement of the expression level.

$$\boldsymbol{y}_t = (y_{t1}, \ldots, y_{tD})'; \quad t = 1, \ldots, T \tag{1}$$

denotes a gene expression vector at measuring point of time t, where y_{tj} is the j-th gene's expression level at t, D the number of genes, and T the number of measuring time-points. Namely, eq. (1) denotes expression time-series consisting of T time points for D genes.

2.2. Probabilistic Model for Linear Dynamical System model

An LDS model is formulated as follows. Time-series of an N-dimensional unobservable internal state variable $\boldsymbol{x}_t$ over discrete time points, $t = 1, \ldots, T$, and those of a D-dimensional observable variable $\boldsymbol{y}_t$ which is derived by linear transformation of $\boldsymbol{x}_t$, are defined by

$$\begin{aligned}
\boldsymbol{x}_t &= \boldsymbol{W}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t; \quad t = 2, \ldots, T, && (2)\\
\boldsymbol{y}_t &= \boldsymbol{V}\boldsymbol{x}_t + \boldsymbol{\eta}_t; \quad t = 1, \ldots, T, && (3)\\
\boldsymbol{x}_1 &\sim \mathcal{N}_N(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \sigma_1^2\boldsymbol{I}_N), && (4)\\
\boldsymbol{\epsilon}_t &\sim \mathcal{N}_N(\boldsymbol{\epsilon}_t|\boldsymbol{0}_N, \sigma_\epsilon^2\boldsymbol{I}_N), && (5)\\
\boldsymbol{\eta}_t &\sim \mathcal{N}_D(\boldsymbol{\eta}_t|\boldsymbol{0}_D, \sigma_\eta^2\boldsymbol{I}_D), && (6)
\end{aligned}$$

where $\boldsymbol{x}_1$ is the initial value of $\boldsymbol{x}$. $\boldsymbol{\epsilon}_t \in \mathcal{R}^N$ and $\boldsymbol{\eta}_t \in \mathcal{R}^D$ are system noise and observation noise, respectively, which obey white Gaussian distributions; $\mathcal{N}_p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{S}) \equiv (2\pi)^{-p/2}|\boldsymbol{S}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$ denotes the probability density function of a p-dimensional multivariate Gaussian with a mean $\boldsymbol{\mu}$, and a covariance matrix $\boldsymbol{S}$. $\boldsymbol{\mu}_1 \in \mathcal{R}^N$ is the prior mean of the initial state variable, $\boldsymbol{W} \in \mathcal{R}^{N\times N}$ evolution matrix, and $\boldsymbol{V} \in \mathcal{R}^{D\times N}$ observation matrix. $\sigma_1^2, \sigma_\epsilon^2$ and σ_η^2 are variances of $\boldsymbol{x}_1$, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$, respectively. $\theta \equiv \{\boldsymbol{\mu}_1, \sigma_1^2, \boldsymbol{W}, \sigma_\epsilon^2, \boldsymbol{V}, \sigma_\eta^2\}$ is the set of model parameters. According to the above formulation, the log likelihood of parameter $\boldsymbol{\theta}$ given the complete data $(\boldsymbol{X}_{1:T}, \boldsymbol{Y}_{1:T})$, where $\boldsymbol{X}_{1:T} \equiv \{\boldsymbol{x}_t\}$ and $\boldsymbol{Y}_{1:T} \equiv \{\boldsymbol{y}_t\}, t = 1, \ldots, N$, is obtained as

$$\begin{aligned}
&p(\boldsymbol{Y}_{1:T}, \boldsymbol{X}_{1:T}|\boldsymbol{\theta}) = \\
&p(\boldsymbol{x}_1|\boldsymbol{\theta})\left[\prod_{t=2}^{T} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta})\right]\left[\prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta})\right]. \quad (7)
\end{aligned}$$

2.3. Property of Observation Matrix

In the $D \times N$ observation matrix $\boldsymbol{V}$, each row vector $\boldsymbol{v}_i \in \mathcal{R}^{1\times N}, i = 1, \ldots, D$, represents the transformation from the internal state variable $\boldsymbol{x}_t$ to the i-th observable variable y_{ti}. This implies that $\boldsymbol{v}_i$ can be considered as an observed feature of the i-th gene.

2.4. Variational Estimation Method

We performed variational estimation assuming a conjugate prior distribution for the model parameter $\boldsymbol{\theta}$. In the variational estimation, we variationally maximized the lower bound $\mathcal{F}[q(\boldsymbol{\theta}), q(\boldsymbol{X})]$ of the log evidence $\ln p(\boldsymbol{Y})$, which was derived by approximating the posterior distribution as a factorized one with respect to the state variable $\boldsymbol{X}$ and the parameter $\boldsymbol{\theta}$; $q(\boldsymbol{X})q(\boldsymbol{\theta}) \approx p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y})$. This lower bound is referred to as variational free energy and defined as

$$\begin{aligned}
\ln p(\boldsymbol{Y}) &= \ln \int p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta})p(\theta)d\boldsymbol{\theta}d\boldsymbol{X} \\
&\geq \int q(\boldsymbol{\theta})q(\boldsymbol{X}) \ln \frac{p(\boldsymbol{Y}, \boldsymbol{X}|\theta)p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})q(\boldsymbol{X})} d\boldsymbol{\theta}d\boldsymbol{X} \\
&\equiv \mathcal{F}\left[q(\boldsymbol{\theta})q(\boldsymbol{X})\right]. \quad (8)
\end{aligned}$$

The variational maximization was done by a VB-EM algorithm which iteratively maximizes $\mathcal{F}[q(\boldsymbol{\theta})q(\boldsymbol{X})]$ with respect to $q(\boldsymbol{X})$ and $q(\boldsymbol{\theta})$ alternately. It is guaranteed that the algorithm converges to a local optimum. Because the maximized value of the free energy approximates the log evidence, it can be utilized as a Bayesian model selection criterion; the most appropriate model can be selected from models with various number of parameters [3].

We defined priors of parameter $\boldsymbol{\theta}$ as

$$\begin{aligned}
p(\boldsymbol{\mu}) &= \mathcal{N}_{\boldsymbol{\mu}}\left(\boldsymbol{\mu}|\boldsymbol{0}, \gamma_0^{-1}\boldsymbol{I}_N\right), && (9)\\
p(\sigma_1^2) &= \mathcal{G}\left(\sigma_1^{-2}|\gamma_0, \gamma_0\tau_{\mu 0}\right), && (10)\\
p(\mathcal{W}) &= \prod_{i=1}^{N} \mathcal{N}_N\left(\boldsymbol{w}_i|\boldsymbol{0}_N, \gamma_0^{-1}\boldsymbol{I}_N\right), && (11)\\
p(\sigma_\epsilon^2) &= \mathcal{G}\left(\sigma_\epsilon^{-2}|\gamma_\epsilon, \gamma_\epsilon\tau_{\mu_\epsilon}\right), && (12)\\
p(\mathcal{V}) &= \prod_{j=1}^{D} \mathcal{N}_D\left(\boldsymbol{v}_j|\boldsymbol{0}_N, \gamma_0^{-1}\boldsymbol{I}_D\right), && (13)\\
p(\sigma_\tau^2) &= \mathcal{G}\left(\sigma_\tau^{-2}|\gamma_\tau, \gamma_\tau\tau_{\mu_\tau}\right), && (14)
\end{aligned}$$

where $\mathcal{G}(\sigma^{-2}|\gamma, \gamma\tau)$ denotes Gamma distribution $\mathcal{G}(\sigma^{-2}|\gamma, \gamma\tau) \equiv \frac{(\gamma\tau)^\gamma(\sigma^{-2})^{\gamma-1}}{\Gamma(\gamma)} \exp[-\gamma\tau\sigma^{-2}]$. $\boldsymbol{w}_i$ and $\boldsymbol{v}_j$ are the i-th row vector of $\mathcal{W}$ and the j-th column vector of $\mathcal{V}$, respectively. To represent vague priors, we set prior hyper parameters as $\gamma_0 = 0.0001$, $\gamma_{\epsilon 0} = \gamma_{\eta 0} = 0.01$, $\tau_{\epsilon 0} = \tau_{\epsilon 0} = 0.01$, and $\tau_{\mu 0}$ was set at the mean variance over genes.

3. Applications

3.1. Yeast cell-cycle data

Our LDS model was applied to a published data set, gene expression profiles of yeast *Saccharomyces cerevisiae* strain *cdc15-2* obtained through a time course on mitotic cell cycle, which were processes by Spellman *et al.* [8]. This data set contains log expression ratios of 6177 genes at 24 time points, measured by cDNA microarrays, and is available at http://cellcycle-www.stanford.edu/.

As data preprocess, we constructed 800 × 19 expression matrix from 800 cell cycle related genes identified by Spellman *et al.* and 19 samples measured at every 10 minutes. There were 1023 (5.3%) missing values appeared in the matrix, and they were imputed by missing value estimation algorithm based on Bayesian PCA [9]. After that, we prepared a smaller matrix consisting of 200 genes which were chosen randomly from the 800 genes, for avoiding heavy computation. (figure **??**).

3.2. Results

3.3. System identification

We prepared ten LDS model structures in which the internal state dimensionality was set at $N = 1, \ldots, 10$, and then the VB-EM algorithm was employed for parameter estimation of each model structure. Because the algorithm may converge to a poor local optimum of the free energy, the best run was selected among 20 runs starting from random initial setting for each model structure. The free energy can be the criterion for selecting the best run. Figure 1 shows the maximum free energy against the dimensionality of internal state variables, $N = 1, \ldots, 5$. According to the free energy maximization, we can say that the optimal dimensionality of the internal state variables is $N = 2$. The free energy values of the model whose internal dimensionality was larger than 5 were further smaller (data not shown).

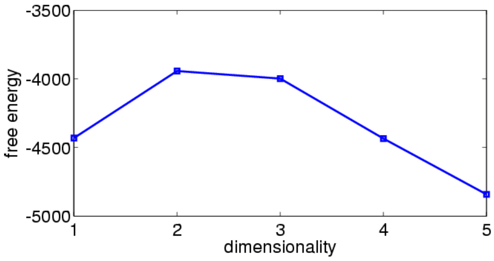

Figure 1: Free energy versus the dimensionality of the internal state variables.

Figure 2 shows the time-series of the internal state variables in each of $N = 1, \ldots, 5$ models, each of which has the maximum free energy among the 20 runs. Each column in the figure corresponds to the dimensionality of the internal state variables. Figure 3 shows the estimated standard deviation of system noise and observation noise in each $N = 1, \ldots, 7$ model.

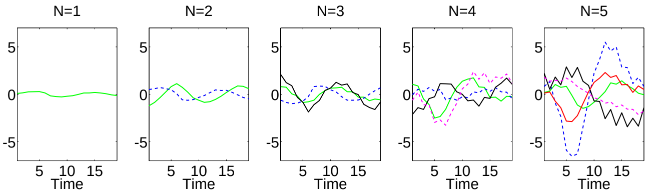

Figure 2: Time-series of the internal state variable $\boldsymbol{x}$.

The model with $N = 1$ is not enough to describe the dynamic nature of the gene expression profile. In the model with $N = 4$ or $N = 5$, we can observe some redundant internal state variables which are approximated by linear combination of other state variables. In the model ($N = 2$) which exhibited the largest free energy among all models, the internal state variables represent two oscillatory behaviors with different phases with each other, corresponding to first-order Fourier base. In this model, the estimated standard deviation of system noise also shows minimum and contributes to the smooth changes of internal state variables.

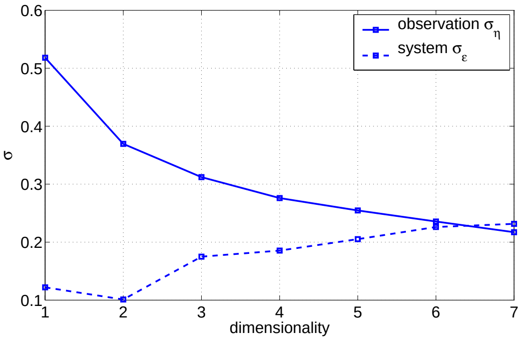

Figure 3: Standard deviation of noise on LDS estimated models for the yeast data set. ($N = 1, \ldots, 7$). Solid and dash lines represents standard deviation of observation noise and system noise, respectively.

Figure 4 shows the scatter plot of two-dimensional row vectors of the observation matrix $\mathcal{V}$ in the case $N = 2$. Each point in the figure corresponds to a single gene and each symbol represents one of the five kinds of phases in the cell-cycle; S, S/G_2, G_2/M, M/G_1or G_1/S, identified by Spellman *et al.*'s analysis. This result suggests that our LDS model with the VB inference automatically extracted characteristics which are closely related to the phase information obtained and classified by Spellman *et al.*

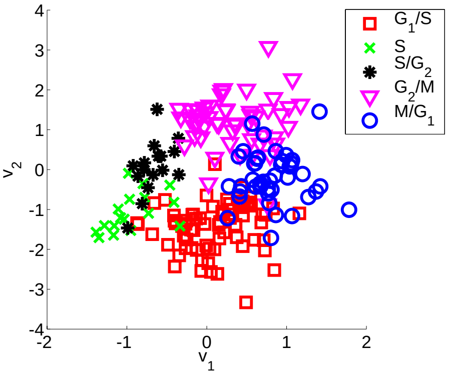

Figure 4: Scatter plot of row vectors of V in the LDS model with $N = 2$. Each symbol represents the known phase information.

4. Discussion

The mitotic cell division is a fundamental physiological phenomenon commonly observed in various cells, which has two interesting properties: periodicity and autonomy. The time period from a mitosis to the next mitosis is called the cell cycle. The cell cycle can be divided into four stages based on macroscopic behaviors. In the molecular level, each gene is activated at a specific phase in the cell cycle and the alternation of expression levels of genes is dynamical, under the influence of environmental factors. Spellman *et al*. [8] assumed the periodicity of gene expression level in the cell cycle, thus employed Fourier bases; sine and cosine waves for modeling gene expression dynamics. The parameters describing the system in their model were the phase and the frequency, and these parameters correspond to the initial internal state variable $\boldsymbol{x}_1$ and the evolution matrix $\mathcal{W}$, respectively, in our LDS model. In addition, two weight parameters correspond to the observation matrix $\mathcal{V}$ in our LDS model. Our Bayesian model selection criterion automatically led to select a model with $N = 2$ whose bases are thus related closely to those used in the model of Spellman *et al*.

In earlier studies of system identification based on the linear state space model from time-series of gene expression profiles, singular value decomposition (SVD) [11] and factor analysis (FA) [1] were used. These studies estimated internal state variables and parameters with simpler assumptions on linear state space models; the noise can be ignored in the system and observation processes. Wu *et al*. [1] employed Bayesian information criterion (BIC) for dimensionality selection of the estimated FA models. Their method was applied to the yeast data set we used, but their optimal dimensionality of the internal state variables was $N = 5$; their optimal model is more complex than ours, $N = 2$. We consider that their FA model may overfit to the data noise. Since our LDS model is formulated as a probabilistic model which involves the system noise and the observation noise, it is statistically robust for the noise contained in the data. This property is great advantage for analyzing noisy data in real world such as gene expression profiles measured by microarrays.

5. Conclusion

We introduced variational estimation for a linear dynamical system model. Our method automatically obtained the optimal bases from noisy time-series. Thus our method can be a tool to explore underlying factors of self-sustaining and cyclical systems like biological organisms.

On the other hand, remaining problems of our method are as follows: we assumed the linearity of living organism system whereas the causal relationships among biological components show non-linearity in general. In addition, the method requires the system to be stationary. We will deal with these problems in a near future work.

References

[1] F. X. Wu, W. J. Zhang and A. J. Kusalik, “Modeling gene expression from microarray expression data with state-space equations” *Pacific Symposium on Biocomputing*, vol.9, pp.581-592, 2004.

[2] J. Yoshimoto, S. Ishii and M. Sato, “System identification based on on-line variational Bayes method and its application to reinforcement learning”, *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003*, Lecture Notes in Computer Science 2714, pp.123-131, 2003.

[3] J. Yoshimoto, S. Ishii and M. Sato, “Hierarchical model selection for NGnet based on variational Bayes inference”, *Artificial Neural Networks - ICANN 2002*, Lecture Notes in Computer Science 2415, pp.661-666, 2002.

[4] Z. Gharahmani and M. J. Beal, “Propagation algorithms for variational Bayesian learning”, *Advances in Neural Information Processing Systems 13*, pp.507-513, 2001.

[5] H. Attias, “Inferring parameters and structure of latent variable models by variational Bayes”, *Proc. of 15th Conference on Uncertainty in Artificial Inteligence*, pp.21-30, 1999.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of Royal Statistical Society B*, vol.39, pp.1-38, 1977.

[7] S. Roweis and Z. Gharahmani, “A unifying review of Linear Gaussian models”, *Neural Computation* vol.11, pp.305-345, 1999.

[8] P. T. Spellman, G. Sherlock, M. Q. Zang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher, , “Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization” , *Molecular Biology of the Cell*, vol.9, pp.3273-3297, 1998.

[9] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, “A Bayesian missing value estimation method”, *Bioinformatics*, vol.19, pp.2088-2096, 2003.

[10] H. de Jong, “Modeling and simulation of genetic regulatory system: a literature review”, *Journal of Computational Biology*, vol.9, pp.67-103, 2002.

[11] T. G. Dewey and D. J. Galas, “Dynamic models of gene expression and classification” ,*Functional & Integrative Genomics*, vol.1, pp.269-278, 2001.