Identification of Apoptosis-Induced Gene Networks from Time Course Expression Profiles with Replicated Measurements

Osamu Hirose ochamu@ims.u-tokyo.ac.jp
Ryo Yoshida yoshidar@ims.u-tokyo.ac.jp
Seiya Imoto imoto@ims.u-tokyo.ac.jp
Satoru Miyano miyano@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Keywords: microarray data, replicated measurements, gene networks, state space model

1 Introduction

To analyze multivariate time course data, a wide variety of models have been proposed in statistics, e.g. the vector autoregressive models. To our knowledge, their applications, however, have been limited owing to the fact that time course of gene expression profiles is fairly short, typically less than 4. To analyze multivariate time course data, a wide variety of models have been proposed in statistics, e.g. the vector autoregressive models. To our knowledge, their applications, however, have been limited owing to the fact that time course of gene expression profiles is fairly short, typically less than 10, whereas the number of targeted genes is ranging from $10^2$ to $10^4$. The length of time course gene expression profiles is not enough to construct such large gene networks. For example, it is obvious that the parameter estimations of the vector autoregressive model from such a short time course often fail due to occurrence of overfitting.

The state space model has a potential to construct large gene networks from time course gene expression profiles [2]. However, for analyzing extremely short time course, e.g. less than 10, the parameter estimations of the state space model still fail due to occurrence of overfitting. To extend applicability of the state space model, we provide a way of using replicated time courses of gene expression profiles which are usually measured in duplicate or triplicate for assessing reproducibility of data. Incorporating the replicated measurements into analysis considerably increases the predictive power of the state space model, and enables to infer more accurate gene networks.

2 Methods

Suppose $y_n^{(l)} \in \mathbb{R}^p$ and $x_n^{(l)} \in \mathbb{R}^k$ are the $l$th replicated expression profile and the $l$th hidden (unobserved) state vector at time $n$ for $l = 1, \ldots, m, n = 1, \ldots, N$. In order to analyze replicated measurements, we extend the conventional state space model as follows:

$$y_n^{(l)} = H x_n^{(l)} + w_n^{(l)}, \quad x_n^{(l)} = F x_{n-1}^{(l)} + v_n^{(l)},$$

where $H$ is a $p \times k$ matrix and $F$ is a $k \times k$ matrix. $w_n^{(l)}$ and $v_n^{(l)}$ are noise terms subjected to $w_n^{(l)} \sim N(0, R)$ and $v_n^{(l)} \sim N(0, I)$. We assume that the covariance $R$ is diagonal for avoiding severe overfitting. To keep the identifiability of parameters, we impose the parameter constraint as $H^T R^{-1} H = \Lambda$, where $\Lambda$ is a $k \times k$ diagonal matrix.

The parameter $(H, F, R)$ is estimated with the EM algorithm. Let $\theta_i = (H_i, F_i, R_i)$ be the estimate of the parameter set after the $i$th iteration of the EM algorithm. Given $\theta_i$ and $Y_N^{(l)} = \{y_1^{(l)}, \ldots, y_N^{(l)}\}$, the parameters are updated with the following formulae, $H_{i+1} = T_{yx} T_{xx}^{-1}, \quad F_{i+1} = T_{xx'}, T_{xx'},$ and $R_{i+1} = (Nm)^{-1} \text{diag}(T_{yx} - T_{yx} T_{xx}^{-1} T_{yx}^T)$ where $T_{yx}, T_{xx}, T_{xx'},$ and $T_{xx'}$ are defined as follows:

$$T_{yx} = \sum_{n=1}^{N} \sum_{l=1}^{m} E[y_n^{(l)} | x_n^{(l)} | Y_N^{(l)}, \theta_i], \quad T_{xx} = \sum_{n=1}^{N} \sum_{l=1}^{m} E[x_n^{(l)} | x_{n-1}^{(l)} | Y_N^{(l)}, \theta_i],$$

$$T_{xx'} = \sum_{n=1}^{N} \sum_{l=1}^{m} E[x_n^{(l)} | x_{n-1}^{(l)} | Y_N^{(l)}, \theta_i], \quad T_{x'x'} = \sum_{n=1}^{N} \sum_{l=1}^{m} E[x_{n-1}^{(l)} | x_{n-1}^{(l)} | Y_N^{(l)}, \theta_i].$$
These posterior expectations are efficiently calculated with the Kalman filtering and smoothing. We repeat this process until a suitable convergence criterion is satisfied.

Based on [3], we define a $p \times p$ matrix $\Psi = R^{-1/2}H \Lambda^{-1}H^T R^{-1/2}$ whose $(i, j)$th element $\Psi_{ij}$ represents the influence from gene $j$ to gene $i$ between consecutive times points. Finally, we construct gene networks with testing whether $\Psi_{ij}$ is zero or not using the permutation test.

3 Application

We constructed the gene network using apoptosis-related gene expression profiles of the Human Umbilical Vein Endothelial Cell [1]. The dataset includes expression profiles of 19,720 probes measured at 0.5, 1.5, 3, 6, 9, 12, and 24 hours after the induction of apoptosis. This experiment was repeated independently three times. We used 1,012 genes composed of 1,000 genes which have the highest coefficient of variations and 12 genes related to apoptosis. We will show the resulted network in our presentation.

Furthermore, we conducted numerical experiments with synthetic triplicate time course data for presenting the predictive power of the proposed method. We set the numbers of genes $p = 1,000$ and the state dimensions $k = 4$. First, we trained the model with synthetic triplicate and single data with time points $N = 7$. In order to evaluate test errors of the fitted models, we generated time course data with time points $N = 20$ drawn by the black line of Figure 1 (left). The green and red lines denote one step ahead predictions of the models trained with single and triplicate time course data, respectively. It is clear that the prediction based on triplicate data is much better than that based on single data.

Figure 1 (right) shows the ROC curves which indicate the correctness of the gene network estimation. Each circle denotes the sensitivity and the false positive rate cutoffed by the p-value 0, 0.01, 0.03, 0.05, 0.1, 0.2, …, 1.0. The ROC curves drawn by the red and blue lines are the results of estimated gene network by triplicate and single data with time points $N = 7$, respectively. Note that the estimation by triplicate data is far more accurate than that estimated by single data. In our presentation, we will show further evidences that the proposed method has the great potential to infer accurate gene networks.

References

