

State-Space Model for Gene Regulatory Networks with Time Delays

Fang-Xiang Wu¹, Wen-Jun Zhang^{1,2}, and Anthony J. Kusalik^{1,3}

Departments of ¹Biomedical Engineering, ²Mechanical Engineering, and ³Computer Science,
University of Saskatchewan, Saskatoon, SK, S7N 5A9, CANADA
faw341@mail.usask.ca; zhangc@engr.usask.ca; kusalik@cs.usask.ca

Abstract

This work proposes a state-space model to account for time delays in gene regulatory network. This model views genes as the observation variables, whose expression values depend on the current internal state variables and any external inputs. The Bayesian information criterion (BIC) and probabilistic principal component analysis (PPCA) are used to estimate the number of internal state variables and their expression profiles from gene expression data. By constructing dynamic equations with time delays for the internal state variables and the relationships between them and the observation variables (gene expression profiles), state-space models for gene regulatory networks with time delays are realized. The parameters of the proposed model may be unambiguously identified from time-course gene expression data with low computational cost. The method is applied to one time-course gene expression dataset, and the modes constructed. The results show that not only is the model (almost) stable, but also it has better prediction accuracy than a model without incorporating time delay.

1. Introduction

Many models have been proposed for gene regulatory networks, but they do not take into account time delay in cellular systems [1-3]. However, analysis of real gene expression data reveals a considerable number of time-delayed interactions suggesting that time delay is ubiquitous in gene regulation [4]. From a biological viewpoint, time delay in gene regulation arises from the delays characterizing the various underlying processes such as transcription, translation and transport. Dasika et al. [4] proposed a mixed integer linear programming framework for inferring time delay in gene regulatory networks. Due to the computational complexity of their algorithm, it is prohibitive to apply to gene regulatory networks with a large number of genes.

In this paper, we extend our earlier work [3] by proposing a state-space model to account for time delays in gene regulatory networks. As previously [3], genes are viewed as the internal state variables, which are estimated by BIC and PPCA from gene expression data (observation data of a cellular system). The model is applied to one time-course gene expression dataset [7]. The results suggest that it is possible to unambiguously determine gene regulatory network with time delays from time-course gene expression datasets. The constructed models are (almost) stable. Compared to the model without time delay, the new model has better prediction accuracy.

2. State-Space Model with Time Delays

The state-space model with time delays can mathematically be described by

$$\begin{cases} \mathbf{z}(t+1) = \sum_{\tau=0}^{\tau_{\max}} \mathbf{A}_{\tau} \cdot \mathbf{z}(t-\tau) + \mathbf{n}_1(t) \\ \mathbf{x}(t) = \mathbf{C} \cdot \mathbf{z}(t) + \mathbf{n}_2(t) \end{cases} \quad (1)$$

where the vector $\mathbf{x}(t) = [x_1(t) \ \dots \ x_n(t)]^T$ consists of the observation variables of the system and $x_i(t)$ ($i=1, \dots, n$) represents the expression level of gene i at time t , where n is the number of genes in the gene regulatory network under consideration. The vector $\mathbf{z}(t) = [z_1(t) \ \dots \ z_p(t)]^T$ consists of the internal state variables of the system and $z_i(t)$ ($i=1, \dots, p$) represents the expression value of internal element i at time t which directly regulates gene expression, where p is the number of the internal state variables. The matrices $\mathbf{A}_{\tau} = [a_{ij\tau}]_{p \times p}$ ($\tau=0, \dots, \tau_{\max}$) are the time translation matrices of the internal state variables or the state transition matrices with time delay τ , while the integer parameter τ_{\max} denotes the maximum time delay accounted for. They provide key information on the influences of the internal variables on each other. The matrix $\mathbf{C} = [c_{ik}]_{n \times p}$ is the transformation matrix between the observation variables and the internal state

variables. The entries of the matrix encode information on the influences of the internal regulatory elements on the genes. Finally, the vectors $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ stand for system error and observation error, respectively. The task of parameter identification in model (1) is to estimate the elements in matrices $\mathbf{A}_\tau = [a_{ij\tau}]_{p \times p}$ ($\tau = 0, \dots, \tau_{\max}$) and $\mathbf{C} = [c_{ik}]_{n \times p}$ such that both the system error and the observation error are minimized.

The building of model (1) from microarray gene expression data may be divided into two phases. Phase 1 extracts the internal state variables and their expression matrix using PPCA [5] to minimize the observation error (i.e., maximize the data likelihood) with BIC [6]. Phase 2 determines the state transition matrices $\mathbf{A}_\tau = [a_{ij\tau}]_{p \times p}$ from the expression matrix of the internal variables to minimize the system error.

To evaluate the model, three indices are employed: the computational cost, the prediction power, and the stability.

The *computational cost*: It is provable that the overall computational complexity of parameter identification of model (1) is $O(n)$. Such computational cost is much lower than that of other existing models such as the Boolean network model and differential/difference model [1,2,3].

The *stability*: As real gene networks are stable, inferred gene network models should be (almost) stable in order to be realistic. For our model, it can be proven that the model (1) is stable if and only if all eigenvalues of the following $(\tau_{\max} + 1) \times (\tau_{\max} + 1)$ block matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{0}_p & \mathbf{I}_p & \cdots & \mathbf{0}_p \\ \vdots & \vdots & \ddots & \mathbf{0}_p \\ \mathbf{0}_p & \mathbf{0}_p & \cdots & \mathbf{I}_p \\ \mathbf{A}_{\tau_{\max}} & \mathbf{A}_{\tau_{\max}-1} & \cdots & \mathbf{A}_0 \end{bmatrix} \quad (2)$$

lie in the unit circle in the complex plane, where \mathbf{I}_p is a $p \times p$ identity matrix and $\mathbf{0}_p$ is a $p \times p$ zero matrix.

The *prediction error*: Let $\hat{\mathbf{X}}$ be a data matrix with the same size as the original data matrix \mathbf{X} , which is computed from an initial state and the model derived from the data matrix \mathbf{X} . The prediction error (P_E) reflects how well $\hat{\mathbf{X}}$ approximates \mathbf{X} , and may be defined as:

$$P_E = \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{X}(i,:) - \hat{\mathbf{X}}(i,:) \right\|^2 / \left\| \mathbf{X}(i,:) \right\|^2 \quad (3)$$

where $\mathbf{X}(i,:)$ is the i -th row vector of gene expression data matrix \mathbf{X} (i.e. expression profile of the i -th gene), and $\left\| \mathbf{X}(i,:) \right\|$ is its Euclidean norm. Accordingly, the smaller the prediction error, the better the model.

3. Computational Experiments and Results

To highlight and evaluate the proposed model, we apply it to one gene expression dataset, and compare the results to our previous model [3]. The dataset consists of the expression data for 701 cell-cycle regulated genes with no missing data at 18 equally-spaced time points in the α -factor synchronized experiment [7]. After normalization, PPCA [5] and BIC [6] indicate that the datasets has 6 internal variables, and further estimate their expression matrix \mathbf{Z} and transformation matrix, \mathbf{C} . We assume $\tau_{\max} = 1$ for the dataset. By Using multivariate regression method [8], matrices \mathbf{A}_0 and \mathbf{A}_1 are estimated from the internal expression matrix, \mathbf{Z} .

To inspect the stability of the model inferred from the dataset, the eigenvalues of the matrix \mathbf{T} in (2) are calculated for the model. Matrix \mathbf{T} has twelve eigenvalues: two real numbers, and five pairs of conjugate complex numbers. All of these eigenvalues except for a single real-valued one lie inside the unit circle in the complex plane. However, the exception is very close to the boundary of the unit circle. Therefore, the model is (almost) stable. Further, when compared to the space-state model without time delays [3], the state-space model with time delay improves on the prediction error by about 70% for the dataset.

References

- [1] Liang, S., et al. "REVEAL, A general reverse engineering algorithm for inference of genetic network architectures" *Pacific Symposium on Biocomputing* 3: 18-29, (1998).
- [2] Chen, T., He, H. L., and Church, G. M. "Modeling Gene Expression with Differential Equations" *Pacific Symposium on Biocomputing* 4: 29-40, (1999).
- [3] Wu, F.X., Zhang, W.J., and Kusalik, A.J. "Modeling Gene Expression from Microarray Expression Data with State-Space Equations" *Pacific Symposium on Biocomputing* 9: 581-592, (2004).
- [4] Dasika, M., et al. "A Mixed Integer Linear Programming (MILP) Framework for Inferring Time Delay in Gene Regulatory Networks" *Pacific Symposium on Biocomputing* 9: 474-485, (2004).
- [5] Tipping, M. E. and Bishop C. M. "Probabilistic principal component analysis" *Journal of the Royal Statistical Society, Series B* 61: 611-622, (1999).
- [6] Burnham, K. P. and Anderson, D. R. "Model selection and inference: a practical information-theoretic approach" New York: *Springer*, (1998).
- [7] Spellman, P. T., et al. "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization" *Mol. Biol.* 9: 3273-3297, (1998).
- [8] Aoki, M. "State Space Modeling of Time Series" 2nd Edition, Berlin: Springer-Verlag, (1990).