

STATE-SPACE MODEL WITH TIME DELAYS FOR GENE REGULATORY NETWORKS

FANG-XIANG WU^{†,*}, W. J. ZHANG[†] and ANTHONY J. KUSALIK^{†,‡}

[†]*Division of Biomedical Engineering, [‡]Department of Computer Science
University of Saskatchewan, Saskatoon, SK, S7N 5A9, Canada*

**faw341@mail.usask.ca*

Received 19 May 2004

Revised 20 July 2004

A gene regulatory network can be considered a dynamic cellular system which describes the behavior (development) of a living cell and depends completely on the current internal state plus any external inputs, if these exist. Although many details inside a cell are not precisely known, gene expression data on a genome scale provide useful insights into such a cellular system. With gene expression data, a wide variety of models, such as Boolean networks and differential/difference equations, have been proposed to model gene regulatory networks. In these previously proposed models, genes are viewed as the internal state variables of a cellular system. This viewpoint has suffered from the underestimation of the model parameters. In addition, these models ignore an important problem with a gene regulatory network — time delay. Instead, this paper proposes a state-space model with time delays for gene regulatory networks. The proposed model views genes as the observation variables, whose expression values depend on the current internal state variables and any external inputs. 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 the internal state variables and the observation variables (gene expression profiles), state-space models with time delays for gene regulatory networks are constructed. The parameters of the proposed model can be unambiguously identified from time-course gene expression data with a lower computational cost. The proposed model is applied to two time-course gene expression datasets, and two gene regulatory networks are inferred, respectively. The analysis shows that the inferred gene regulatory networks have several features of the real gene regulatory networks, such as the stability, the robustness, and the periodicity. Further, compared to state-space models without time delays, the proposed model with time delays has better prediction accuracy.

Keywords: Gene regulatory network; cellular system; state-space model; time delay; gene expression; PPCA; BIC.

*Corresponding author.

1. Introduction

A gene regulatory network is a dynamic model to describe a cellular system in which a large number of different substances (such as mRNA, proteins, and metabolites) in a living cell interact and regulate cellular behaviors and functions. The unraveling of such cellular systems has proven useful in disease diagnosis and genomic drug design. Although currently we do not know exactly how these substances interact inside a cell, gene expression levels on the genomic scale produced by microarray or other high throughput measurement technologies provide very useful insights into a cellular system (ideally, we also want to measure levels of proteins and metabolites). Using gene expression data, a wide variety of models, such as Boolean networks [1, 2, 25, 32] and differential/difference equations [10, 12, 15], have been proposed to model cellular systems.

In a Boolean network model, a gene's expression (state) is simplified to being either completely "on" or "off". These states are often represented by the binary values 1 and 0, respectively, and the state of a gene is determined by a Boolean function of the states of all genes in the network. As the system proceeds from one time point to the next, the current states of all genes are used as input to Boolean functions which specify whether the states of genes will be "on" or "off" at the time point. Somogyi and Sniegoski [32] showed that such Boolean networks have features similar to those in biological systems, such as global complex behavior, self-organization, stability, redundancy, and periodicity. Liang *et al.* [25] described an algorithm for inference of gene network architectures in the context of Boolean network models. Their computational experiments showed that a small number of state transition pairs are sufficient to infer the original observations. Akutsu *et al.* [1] devised a much simpler algorithm for the same problem and proved that if the connectivity degree of genes (i.e., the maximum number of genes connected to each gene) is bounded by a constant h , only $O(\log n)$ state transition pairs (from all possible 2^n pairs) are necessary and sufficient to identify the original Boolean network of n genes correctly with high probability. Their algorithm was claimed to have time complexity $O(mn^{h+1})$ where m is the number of time points in gene expression dataset [2]. However, a further look finds that the "big O " notation hides a very large coefficient of 2^{2^h} in $O(mn^{h+1})$ [1, 2, 36].

In addition to Boolean network models, Chen *et al.* [10] proposed a differential equation model while D'haeseleer *et al.* [15] proposed a difference equation model. The task of identification in these models is the extraction of the "gene regulatory matrix" [10, 12, 15, 35] which contains n^2 elements for a gene network with n genes. Since the size of current gene expression datasets is typically much less than n^2 [6, 31], these models are usually under-determined. The existing literature describes two ways to address this issue. One way is to use a nonlinear interpolation scheme [15] to make up enough data for identifying n^2 elements. Such an interpolation scheme is *ad hoc*. Therefore, the reasonableness of the model built from such interpolated data is suspect. In addition, there exists a problem of "dimensional

disaster” when the number of genes in a model is large, for example, hundreds or thousands. The other way is to limit the connectivity degree of genes to be small, as in Boolean network models, so that the gene regulatory matrix is sparse. As a result, the number of identified parameters is much less than n^2 and the gene regulatory matrix becomes identified with current gene expression datasets. It can be proven [10] that the differential/difference model can be constructed in $O(n^{h+1})$ time, where h is the connectivity degree of genes.

In order that the parameters of the models are identifiable, both Chen [10] and Akutsu [1, 2] assumed that all genes have a fixed maximum connectivity degree h (often small). This assumption obviously contradicts the biological reality. For instance, some genes are known to have many regulatory inputs, while others are known to have not more than a few [6]. Another shortcoming of these methods is that the fixed maximum connectivity degree h is chosen in an *ad hoc* manner. Although de Hoon *et al.* [12] used Akaike’s Information Criterion (AIC) to determine the connectivity degree h of each gene for Chen’s differential model [10], they have not presented any efficient algorithm to identify the parameters of their differential equation model; the brute-force algorithm [12] has a computational complexity of $O(2^{n^2})$, where n is the number of genes in the model. Therefore, for biologically realistic regularity networks, the computational complexity of de Hoon’s method is prohibitive.

The state-space model is one of the most powerful methods to describe a dynamic system and has been widely employed for engineering control systems [9]. A state-space model consists of internal variables, external input variables, and output (observation) variables. Figure 1 shows a typical state-space model of a cellular system. In a state-space model, the observation variables typically depend on the internal variables, while the change in the internal variables is completely determined by the current internal variables plus any external inputs, if these exist. Interestingly, the aforementioned models are a variation of state-space models. However, in these models, genes were viewed as the internal state variables as well as observation variables of a cellular system, and their expression levels were the values of both the internal state variables and the observation variables. This viewpoint has suffered from the underestimation of the model parameters as pointed out previously. Actually, not all genes (their products, proteins) directly regulate gene expressions in a gene network since only a part of genes are translated into regulatory proteins which regulate gene expression while others are translated into structural proteins which do not regulate gene expression, but construct the tissues [3, 6, 26]. To explore this biological knowledge, recently a state-space model for gene regulatory networks was proposed [37], in which genes are viewed as the observation variables and gene expression dynamics is governed by a group of the internal variables. The Bayesian information criterion (BIC) was employed to determine the number of the internal variables and the factor analysis was used to estimate their expression profiles from the observation values of a cellular system, i.e., gene expression data.

In addition, the previous models have not taken into account time delay in a cellular system. However, the real microarray data example reveals a considerable number of time delayed interactions, suggesting that time delay is ubiquitous in gene regulation [11]. From a biological viewpoint, time delay in gene regulation arises from the delays characterizing the various underlying processes such as transcription, translation and transport. For example, time delays in regulation may stem from the time taken for the transport of a regulatory protein to its site of action. Dasika *et al.* [11] 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 it to gene regulatory networks with a moderate number of genes as considered in this paper.

In this paper, we extend our earlier work [37] by proposing a state-space model with time delays for gene regulatory networks. As before [37], 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 two time-course gene expression datasets [33]. The results suggest that it is possible to determine unambiguously gene network with time delays from time-course gene expression datasets. The inferred gene regulatory networks have several features of the real gene regulatory networks, such as the stability, robustness, and the periodicity. Compared to the model without time delay, the new model has better prediction accuracy.

2. Methods

2.1. State-space model with time delays

In Boolean network and differential/difference models for gene regulatory network, genes are viewed as state variables in a cellular system. This makes parameter identification of the models impossible without additional subjective assumptions when using the current volume of microarray gene expression datasets in which the number of genes is much larger than the number of time points. In addition, these models assume that regulatory relationships among genes are direct; for example, gene *j* directly regulates gene *i* with weight w_{ij} [10, 15]. In fact, genes may not be directly regulated in a cellular system, but by some regulatory internal variables [3, 6, 17].

In the state-space model for gene regulatory networks as shown in Fig. 1, genes are viewed as the observation variables, which are governed by some regulatory internal variables. Change in the internal states depends completely on the current internal states plus any external inputs, if these exist. The state-space model with time delays can mathematically be described by

$$\begin{cases} \mathbf{z}(t + 1) = \sum_{\tau=0}^{\tau_{\max}} \mathbf{A}_{\tau} \cdot \mathbf{z}(t - \tau) + \mathbf{n}_1(t), \\ \mathbf{x}(t) = \mathbf{C} \cdot \mathbf{z}(t) + \mathbf{n}_2(t), \end{cases} \tag{2.1}$$

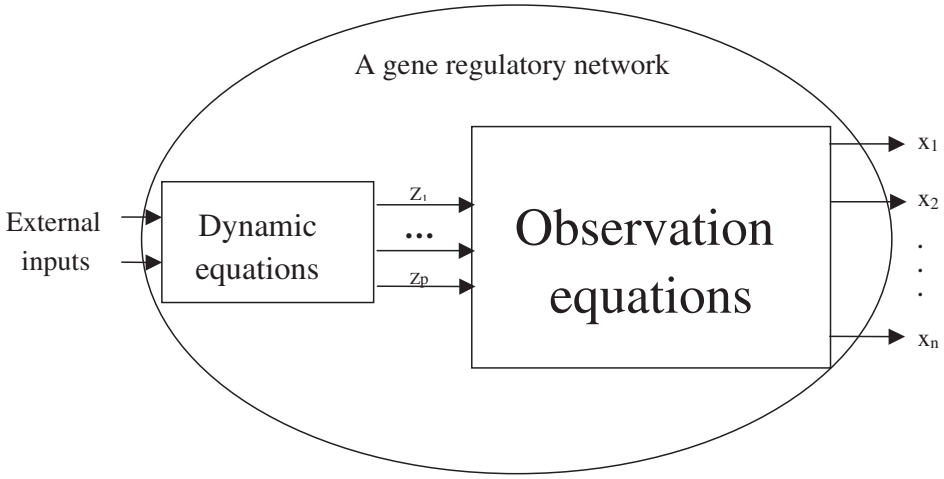


Fig. 1. A state-space model for gene regulatory networks, where x_i ($i = 1, \dots, n$) is the observation variables while z_i ($i = 1, \dots, p$) is the state variables.

where the vector $\mathbf{x}(t) = [x_1(t) \cdots x_n(t)]^T$ 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 $\mathbf{z}(t) = [z_1(t) \cdots z_p(t)]^T$ 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 regulatory internal variables on the genes. Finally, the vectors $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ stand for the system error and the observation error, respectively.

2.2. Model identification

The task of parameter identification in model (2.1) 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 with some senses. Let \mathbf{X} be the gene expression data matrix with n rows and m columns, where n and m are the numbers of the genes and the measuring time points, respectively. The building of model (2.1) from microarray gene expression data \mathbf{X} may be divided into two phases. Phase one extracts the internal state variables and their expression

matrix \mathbf{Z} with p rows and m columns from the data matrix x and computes the transformation matrix \mathbf{C} such that

$$\mathbf{X} = \mathbf{C} \cdot \mathbf{Z}. \tag{2.2}$$

Note that the matrices \mathbf{C} and \mathbf{Z} are dependent. After \mathbf{Z} is identified, \mathbf{C} may be calculated by formulae $\mathbf{C} = \mathbf{X} \cdot \mathbf{Z}^+$, where \mathbf{Z}^+ is a unique Moore–Penrose generalized inverse of the matrix \mathbf{Z} . Phase two builds the difference equations of the internal states, i.e., determines the state transition matrix \mathbf{A} from the expression matrix \mathbf{Z} . Phase one minimizes the observation error (i.e., maximize the data likelihood) with BIC while phase two minimizes the system error.

2.2.1. *Extraction of the internal variables*

Phase one is a key in the process of building model (2.1). There are many methods for latent variable analysis that may be used to extract the internal state variables and compute the transformation matrix from the gene expression data, i.e., to establish Eq. (2.2). For example, one may employ singular value decomposition [4, 20], where the characteristic modes or eigengenes may be viewed as the internal variables. However, in typical applications of singular value decomposition, the number of such internal variables is chosen in an *ad hoc* fashion, with the result that transformation matrix \mathbf{C} and the expression data matrix of the internal variables \mathbf{Z} are decided subjectively rather than from the data themselves. In our previous work [37], the maximum likelihood factor analysis and the EM algorithm [7, 24] were employed to extract the internal state variables and to compute the transformation matrix from the gene expression data. Nevertheless, the EM algorithm for maximum likelihood estimate may fall into a local maximum [14]. Tipping and Bishop [34] developed a probabilistic principal component analysis (PPCA) and proposed two methods for PPCA: maximum likelihood algorithm and EM algorithm. Further, they proved that the maximum likelihood algorithm for PPCA can find the global maximum.

In this paper, we employ PPCA [34] to extract the internal variables from time-course gene expression datasets, where \mathbf{X} is the $n \times m$ observation data matrix, each row of which is an observation sample; \mathbf{C} is the $n \times p$ transformation matrix, each row of which is a realization of latent variables; and \mathbf{Z} is the $p \times m$ loaded matrix, each row of which represents the expression profile of an internal state. It is assumed that the sample mean has been shifted to zero. The log-likelihood for PPCA model is expressed by

$$L = -\frac{n}{2} \{m(\ln 2\pi) + \ln |D| + \text{tr}(D^{-1}S)\}, \tag{2.3}$$

where $D = Z^T Z + \sigma^2 I$ and $\mathbf{S} = \mathbf{X}' * \mathbf{X} / n$. For the given number of internal variables, k , the log-likelihood for the PPCA model finds its global maximum [34]

$$L_k = -\frac{n}{2} \left\{ \sum_{j=1}^k \ln(\lambda_j) + (m - k) * \ln \left(\sum_{j=k+1}^m \lambda_j / (m - k) \right) + m(\ln(2\pi) + 1) \right\}, \tag{2.4}$$

when

$$\mathbf{Z}_k = \mathbf{R}(\mathbf{Q}_k - \sigma^2 \mathbf{I}_k)^{1/2} \mathbf{U}_k^T, \tag{2.5}$$

where λ_j ($j = 1, \dots, k$) are the first k largest eigenvalues of the sample variance matrix \mathbf{S} , the matrix \mathbf{Q}_k is a $k \times k$ diagonal matrix, whose diagonal elements are these λ_j ($j = 1, \dots, k$), \mathbf{U}_k is a $m \times k$ matrix, each column of which is a corresponding eigenvector of \mathbf{S} , \mathbf{I}_k is a $k \times k$ identity matrix, \mathbf{R} is an arbitrary $k \times k$ orthogonal matrix, and $\sigma^2 = \sum_{j=k+1}^m \lambda_j / (m - k)$.

Note that if $\{\mathbf{C}, \mathbf{Z}\}$ is an optimum solution of Eq. (2.2), $\{\mathbf{C}\mathbf{S}^{-1}, \mathbf{S}\mathbf{Z}\}$ is also its optimum solution, where \mathbf{S} is any $k \times k$ non-singular matrix. However, it can be proved that the state-space models from $\{\mathbf{C}, \mathbf{Z}\}$ and $\{\mathbf{C}\mathbf{S}^{-1}, \mathbf{S}\mathbf{Z}\}$ are equivalent [9]. Therefore, one may always normalize the expression profiles of the internal state variables. For the optimum number of internal state variables, k , since $\mathbf{R}(\mathbf{Q}_k - \sigma^2 \mathbf{I}_k)^{1/2}$ is a $k \times k$ non-singular matrix, we can take

$$\mathbf{Z} = \mathbf{U}_k^T, \tag{2.6}$$

as the expression profiles of the internal state variables. Further, the corresponding transformation matrix \mathbf{C} may be calculated.

2.2.2. Determination of the number of internal variables

From Eq. (2.4), the values of the maximum log-likelihood for the PPCA model increase with the increase of the number of internal state variables, k . The redundant internal state variables may result in a complicated model. Since the PPCA has a solid probabilistic foundation, we can employ some model selection criteria to determine the number of internal state variables. The model selection criteria which have been used in other model selection problems include the generalized likelihood ratio test (GLRT), the Akaike's information criterion (AIC), and the Bayesian information criterion (BIC) [8]. Both GLRT and AIC have a similar drawback; as the sample size increases there is an increasing tendency to accept a more complex model [14]. On the other hand, the BIC takes the sample size into account. Although the BIC method is developed from a Bayesian standpoint, the result is insensitive to the prior distribution for adequate sample size. Thus a prior distribution does

not need to be specified [28, 30], which simplifies the method. For each model, the BIC is calculated as

$$BIC(k) = 2 \cdot L_k - \ln(n) \cdot \nu_k, \tag{2.7}$$

where n is the sample size (the number of genes) and $\nu_k (=mk + 1)$ is the number of parameters in the PPCA model. The model with the largest BIC is chosen. BIC avoids over-fitting the model to the data. Since the term $nm(\ln(2\pi) + 1)/2$ in (2.4) is a constant for a given dataset, the calculation of BIC may be simplified as

$$BIC(k) = -n \left\{ \sum_{j=1}^k \ln(\lambda_j) + (m - k) \cdot \ln \left(\sum_{j=k+1}^m \lambda_j / (m - k) \right) \right\} - \ln(n) \cdot (mk + 1) \tag{2.8}$$

We denote the optimum number of internal variables selected with BIC in (2.8) by p hereafter.

2.2.3. Identification of the internal state model

After obtaining the expression data matrix of the internal variables z and the transformation matrix C in phase one, we develop the internal state transition equation (internal state model)

$$\mathbf{z}(t + \Delta t) = \sum_{\tau=0}^{\tau_{\max}} \mathbf{A}_\tau \cdot \mathbf{z}(t - \tau), \tag{2.9}$$

in model (2.1) from the data matrix \mathbf{Z} in phase two. Each state transition matrix \mathbf{A}_τ ($\tau = 0, \dots, \tau_{\max}$) contains p^2 unknown elements while the matrix \mathbf{Z} contains $m \cdot p$ known expression data points. If $(\tau_{\max} + 1)p > m$, Eq. (2.9) will be underdetermined. To find suitable state transition matrices, some additional conditions are necessary [10–12]. Using BIC, the number of chosen internal variables p generally is less than the number of time points m . Therefore, these matrices are often identifiable if there are just a few time delays (e.g., $\tau_{\max} \leq 1$) accounted for.

For equally spaced measurements of gene expression, the multivariable linear regression method [5, 19] may be used to identify state transition matrices, \mathbf{A}_τ ($\tau = 0, \dots, \tau_{\max}$). For unequally spaced measurements, the problem becomes nonlinear, and it is necessary to determine these matrices by using an optimization technique, such as those in Chapter 10 of Press’s text [27].

2.3. Model evaluation

Due to limitations in the understanding of real gene regulatory networks, it is difficult (if possible) to validate the models for inferred gene regulatory networks completely by biological experiments. Wessels *et al.* [35] proposed six indices to evaluate the models for gene regulatory networks from the viewpoint of bioinformatics. Some

of these indices are inapplicable to evaluation of the gene regulatory network models on real-life gene expression datasets because the real gene regulatory networks that created these data are unknown. In the following, five indices are introduced, including the computational cost, the prediction power, and the stability [35], the robustness, and the periodicity.

The *computational cost*: In phase one, the computational cost is bounded by the maximum likelihood method for the PPCA and is $O(mn + m^3)$ [34]. In phase two, the computational cost is $O(mp + p^3)$. Since both m and p are much smaller than n , overall the computational cost of the state-space model identification is $O(n)$, i.e., linear in the number of genes in a model. Such computational cost is much cheaper than that of other existing models such as the Boolean network model and differential/difference model.

The *stability*: due to the limited energy and storage within a cell, concentrations of gene expression products such as mRNA should remain bounded. All real gene networks are therefore stable. Consequently, the inferred gene network models should also be (almost) stable in order to be realistic. For our model, this is equivalent to dynamic Eq. (2.9) being stable. It can be proven that the dynamic model (2.9) is stable if and only if all eigenvalues of the following $(\tau_{\max} + 1) \times (\tau_{\max} + 1)$ block matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{O}_p & \mathbf{I}_p & \cdots & \mathbf{O}_p \\ \vdots & \vdots & \ddots & \mathbf{O}_p \\ \mathbf{O}_p & \mathbf{O}_p & \cdots & \mathbf{I}_p \\ \mathbf{A}_{\tau_{\max}} & \mathbf{A}_{\tau_{\max}} - \mathbf{1} & \cdots & \mathbf{A}_0 \end{bmatrix}, \tag{2.10}$$

lie inside the unit circle in the complex plane, where \mathbf{I}_p is a identity matrix and \mathbf{O}_p is a $p \times p$ zero matrix, and \mathbf{A}_τ ($\tau = 0, \dots, \tau_{\max}$) are state transition matrices with time delay τ in Eq. (2.9).

The *periodicity*: Certain biological processes are periodic. The cell-cycle and circadian clock, for example, repeat at well-defined and reliable intervals. Studies have shown that gene regulatory networks associated with these periodic biological processes are themselves rhythmic [6, 21, 23]. Therefore, the inferred gene regulatory networks associated with these periodic biological processes should be periodic at its stable states. Accordingly, the periodicity of system (2.1) at its stable state is determined by its dominant eigenvalues [the eigenvalues of matrix \mathbf{T} in (2.10) whose moduli are the largest].

The *robustness*: The robustness of a gene regulatory network is understood as its insensitivity to noise or disturbance. It is obvious that a real gene regulatory network has robustness [18, 22]. Therefore, the inferred gene regulatory network should be robust. Accordingly, the stability of a linear system implies some of its robustness [9]. Note that the stability, the robustness, and the periodicity of the system (2.1) are all related to the eigenvalues of matrix \mathbf{T} in (2.10).

The *prediction power (error)*: Let $\hat{\mathbf{X}}$ be a data matrix with the same size as the original data matrix X , which is computed from an initial state and the

model derived from the data matrix X . The prediction error reflects how well $\hat{\mathbf{X}}$ approximates \mathbf{X} . The prediction error P_E may be defined as:

$$P_E = \frac{1}{n} \sum_{i=1}^n \|\mathbf{X}(i, :) - \hat{\mathbf{X}}(i, :)\|^2 / \|\mathbf{X}(i, :)\|^2, \tag{2.11}$$

where $\mathbf{X}(i, :)$ is the i th row vector of gene expression data matrix X (i.e., the expression profile of the i th gene). $\|\mathbf{X}(i, :)\|$ is the Euclidean norm of the vector $\mathbf{X}(i, :)$. Intuitively, the smaller the prediction error, the bigger the prediction power. Wessels *et al.* [35] define the prediction power as:

$$P_P = 1 / (1 + E_{MSE}),$$

where $E_{MSE} = \frac{1}{nm} \sum_{i=1}^n \|\mathbf{X}(i, :) - \hat{\mathbf{X}}(i, :)\|^2$. Obviously, the scale of \mathbf{X} 's elements influences the value of E_{MSE} and further influences the value of P_P . For example, one may always multiply by a smaller constant to decrease E_{MSE} and thus increase P_P while the model has no improvements. On the other hand, P_E in (2.11) is invariant to the scale of \mathbf{X} . Therefore, it is more reasonable to evaluate the models. We will use the definition of the prediction error by (2.11) to evaluate the models.

3. Computational Experiments and Results

To evaluate the proposed model, we apply it to two gene expression datasets, and compare the result to a previous model [37]. These two datasets are from Spellman *et al.*'s experiment for studying cell cycle-regulated genes in yeast *Saccharomyces cerevisiae* [33]. Dataset 1 consists of the expression data for 701 cell-cycle regulated genes with no missing data at a total of equally-spaced 18 time points in the α -factor-synchronized experiment. Dataset 2 consists of the expression data for 789 cell cycle regulated genes with no missing data at a total of equally-spaced 14 time points in the elutriation-synchronized experiment. These two dataset is available at <http://genome-www.stanford.edu/SVD/>.

Before PPCA, a \log_2 -transformation as applied to all original intensity ratios (Cy5/Cy3) of gene expression, and the expression profile for each gene was normalized to have a median of 0 and a standard deviation of 1. Further we normalize the expression values of all genes on each microarray so as to have a mean of 0 and a standard deviation of 1. Thus in PPCA, we do not need to estimate the mean in the PPCA model [24].

The maximum likelihood algorithm for PPCA [24] is employed to analyze the two datasets. For a variety of number k of internal state variables, $BIC(k)$ is calculated by Eq. (2.6). Figures 2 and 3 depict the profiles of BIC with respect to the number of internal variables for Datasets 1 and 2, respectively. With BIC, one may conclude that 6 is the optimum number of internal variables for Dataset 1 from Fig. 2 while 4 is the optimum number of internal variables for Dataset 2 from Fig. 3.

After the numbers of the internal variables are found, the expression matrices of the internal state variables, Z 's, may be given by (2.6), and further so are the

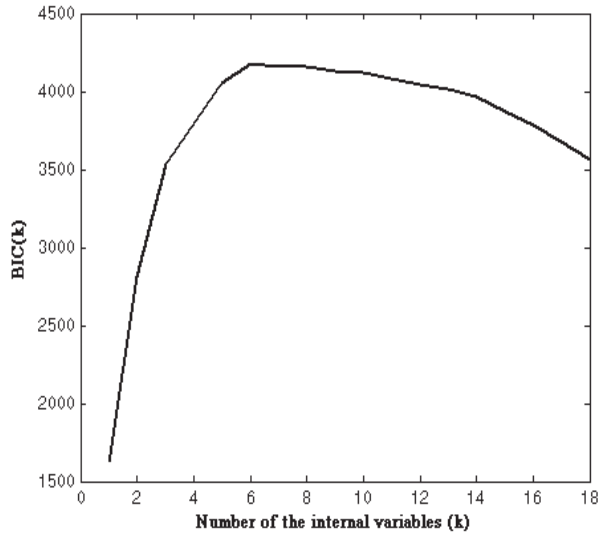


Fig. 2. Plot of BIC with respect to the number of internal variables for Dataset 1.

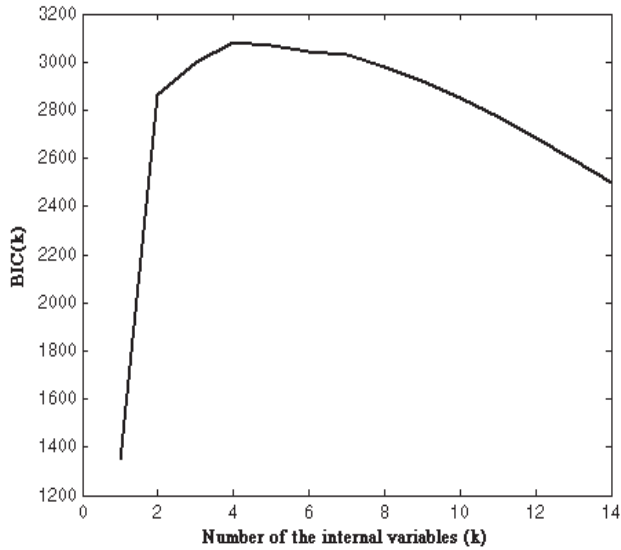


Fig. 3. Plot for BIC with respect to the number of internal variables for Dataset 2.

corresponding transformation matrices, C 's. Since these two datasets are collected at equally spaced time points, the multivariate regression method [8] is employed to determine the state transition matrices A_τ ($\tau = 0, \dots, \tau$) in the models from the internal expression matrices, Z 's. In this work, we take $\tau_{\max} = 1$ for both datasets.

To inspect the stability, the robustness, and the periodicity of these two inferred gene regulatory networks for genes in these two datasets, we calculate the eigenvalues of the matrix T in (2.10) for inferred gene regulatory networks, respectively. For Dataset 1 with $\tau_{\max} = 1$, the matrix T in (2.10) has 12 eigenvalues: two real numbers, 1.0073 and -0.4074 ; and five pairs of conjugate complex numbers, $0.6244 \pm 7757i$, $0.7282 \pm 0.6498i$, $-0.8580 \pm 0.0588i$, $0.1448 \pm 0.8083i$, $0.6498 \pm 0.4086i$. All of these eigenvalues except for the first real eigenvalue lie inside the unit circle in the complex plane. However, the first real eigenvalue is very close to the boundary of the unit circle as shown in Fig. 4. This means that the inferred regulatory network for genes in Dataset 1 is almost stable and robust. Furthermore, the dominant eigenvalues of the network are two pairs of conjugate complex numbers and a real number which are very close to the unit circle. Accordingly, this implies that at the stable states, the network behaves periodically. This result is not surprising because the genes in Dataset 1 are cell-cycle regulated.

For Dataset 2 with $\tau_{\max} = 1$, the matrix T in (2.10) has eight eigenvalues: two real numbers, 0.5344 and -0.2357 ; and three pairs of conjugate complex numbers: $-0.8079 \pm 0.6862i$, $0.8268 \pm 4940i$, and $0.0480 \pm 0.8810i$. All of these eigenvalues except for $-0.8079 \pm 0.6862i$ lie inside the unit, but their modulus is 1.0600 and is very close to the unit circle in the complex plane as shown in Fig. 5. This means the inferred regulatory network for genes in Dataset 2 is almost stable and robust. Furthermore, the dominant eigenvalues of the network are two pairs of conjugate complex numbers which are very close to the unit circle. Accordingly, this implies that at the stable states, the inferred network behaves periodically. Again, this result is not surprising because the genes in Dataset 2 are cell-cycle regulated as well.

Figures 6 and 7 depict comparisons of the internal state profiles estimated by PPCA and predicted by the dynamic model (2.9) for two datasets. These figures show that two kinds of profiles match very well for both datasets. Furthermore, to quantitatively evaluate the state-space models with time delays, we also calculate their prediction errors P_E , and compare against our previously proposed state-space models without time delays [37]. The results of these calculations are listed in Table 1, where the *improvement* is defined as

$$P_E \text{ (improvement)} = \frac{P_E \text{ (without time delay)} - P_E \text{ (with time delay)}}{P_E \text{ (without time delay)}}. \quad (3.1)$$

From Table 1, the PREDICTION error of the space-state model without time delays [37] is 0.0844, while the prediction error of the model in this paper is 0.0258 for Dataset 1. Comparing the state-space model without time delays, the state-space model with time delays improved the prediction error by about 70% for Dataset 1. Similarly, for Dataset 2 the state-space model with time delays improved the prediction error by about 60%, as compared to the state-space model without time delays. These results illustrate that the state-space model with time delays outperforms the model without time delays [37] for gene regulatory networks.

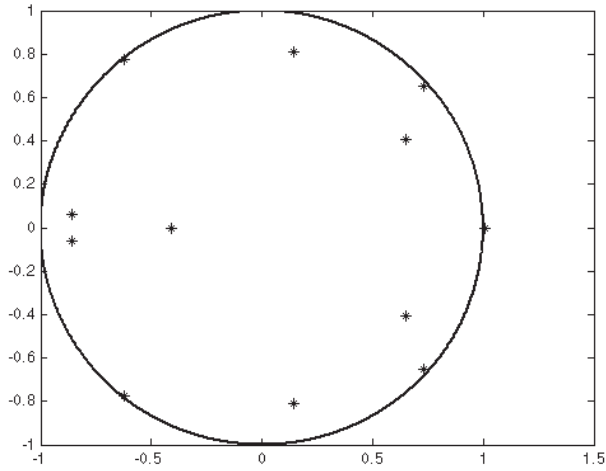


Fig. 4. The distribution of eigenvalues of the inferred gene network from Dataset 1.

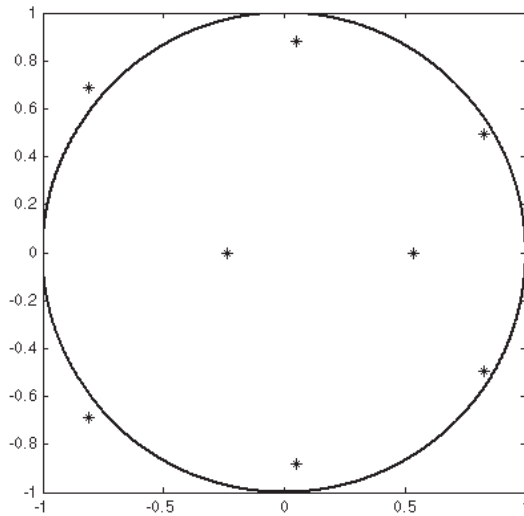


Fig. 5. The distribution of eigenvalues of the inferred gene network from Dataset 2.

4. Conclusion and Discussion

This paper has proposed a state-space model with time delays for gene regulatory networks. Applications of the presented model to two previously published gene expression datasets has showed that some features of the models are consistent with biological knowledge. For example, genes are regulated by some regulatory internal variables [6, 37], and the inferred gene regulatory networks have stability,

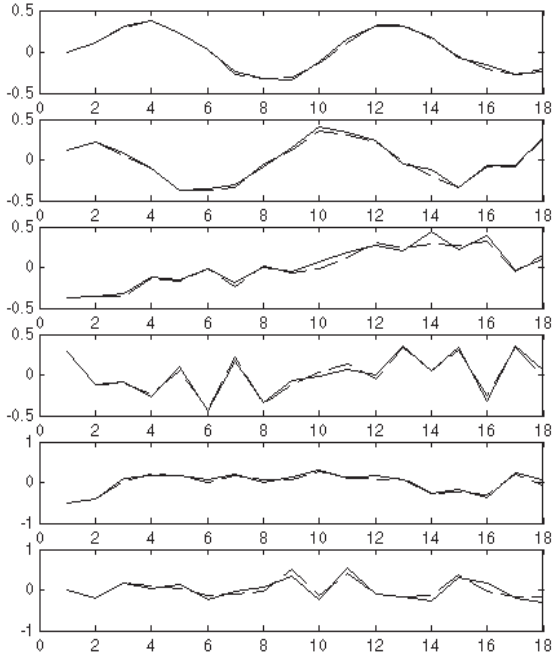


Fig. 6. A comparison of six internal state expression profiles estimated by PPCA and predicted by the dynamic model (2.9) for Dataset 1. The solid line: estimated profiles; and the dash lines: predicted profiles.

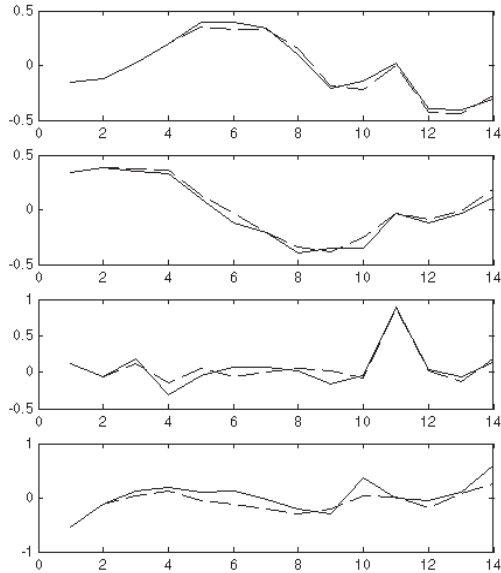


Fig. 7. A comparison of four internal state expression profiles estimated by PPCA and predicted by the dynamic model (2.9) for Dataset 2. The solid line: estimated profiles; and the dash lines: predicted profiles.

Table 1. Comparison of prediction power between the state-space model with time delays and without time delays.

P_E	Dataset 1	Dataset 2
Without time delays	0.0844	0.1286
With time delays	0.0258	0.0519
Improvement (%)	69.43	59.64

robustness [18, 22], periodicity [21, 23] and time delays [13, 29], which are the properties a real gene regulatory network has. Further, compared to the state-space model without time delays [37], the space-state model with time delays has more prediction power.

Compared to Boolean network models and differential/difference models, the proposed model (2.1) has the following characteristics. First, genes are viewed as observation variables rather than internal state variables. In fact, microarray experiments provide us only information of observation variables (gene expression data), not direct information about the internal variables of a cellular system. Therefore, the viewpoint in this model is more reasonable. Furthermore, this viewpoint yields a lower computational cost for model identification as compared to models such as Boolean network models [1, 2, 25, 32] and differential/difference models [10, 12, 15]. Second, the proposed model takes into consideration time delays in gene regulatory networks. Finally, from a biological angle, the proposed model (2.1) can capture the fact that genes may be regulated by some regulatory internal variables [6, 17].

It is interesting to compare the proposed model with the Bayesian network model. Hartemink *et al.* [17] proposed a Bayesian network model for gene regulatory networks. Their Bayesian network model permits the latent variables capturing unobserved factors and employs the BIC to select the model. However, the latent variables there are some known regulatory proteins, and the regulatory relationships among genes and the latent variables in the candidate models are also known. For many practical situations, such pieces of information are unknown. In addition, with their model the number of model parameters exponentially increases with the number of genes in a network. Thus the computational complexity of the model identification prohibits their model to be applied to the network with a moderate number of genes.

On the other hand, our model does have some shortcomings. For example, the model is linear and can only capture the primary linear components of a biological system, which may be nonlinear. However, with current scale of gene expression data, it is too difficult (if possible) to construct a nonlinear model. In principle, any nonlinear system can be approximated by a linear system. Studies have showed that a linear model can capture some key characteristics of nonlinear gene regulatory systems [16]. In the near term, the proposed model will be applied to more datasets, and its biological features need to be explored further.

In addition, one important exercise of future work is to study the biological relevance of the internal variables. According to the principle of gene regulation process [3], the internal variables should reflect the information of the regulatory proteins, which involve the regulation process of genes in the network. Unlike the Bayesian network model [17], such regulatory proteins are unknown in our model. Fortunately, with advances in proteomics [26] it is possible to employ the expression of proteins involving a gene regulatory process to investigate the biological meaning of the internal variables in our model. This goal requires closer collaboration with molecular biologists.

We cannot expect to obtain perfect gene regulatory network models, which can completely explain organismal or suborganismal behaviors, from the current volume of gene expression datasets at this time. On the other hand, any subjective assumptions-enforced models may result in misunderstanding (or misinterpreting) organismal or suborganismal behaviors. Using the proposed model one may infer sound generic regulatory networks (which are what data can tell us without any subjective assumptions) from the current volume of gene expression datasets. We believe that the model proposed in this paper, along with the results of the applications to two datasets, has advanced the art of gene regulatory network modeling from time-course gene expression datasets.

Acknowledgments

We thank Natural Sciences and Engineering Research Council of Canada (NSERC) for partial financial support of this research. The first author thanks University of Saskatchewan for funding him through a graduate scholarship award. We also want to thank anonymous referees for their valuable comments.

References

- [1] Akutsu T *et al.*, Identification of gene networks from a small number of gene expression patterns under the Boolean network model, *Pac Symp Biocomput* 4:17–28, 1999.
- [2] Akutsu T, Miyano S, Kuhara S, Algorithms for identify Boolean networks and related biological networks based on matrix multiplication and fingerprint function, *Proc 4th Annu Int Conf Res in Comput Mol Bio*, Tokyo, Japan, pp. 8–14, 2000.
- [3] Alberts B *et al.*, *Essential Cell Biology*, Garland, New York, 1998.
- [4] Alter O, Brown PO, Botstein D, Singular value decomposition for genome-wide expression data processing and modeling, *Proc Natl Acad Sci USA* 97:10101–10106, 2000.
- [5] Aoki M, *State Space Modeling of Time Series*, 2nd edn., Springer-Verlag, Berlin, 1990.
- [6] Baldi P, Hatfield GW, DNA microarrays and gene expression: From experiments to data analysis and modeling, Cambridge University Press, New York, 2002.
- [7] Bubin DB, Thayer DT, EM algorithms fro ML factor analysis, *Psychometrika* 47:69–76, 1982.

- [8] Burnham KP, Anderson DR, *Model Selection and Inference: A Practical Information-Theoretic Approach*, Springer, New York, 1998.
- [9] Chen CT, *Linear System Theory and Design*, 3rd edn., Oxford University Press, New York, 1999.
- [10] Chen T, He HL, Church GM, Modeling gene expression with differential equations, *Pac Symp Biocomput* **4**:29–40, 1999.
- [11] Dasika M *et al.*, A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks, *Pac Symp Biocomput* **9**:474–485, 2004.
- [12] de Hoon MJL *et al.*, Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations, *Pac Symp Biocomput* **8**:17–28, 2003.
- [13] de Jong H, Modelling and simulation of genetic regulatory systems: A literature review, *J Comput Bio* **9**:67–103, 2002.
- [14] Dempster AP, Laird NM, Rubin DB, Maximum likelihood from incomplete data via the EM algorithm, *J Roy Stat Soc B* **39**:1–38, 1977.
- [15] D’haeseleer P *et al.*, Linear modeling of mRNA expression levels during CNS development and injury, *Pac Symp Biocomput* **4**:41–52, 1999.
- [16] Gardner TS, di Bernardo D, Lorenz D, Collins JJ, Inferring genetic networks and identifying compound mode of action via expression profiling, *Science* **301**:102–105, 2003.
- [17] Hartemink AJ *et al.*, Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, *Pac Symp Biocomput* **6**:422–433, 2001.
- [18] Hartwell LH *et al.*, From molecular to modular cell biology, *Nature* **402**:C47–52, 1999.
- [19] Harvey AC, *Time Series Models*, 2nd edn., The MIT Press, Cambridge MA, 1993.
- [20] Holter NS *et al.*, Dynamic modeling of gene expression data, *Proc Natl Acad Sci USA* **98**:1693–1698, 2001.
- [21] Kauffman SA, *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, Oxford, 1993.
- [22] Kitano H, Computational systems biology, *Nature* **420**:206–210, 2002.
- [23] Langmead CJ *et al.*, Phase-independent rhythmic analysis of genome-wide expression patterns, *Proc Ann Int Conf Res Comput Mol Bio*, Washington DC, USA, pp. 205–215, 2002.
- [24] Lawley DN, Maxwell AE, *Factor Analysis as a Statistical Method*, 2nd edn., Butterworth, London, 1971.
- [25] Liang S *et al.*, REVEAL, A general reverse engineering algorithm for inference of genetic network architectures, *Pac Symp Biocomput* **3**:18–29, 1998.
- [26] Liebler DC, *Introduction to Proteomics*, Humana Press, Totowa, NJ, 2002.
- [27] Press WH *et al.*, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edn., Cambridge University Press, Cambridge, UK, 1992.
- [28] Raftery AE, Choosing models for cross-classification, *Am Sociol Rev* **51**:145–146, 1986.
- [29] Rosenfeld N, Alon U, Response delays and the structure of transcription networks, *J Mol Bio* **329**:645–654, 2003.
- [30] Schwarz G, Estimating the dimension of a model, *Ann Stat* **6**:461–464, 1978.
- [31] Sherlock G *et al.*, The Stanford microarray database, *Nucleic Acids Res* **29**:152–155, 2001.
- [32] Somogyi R, Sniegoski CA, Modeling the complexity of gene networks: Understanding multigenic and pleiotropic regulation, *Complexity* **1**:45–63, 1996.

- [33] Spellman PT *et al.*, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol Biol* **9**:3273–3297, 1998.
- [34] Tipping ME, Bishop CM, Probabilistic principal component analysis, *J Roy Stat Soc B* **61**:611–622, 1999.
- [35] Wessels LFA, Van Someren EP, Reinders MJT, A comparison of genetic network models, *Pac Symp Biocomput* **6**:508–519, 2001.
- [36] Wu FX, Analysis and modeling of gene expression data from DNA microarray experiments, Technical Report, Division of Biomedical Engineering, University of Saskatchewan, Oct., 2003.
- [37] Wu FX, Zhang WJ, Kusalik AJ, Modeling gene expression from microarray expression data with state-space equations, *Pac Symp Biocomput* **9**:581–592, 2004.