A Comparative Study of the Time-Series Data for Inference of Gene Regulatory Networks Using B-Spline

Haixin Wang, James E. Glover and Lijun Qian

Abstract—In this paper, the quantitative analysis of time-series gene expression data on inference of gene regulatory networks is performed. Time-series gene data are modeled by the B-Spline algorithm to improve the overall smooth expression curves which can further reduce over-fitting. The effect of the different sizes of observed time-series data on gene regulatory networks inference is analyzed. The stochastic errors introduced by the B-Spline algorithm to the system are evaluated. The precision of different sizes of time-series data on parameter estimations is compared. With application of the B-Spline to generate continuous curves, simulation results can be much more accurate and inference results are significantly improved. Both synthetic data and experimental data from microarray measurements are used to demonstrate the effectiveness of the proposed method.

I. INTRODUCTION

Gene-regulatory networks (GRNs) are collections of DNA segments in a cell which interact with each other and with other substances in the cell, thereby governing gene transcriptions. In light of the recent development of high-throughput DNA microarray technology, it becomes possible to discover GRNs, which are complex and nonlinear in nature. Specifically, the increasing existence of microarray time-series data makes possible the characterization of dynamic nonlinear regulatory interactions among genes. The modeling, analysis, and control of GRNs are critical for finding medicine for gene-related diseases.

One issue during the inference of GRNs is that not enough time-series experiment data are available [1]. This makes many inference processes not realizable. Many models are restricted by limited data. B-Spline can reconstruct unobserved gene expression data or recover the missing data. Compared to the B-Spline algorithm, simple algorithms, for example linear interpolation, can lead to poor estimation for the missing time-series data, especially when the sampled data is not uniform [2].

In this paper, the ordinary differential equation (ODE) model using S-system is adopted to infer GRNs [3]. Different sizes of time-series data are analyzed. The B-Spline algorithm is introduced to analyze the raw data. The effects of the time-series data to infer gene regulatory networks are analyzed in synthetic model and microarray experimental expression data.

Haixin Wang and James E. Glover are with the Department of Mathematics and Computer Science, Fort Valley State University, Fort Valley, Georgia, USA (email: wangh@fvsu.edu).

Lijun Qian is with the Department of Electrical and Computer Engineering, Prairie View A&M University, Prairie View, Texas, USA (email: LiQian@pvamu.edu).

II. INTRODUCTION TO B-SPLINE ALGORITHM

In numerical analysis, B-Spline has been widely used in image processing, computer-aided design and statistical analysis [4], [5]. It is one of the most flexible formulations for many engineering applications. A spline curve is a sequence of curve segments that are connected together to form a single continuous curve. In this study, it is a time-related sequence [6]. The basis B-Splines of degree m can be expressed as

$$ S(t) = \sum_{i=0}^{N-m-1} P_i b_{i,k}(t), \quad t \in [d_m, d_{k-m}] $$

where \( k \) is the order of the polynomial basis, \( P_i \) are called control points, and the normalized B-Spline basis \( b_{i,k}(t) \) can be expressed using the Cox-deBoor recursion formula [6]:

$$ b_{i,1}(t) = \begin{cases} 1, & d_i \leq t \leq d_{i+1} \\ 0, & \text{otherwise} \end{cases} $$

$$ b_{i,k}(t) = \frac{(t-d_i) b_{i,k-1}(t) + (d_{i+k} - t) b_{i+1,k-1}(t)}{d_{i+k} - d_i}. $$

Given \( N + 1 \) time series datasets with knots \( (d_i) \) that satisfying

$$ d_0 \leq d_1 \leq \cdots \leq d_N, $$

the corresponding control points \( (P_i) \) need to be calculated. Suppose \( t \in [d_m, d_{k-m}] \) and \( P_i^0 = P_i \) for \( i = m, \cdots, k-m \)

$$ P_i^h = (1 - \alpha_{h,i}) P_{i-1}^h + \alpha_{h,i} P_i^{h-1} $$

where \( h = 1, \cdots, k; \quad i = 0, \cdots, N - m - 1; \)

$$ \alpha_{h,i} = \frac{t - P_i}{P_{i+k+1-h} - P_i}. $$

A polygon can be constructed by connecting the control points with line segments. From those control points, the corresponding data points can be calculated.

In this study, we select the uniform cubic B-Spline to insert time-series data, and \( k \) is chosen to be 4. Then the basis can be expressed as

$$ S_i(t) = \begin{bmatrix} u^3 & u^2 & u & 1 \end{bmatrix} \begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 0 & 3 & 0 \\ 1 & 4 & 1 & 0 \end{bmatrix} \begin{bmatrix} P_{i-1} \\ P_i \\ P_{i+1} \\ P_{i+2} \end{bmatrix} $$

where \( u \in [0, 1] \).

The general process of inserting time series data via B-spline algorithm is shown in Algorithm 1.

1. **Algorithm 1:**
Algorithm 1 General time-series data processing by B-Spline

Require:
The raw data: \(d_0 \cdots d_N\)
The insertion data length between two insertion data points: \((d_i, d_{i+1})\)
Spline order: \(k = 4\)

Ensure: The data set \(S_i\)
1) Calculate the control points \(P_i\) from the original data \(d_i\)
2) Determine the insertion data \(S_i\) according to equation (1) with initial conditions: spline order and insertion data length

III. The adopted ODEs model to infer gene regulatory networks

In general, modeling gene regulatory networks is a nonlinear identification problem. Assuming there are \(N\) genes of interest and \(x_i\) denotes the state (such as the microarray reading) of the \(i^{th}\) gene, then the dynamics of the GRN may be modeled as

\[
\frac{dx_i}{dt} = f_i(x_1, x_2, \ldots, x_N) \quad i = 1, 2, \ldots, N
\]  

(2)

where the nonlinear functions \(f_i\) need to be determined from time-series microarray measurements. In this study, we assume that the functions \((f_i, \forall i)\) are in the form [7]:

\[
f_i = \sum_{j=1}^{L_i} \left[ w_{ij} \Omega_{ij}(x_1, x_2, \ldots, x_N) \right] \quad i = 1, 2, \ldots, N
\]  

(3)

where \(L_i\) is the number of terms in \(f_i\), \(w_{ij}\) are the parameters to be estimated and \(\Omega_{ij}(x_1, x_2, \ldots, x_N)\) is the \(j^{th}\) component of the nonlinear function \(f_i\). A two-step nested optimization procedure is proposed to identify the nonlinear differential equation for each individual gene. Genetic programming (GP) is applied to determine the nonlinear parameters (global optimization) and then the corresponding parameters associated with each term are estimated by Kalman filtering (local optimization) in each iteration. Such a decomposition of the problem into a structural part solved by GP and a parameter optimization part solved by Kalman filtering reduces the complexity significantly and speeds up convergence. In this paper the S-system is adopted for function \(f_i\). The S-system model is given by [3]:

\[
\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{N} x_j^{g_{i,j}} - \beta_i \prod_{j=1}^{N} x_j^{h_{i,j}}, \quad (i = 1, \ldots, N)
\]  

(4)

where \(x_i\) is the state variable, \(\alpha_i\) and \(\beta_i\) are the positive rate constants, \(g_{i,j}\) and \(h_{i,j}\) are the exponential parameters called kinetic orders. If \(g_{i,j} > 0\), gene \(j\) will induce the expression of gene \(i\). On the contrary, gene \(j\) will inhibit the expression of gene \(i\) if \(g_{i,j} < 0\). \(h_{i,j}\) will have the opposite effects on controlling gene expressions compared to \(g_{i,j}\). S-system is a quantitative model which is characterized by power-law functions. It has the rich structure capability of capturing various dynamics in many biochemical systems [8].

In addition, the S-system model has been proven to be successful in modeling GRNs [9], [10], [11], [12], [13], [14]. Hence, the S-system model is adopted for modeling GRNs in this paper. More importantly, the B-Spline algorithm is introduced into the process to improve the quality of the time-series data sets. The general process of inferring the gene regulatory networks is shown in Fig. 1.

IV. The analysis of time-series data using B-Spline

In this section, the effects of gene expression data using the B-Spline algorithm are discussed in the optimization estimation process.

A. The effect of the sampling data size to infer GRNs

For microarray experiments, the time-series data are important to infer gene regulatory networks. The sampling data sizes from the experiment have a significant impact on the resulted GRNs. For insufficient data sizes, the inferred GRNs can have errors. Hence, the effects of the various data sizes on the inference of GRNs are analyzed in this section. In order to show the relation, the coefficient of determination (CoD) is applied [15]. It provides a measure of how well future outcomes are likely to be predicted by the model. The
The Number of Sampling Data

Average CoD

B−Spline

B−Spline

Linear

data4

Fig. 2. The effect of the sampling data size on the estimator.

CoD is given by

\[ SS_{err} = \sum_{i=0}^{N} (y_i - f_i)^2 \]

\[ SS_{tot} = \sum_{i=0}^{N} (y_i - \bar{y})^2 \]

\[ R^2 = 1 - \frac{SS_{err}}{SS_{tot}} \]

where \( SS_{err} \) is the sum of the squared errors between observed data and data from the inferred model. \( SS_{tot} \) is the sum of the squared errors between observed data and its mean. \( R^2 \) is the CoD that indicates how well the regressing line fits the data. An \( R^2 = 1 \) indicates that the regression line perfectly fits the data. \( y_i \) is the observed data, \( f_i \) is the modelled data and \( \bar{y} \) is the mean of the observed data. The process is shown by Algorithm 2.

Algorithm 2  The general process of analyzing the effect of data size to infer GRNs

Require:

The observed data \( x_i \) with size \( n \)

Ensure: CoD of each GRNs

1) Apply two-step method to infer GRNs
2) Analyze CoD for each GRNs

As an example, the result is shown in Fig. 2 for the following synthetic model:

\[ x_1' = x_1^{1.5} x_2^{1.2} - x_2^{0.2} \]

\[ x_2' = x_1^{0.1} - x_1 x_2^{0.5} \]

From Fig. 2, the averaged CoD from the estimator becomes larger with the increased amount of sampling data. The CoD does not change much after certain number of samples, which implies that the CoD has an upper bound. It is observed that increasing the number of sampling points yields more exact estimates, within the upper bound. Note that the number of samples in the time-series data from microarray experiment is usually very small. Therefore, it is necessary to introduce the B-spline into the process of inference of GRNs. It is also noticed that B-Spline can reach a better result than that of a linear interpolation algorithm before approaching the upper bound.

B. The data error introduced by the B-Spline

B-Spline algorithm is introduced to interpolate more data from the original sampling data with some missing data. It is necessary to analyze whether or not the B-Spline algorithm introduces cumulative data errors. The mean square errors \( (MSE = \sum_{i=0}^{N} (y_i - f_i)^2) \) for different sizes of the data in B-Spline algorithm are stochastically calculated and compared to the observed data. The result is shown in Fig.3. The synthetic data range is \([0.1, 1]\). The MSE range by B-Spline algorithm is \([0.02, 0.25]\). The MSE range by linear interpolation is \([0.029, 0.81]\). From Fig. 3, it is observed that the MSE remains very small and it does not increase with interpolated data by B-Spline algorithm. Therefore, B-Spline algorithm introduce very limited data errors, which have little effect on the inference result of GRNs.

C. The effect of data size on parameter optimization estimation

After applying B-Spline to interpolate data, it is necessary to analyze the parameter estimation under different interpolation sizes. The general process is shown by Algorithm 3.

The effect of the total data size is analyzed using the same synthetic model. The data size is the total data size after applying the B-Spline algorithm with different sampling sizes. The stochastic CoD is analyzed with different intervals from two adjacent control points. The result is shown by Fig.4. It clearly shows that the CoD is larger with larger sampling size. However, both interpolation algorithms are upper bounded. Fig. 4 shows that B-Spline can get much better results in comparison to linear interpolation.
Algorithm 3 The general process to analyze the effect of data size on parameter optimization

Require:
The observed data $x_i$

Ensure: Stochastical errors of the estimated parameters
  1) Choose the sampling size $N$ for B-Spline algorithm
  2) Apply B-Spline algorithm with the size $N$ sampling points
  3) Apply Kalman Filter to calculate the parameters
  4) Calculate the stochastical errors of the estimated parameters

![Fig. 4. The effect of interpolation data size on the parameter estimator.](image)

V. YEAST DATA SIMULATION RESULTS

During this part of the simulation, we consider time-series gene-expression data corresponding to yeast protein synthesis. Five genes (HAP1($x_1$), CYB2($x_2$), CYC7($x_3$), CYT1($x_4$), COX5A($x_5$)) are selected because the relations among them have been revealed by biological experiments [16]. After the B-Spline algorithm is applied to the experimental data, the trace of the time-series microarray measurement data from [17] is shown by Fig. 5. Because the processes using B-splines algorithm generates enough data from the original experiment data, inference results for each iteration are stable. Therefore, unlike [3], the complicated processes using Z-scores are not necessary to process the simulation results, and B-splines is useful for the experiment data processing which improves the inference with less complicated process and more stable results.

The relationships among the 5 genes are shown by the branch pathway model given in Fig. 6. We observe that HAP1 represses CYC7, and CYB2 activates CYC7. It is also observed that HAP1 activates COX5A and CYT1. These observations are in agreement with the biological experiment findings in [16], [18].

VI. CONCLUSIONS

In this paper, the effects of the B-Spline algorithm on the gene time-series data and inference process are analyzed.

![Fig. 5. The dynamics of expression level of the 5 yeast genes.](image)

![Fig. 6. The branch pathway model of the 5 genes in yeast.](image)

Both synthetic and yeast microarray data and models are analyzed. The introduction of B-Spline algorithm yields better results in both cases. In further studies, B-Spline algorithm will be applied to analyze the effects of noisy data on genetic programming during the inference process.

REFERENCES


