Chemometrics and Intelligent Laboratory Systems 138 (2014) 57–63

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

An extended fractional Kalman filter for inferring gene regulatory networks using time-series data

Yongqing Zhang [a,*], Yifei Pu [a], Haisen Zhang [b], Yong Cong [c], Jiliu Zhou [a]

[a] College of Computer Science, Sichuan University, Chengdu 610065, PR China
[b] Department of Mathematics, Sichuan University, Chengdu 610064, PR China
[c] College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou 730000, PR China

ARTICLE INFO

Article history:
Received 9 October 2013
Received in revised form 17 June 2014
Accepted 9 July 2014
Available online 18 July 2014

Keywords:
Gene regulatory network
Extended fractional Kalman filter
Parameter estimation
Time-series data

ABSTRACT

In recent years, inference of gene regulatory networks has received ever increasing attention in the systems biology field. In this paper, for the first time, a fractional gene regulatory algorithm by extended fractional Kalman filter (EFKF) is proposed to estimate the hidden states as well as the unknown static parameters of the model, which can provide insight into the underlying regulatory relations among genes in the biological system. In the proposed method, gene regulatory networks are inferred via evolutionary modeling based on time-series microarray measurements. The gene regulatory network is considered as a fractional order discrete stochastic dynamic model that consists of the gene measurement equation and the gene regulation equation. After specifying the model structure, we apply the EFKF algorithm for identifying both the model parameters and the actual value of gene expression levels. In this paper, the main advantages of using fractional order systems, increasing the flexibility and improving the accuracy of the system state equation in EFKF are highlighted. The performance of the EFKF algorithm is compared with EKF and other nonlinear algorithms in predicting the parameters of gene regulatory networks from synthetic data and real biological data. Extensive computer simulations illustrate that the proposed algorithm outperforms EKF and other methods, and therefore, it can serve as a natural framework for inference gene regulatory networks with a nonlinear structure.

1. Introduction

Fractional calculus, which is first mentioned by Leibniz and L'Hospital in 1695, has been studied by many mathematicians for a long time. For more details, refer to the books by Oldham and Spanier [1]. During the last 20 years, fractional calculus has started to enter in more and more application fields, including physics, chemistry, materials science, viscoelasticity, electrical circuits, engineering and biology [2–6]. Fractional order models are now being applied to a wide range of problems in bioengineering [2].

Gene expression is the process of generating functional gene products, for example, mRNA and protein. Hence the level of gene functionality can be measured using microarrays or gene chips to produce gene expression data [7]. Measuring the levels of gene expression in different conditions is meaningful in medical diagnosis, treatment, and drug design [8]. Many gene expression experiments produce time-series data with only a few time points owing to the high measurement costs. Therefore, it becomes significant to predict the behavior of gene regulatory networks (GRNs) through modern computing technology. Recently, many algorithms and mathematical models were proposed to predict gene regulatory networks from time-series data [9,10], such as Boolean network [11], Dynamic Bayesian networks [12], neural networks [13], differential equations [14], state-space model [15], and stochastic model [16].

It is well known that gene expression is a complex nonlinear dynamic system. Parameter estimation in nonlinear dynamic systems is extremely important, but also extremely difficult. Some researchers use the S-system model to perform analysis of genetic network [17,18]. It has been successfully used in some biochemical networks, but encounters high-dimensionality problems when used to analyze large-scale genetic networks.

A remarkable feature of time-series gene express data is that the number of the time points is usually much smaller than that of the number of genes. So one of the most significant challenges in GRNs is to build a model that can analyze such a high-dimensional and short-length time-series data. In general, gene expression systems are partially observed. Therefore, a natural way to infer dynamic gene regulatory networks is to employ nonlinear state-space models that consist of two types of equations: system equations and observation equations [19]. The well-known extended Kalman filter (EKF) has been widely used in the state estimation of nonlinear dynamic systems from noisy measurements. Wang [20]has applied EKF to model nonlinear dynamic GRNs via short gene expression time series. The GRNs are considered

* Corresponding author.
E-mail address: zhangyongqingscu@hotmail.com (Y. Zhang).

http://dx.doi.org/10.1016/j.chemolab.2014.07.007
0169-7439/

as a nonlinear dynamic stochastic model that consists of the gene measurement equation and the gene regulation equation. Sun [21]has proposed EKF for estimation of parameters in nonlinear state-space models of biochemical networks. He discussed in detail how to develop a general framework for modeling biochemical networks and how to estimate the parameters in the models. Qian [22]has presented inference of noisy nonlinear differential equation models for GRNs using genetic programming and Kalman filtering. These research works demonstrate that EKF cannot only develop mathematical models but also estimate their parameters in gene regulatory networks. However, the differential order of the system equation in EKF is fixed as an integer which limits its applications and EKF only relies on the current state value. So we proposed a fractional gene regulatory algorithm by fractional calculus. The fractional calculus is a generalization of the traditional differential calculus for a case when integrals and derivatives are in not only integers but also fractional order. It is well known that fractional differential equations are useful because of their nonlocal character [23], i.e., the next state of a system not only depends on its current state but also on its historical states starting from the initial time. This is closer to reality and is therefore the main reason that fractional differential equations have become more popular.

This paper assumes that the GRNs obey a nonlinear fractional differential equation with additive Gaussian white noise. The gene expressions are assumed to evolve following a sigmoid squash function, whereas a linear function is considered for the microarray data. After specifying the model structure, we apply the extended fractional Kalman filter (EFKF) [24] to estimate the model parameters and hidden states of the nonlinear model using time-series data. A synthetic data and two real microarray time-series data from the yeast protein synthesis and SOS DNA Repair network of *Escherichia coli* are used to test our method. Result show that this method is capable of improving the prediction accuracy of microarray time-series dataset. Our major contributions in this study can be summarized as follows. (1) An EFKF is firstly presented to estimate the hidden states and parameters of the nonlinear model in GRNs. (2) The performance of our algorithm is evaluated for time-series data contrasting with the EKF and other nonlinear algorithms. It is demonstrated that our method is more effective than previous methods.

This paper is organized as follows: Section 2 illustrates the fractional gene regulatory algorithm by extended fractional Kalman filter and how to analyze the parameters for EFKF. Simulation results are given in Section 3. The performance of EFKF method is compared with EKF and other nonlinear algorithms. Section 4 contains some concluding remarks.

2. Materials and methods

2.1. System model and problem statement

Let $\mathbb{R}$ be the set of real numbers. The fractional order Grünwald–Letnikov difference of a function $x : \mathbb{R} \to \mathbb{R}$ is given by the following equation [24]:

$$!^{\gamma}x(k) = \frac{1}{\delta^{\gamma}}\sum_{j=0}^{k}(-1)^{j}\binom{\gamma}{j}x(k-j), \tag{1}$$

where $\gamma \in \mathbb{R}$ is a fractional order, and δ is a sampling time later equal to 1, k is the number of samples for which the derivative is calculated. The factor $\binom{\gamma}{j}$ can be obtained from:

$$\binom{\gamma}{j} = \begin{cases} 1, & j = 0 \\ \dfrac{\gamma(\gamma-1)\cdots(\gamma-j+1)}{j!}, & j>0 \end{cases}. \tag{2}$$

According to this definition, it is possible to obtain a discrete equivalent of the derivative (when γ is positive), a discrete equivalent of integration (when γ is negative).

Let n be the number of genes. We assume that the gene expression follows the fractional order discrete stochastic dynamical system:

$$\begin{cases} \Delta^{\gamma}x(k+1) = f(x(k)) + \omega(k), \\ x(k+1) = \Delta^{\gamma}x(k+1) - \sum_{j=1}^{k+1}(-1)^{j}\gamma(j)x(k+1-j), \quad k = 1,2,\cdots,m, \\ y(k) = h(x(k)) + \nu(k), \end{cases} \tag{3}$$

where

$$\gamma_k = diag\left[\binom{\alpha_1}{k}\cdots\binom{\alpha_n}{k}\right],$$
$$\Delta^{\gamma}x(k+1) = \begin{bmatrix} \Delta^{\alpha_1}x_1(k+1) \\ \cdots \\ \Delta^{\alpha_n}x_n(k+1) \end{bmatrix},$$

α_i is the order of system equation with respect to the i-th gene, $i = 1,2,\cdots, n$. m is the total number of data points, $\omega(k)$ and $\nu(k)$ are assumed to be zero-mean Gaussian white noise with covariance $Q(k)$and $R(k)$, respectively, i.e., $\omega(k) \sim N(0, Q(k))$, $\nu(k) \sim N(0, R(k))$. $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ and $h(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ are some proper nonlinear functions. The nonlinear function $f(\cdot)$ and $h(\cdot)$ can be linearized according to the Taylor series expansion.

Setting $f(x(k)) = [f_1(x_1(k)), f_2(x_2(k)), \cdots, f_n(x_n(k))]^T$. To capture the gene interactions effectively, we assumed [25]

$$f_i(x) = \sum_{j=1}^{n} a_{ij}g_j(x), \quad i = 1,2,\cdots,n, \tag{4}$$

where $\boldsymbol{A} = (a_{ij})_{n \times n}$ represents the nonlinear regulatory relationship among genes; and the nonlinear function $g_j(\cdot)$ is given by

$$g_j(x) = \frac{1}{1+\exp(-x)}, \quad x \in \mathbb{R}^n. \tag{5}$$

A discrete linear Gaussian model for the microarray data is considered which can be expressed at the ith time instant as [20]

$$y_i(k) = x_i(k) + \nu_i(k), \quad i = 1,2,\cdots,n, \quad k = 1,2,\cdots,m, \tag{6}$$

where $y(k) = [y_1(k), y_2(k), \cdots, y_n(k)]^T$. That is, $h \equiv I$, which is the identity matrix.

In our model, the $\theta = [a_{11}, a_{21}, \cdots, a_{n1}, a_{12}, a_{22}, \cdots, a_{n2}, a_{1n}, a_{2n}, \cdots, a_{nn}]$ are parameters to be identified. It is also worth pointing out that we can identify the n state variables as well.

2.2. Extended fractional Kalman filter

The extended fractional Kalman filter (EFKF), which has been studied in [24], is an important and fascinating algorithm in nonlinear theory which is extended to fractional systems. We list this algorithm in the following:

Theorem 2.1. [24]*For the nonlinear fractional order stochastic discrete state–space system given by Eq.(3), the extended fractional Kalman filter (EFKF) is given by the following equations:*

$$\Delta^{\gamma}\tilde{x}(k+1) = f(\hat{x}(k)), \tag{7}$$

$$\tilde{x}(k+1) = \Delta^{\gamma}\tilde{x}(k+1) - \sum_{j=0}^{k+1}(-1)^{j}\gamma_j\hat{x}(k+1-j), \tag{8}$$