Computational Statistics and Data Analysis 94 (2016) 63–74

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A likelihood-free filtering method via approximate Bayesian computation in evaluating biological simulation models

Takanori Hasegawa [a,*], Atsushi Niida [b], Tomoya Mori [a], Teppei Shimamura [c], Rui Yamaguchi [b], Satoru Miyano [b], Tatsuya Akutsu [a], Seiya Imoto [b,*]

[a] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, Japan
[b] Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, Japan
[c] Division of Systems Biology, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya, Aichi, Japan

ARTICLE INFO

Article history:
Received 18 November 2014
Received in revised form 23 June 2015
Accepted 4 August 2015
Available online 12 August 2015

Keywords:
Approximate Bayesian computation
Nonlinear state space model
Biological simulation
Gene expression

ABSTRACT

For the evaluation of the dynamic behavior of biological processes, e.g., gene regulatory sequences, we typically utilize nonlinear differential equations within a state space model in the context of genomic data assimilation. For the estimation of the parameter values for such systems, the particle filter can be a strong approach in terms of obtaining their theoretically exact posterior distributions of the parameter values. However, it has some drawbacks for dealing with biological processes in practice: (i) the number of unique particles decreases rapidly since the dimension of the parameter vector and the number of observed time points are higher than its capability, (ii) it cannot be applied when the likelihood function is analytically intractable, and (iii) the prior distributions of the parameter values are often arbitrary determined. To address these problems, we propose a novel method that utilizes the approximate Bayesian computation in filtering the data and self-organizing ensemble Kalman filter in constructing the prior distributions of the parameter values. Simulation studies show that the proposed method can overcome these problems; thus, it can estimate the posterior distributions of the parameter values with automatically setting prior distributions even for the cases that the particle filter cannot perform good results. Finally, we apply the method to real observation data in rat circadian oscillation and demonstrate the usefulness in practical situations.

1. Introduction

For the simulation of the dynamic behavior of biological processes, e.g., gene regulatory systems and chemical reactions, nonlinear differential equations have been typically utilized in the field of systems biology (Elowitz and Leibler, 2000; de Jong, 2002; Quach et al., 2007; Opper and Sanguinetti, 2010; Hasegawa et al., 2011). Thus, the changes of the concentration of biomolecules are represented as differential equations and their parameter values are determined to be consistent with observation data from biological experiments. Especially, in the context of genomic data assimilation using nonlinear state space model (SSM), the particle filter has been applied to estimate the parameter values of such nonlinear systems (Kitagawa, 1996; Nagasaki et al., 2006; Hasegawa et al., 2014).

* Corresponding authors.
E-mail addresses: t-hasegw@kuicr.kyoto-u.ac.jp (T. Hasegawa), aniida@ims.u-tokyo.ac.jp (A. Niida), tmori@kuicr.kyoto-u.ac.jp (T. Mori), shimamura@med.nagoya-u.ac.jp (T. Shimamura), ruiy@ims.u-tokyo.ac.jp (R. Yamaguchi), miyano@ims.u-tokyo.ac.jp (S. Miyano), takutsu@kuicr.kyoto-u.ac.jp (T. Akutsu), imoto@ims.u-tokyo.ac.jp (S. Imoto).

http://dx.doi.org/10.1016/j.csda.2015.08.003
0167-9473/

While the particle filter can estimate theoretically exact posterior distributions of the parameter values, there exist some drawbacks in evaluating biological simulation models: (i) the number of unique particles decreases rapidly since the dimension of the parameter vector and the number of observed time points are higher than its capability, (ii) it is not applicable when the likelihood function is analytically intractable, and (iii) the prior distributions of the parameter values are arbitrary determined since we have less information about them in many situations. These problems make it unable to estimate appropriate posterior distributions of the parameter values in many practical situations.

To overcome these problems, we propose a novel method utilizing the approximate Bayesian computation (ABC) in estimating the conditional distributions of the hidden state variables and self-organizing ensemble Kalman filter (self-EnKF) in constructing the prior distributions of the parameter values. Let $\boldsymbol{y}$ and $\boldsymbol{\theta}$ be observation data and a particular parameter vector. The original ABC was developed to estimate the posterior distribution of the parameter vector $p(\boldsymbol{\theta}|\boldsymbol{y})$ when the likelihood function $p(\boldsymbol{y}|\boldsymbol{\theta})$ is analytically intractable or difficult to calculate (Pritchard et al., 1999; Beaumont et al., 2002) and several extensions have been proposed utilizing either one of Markov chain Monte Carlo and sequential Monte Carlo approaches (Marjoram et al., 2003; Del Moral et al., 2006, 2012). Recently, ABC has been further extended to biological time-series data using differential equations (Toni et al., 2009; Toni and Stumpf, 2010) and nonlinear SSM in estimating the conditional distributions of the hidden state variables, known as the filtering step, as a framework of the ABC filtering (Jasra et al., 2012). Through the framework of the ABC filtering, the problems (i) and (ii) can be partially overcome; however, the method still cannot be directly applied to the evaluation of biological simulation models since (a) the dimension of the parameter vector and the number of observed time points are higher than the case that the previous ABC filtering approach can deal with and (b) the previous approach cannot handle replicated observations which are common in time-course biological data. To fully overcome the problems (i)–(iii), we develop a novel ABC filtering method with three extensions as contributions of this papers; thus, applying ABC filtering within Gibbs sampling procedure to obtain the posterior distribution of the parameter vector, suggesting an adaptive way to choose ϵ in dealing with replicated observations, developing self-EnKF for the construction of the prior distributions.

Through three simulation studies using two synthetic data based on differential equations describing a circadian oscillation with Gaussian and non-Gaussian observation noises, we show the effectiveness of the proposed method compared to the particle filter and the previous ABC filtering. At first, we demonstrate the usefulness of the ABC filtering to estimate the posterior distribution of a two dimensional parameter vector when the likelihood is intractable, thus, the particle filter cannot perform good results. Second, we show the applicability of the proposed method within the Gibbs sampling procedure for the problem of estimating a high dimensional parameter vector compared to the particle filter and the previous ABC filtering. Third, the validity of self-EnKF is shown for the automatic construction of the prior distributions of the parameter values. Finally, as an application example, we apply the method to a rat circadian oscillation pathway and estimate the posterior distributions of the parameter values, and then represent the expectation values of the hidden state variables for each time point.

2. Genomic data assimilation

2.1. A description of biological simulation models

To represent the dynamic behavior of biological processes, we apply a stochastic differential equation as a biological simulation model. Let $\boldsymbol{x}(\tau)$ be a p-dimensional vector including the concentration of biomolecules, e.g., mRNA and metabolite, as a function of time τ $(0 \leq \tau \leq T)$. Then, the time evolution of $\boldsymbol{x}(\tau)$ is described as

$$\frac{d\boldsymbol{x}(\tau)}{d\tau} = f(\boldsymbol{x}(\tau), \boldsymbol{\theta}) + \boldsymbol{v}, \tag{1}$$

where f is a function representing bimolecular reactions, $\boldsymbol{\theta}$ is a ν-dimensional parameter vector including, e.g., a synthesis rate of mRNA, and $\boldsymbol{v}$ is a system noise according to a Gaussian probability density $p(\boldsymbol{v})$. Let $\boldsymbol{y}_t \in R^q$ $(t \in \mathcal{T})$ be observation data of the concentration of biomolecules at time t where $\mathcal{T}$ is the entire set of observation time points. Here, we postulate $\mathcal{T} = \{1, \ldots, T\}$ for brevity. Since observation data is measured with observational noise $\boldsymbol{w}_t \sim p(\boldsymbol{w}_t)$, we have mathematical equations representing biological systems as

$$\boldsymbol{x}_t = \int_{t-1}^{t} f(\boldsymbol{x}(\tau), \boldsymbol{\theta}) d\tau + \boldsymbol{x}_{t-1} + \boldsymbol{v}_t, \tag{2}$$

$$\boldsymbol{y}_t = H\boldsymbol{x}_t + \boldsymbol{w}_t, \tag{3}$$

where $\boldsymbol{x}_t \in R^p$ is a hidden state vector representing the concentration of biomolecules at time t, H is a $q \times p$ observation matrix in which each row vector has only one active element and $\boldsymbol{v}_t \sim N(0, Q)$ is a system noise with a diagonal matrix Q. Eqs. (2) and (3) are called the nonlinear state space model that can assimilate observation data to a simulation model. Generally, the purpose of assimilating the data is to estimate the posterior distribution of the parameter vector $p(\boldsymbol{\theta}|Y_T = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\})$ and the posterior probability of a simulation model $\mathcal{M}$ as $p(\mathcal{M}|Y_T)$ represented by

$$p(\mathcal{M}|Y_T) = \int p(\boldsymbol{\theta}|Y_T) d\boldsymbol{\theta}. \tag{4}$$