ANALYSIS AND IMPLEMENTATION OF A MLE-BASED STOCHASTIC ALGORITHM FOR PARAMETER ESTIMATION

Irina BADRALEXI[1]

For as long as mathematical modelling has existed, parameter estimation has been a fundamental problem. In this article, we make use of an existing algorithm for parameter estimation in discreetly observed stochastic systems and apply it on different known biological models. Our main purpose is to investigate the potential of the considered algorithm by comparing the results we obtain to the results from literature. We also discuss the computational cost and possible optimizations that can be included.

Keywords: parameter estimation, stochastic algorithm, computational cost.

1. Introduction

The sole purpose of a mathematical model is to describe, as authentic as possible, a phenomenon. During the modelling phase, selecting which parameters should be included in the model influences its degree of accuracy and robustness. In most cases, though, the numerical values of the parameters are not known. Thus, estimating these values, so that the model behaves alike to the studied phenomenon, is a crucial step.

In this paper we focus on a method for parameter inference in discretely observed complex stochastic systems that can be modelled as a continuous-time discrete-state Markov process. It is based on estimating the likelihood function and its gradient with respect to the parameters. The mathematical framework and full description of the method can be found in [3].

The author remarks on the broad nature of the method and states that is can be applied to any discretely observed continuous-time Markov process with an explicit functional form of the transition rates. Moreover, this approach is also suitable for parameter estimation in the case of partial observations.

In what follows, we will use the method from [3] to estimate the parameters of some biological models from literature. For this purpose, we are going to offer a brief presentation of the mathematical concepts in a biological context and a short sketch of the algorithm. For more details, see [3].

[1] Assistant Professor, University POLITEHNICA of Bucharest, Romania, e-mail: irina.badralexi@gmail.com

2. Parameter estimation for biological systems

Consider K species of molecules $S_1, S_2, ..., S_K$, each having a population of $x_1, x_2, ..., x_K$ molecules at a certain time t. In the system containing our species of molecules, only M reactions $R_1, R_2, ..., R_M$ can take place. This system can be modelled by a continuous-time discrete-state Markov process. The state vector for the reaction system at a time moment t is: $X(t) = (x_1, x_2, ..., x_K)$.

We assume that the reaction model is governed by a set of parameters $\Theta = \{\theta_1, \theta_2, ..., \theta_r\}$, with $r \geq M$ (such that each reaction depends on at least one parameter).

The author of [3] assures us that the hypotheses of the mathematical framework (which is more general in nature) hold, in particular, for any biological system that obeys the mass-action law. This is due to the fact that the likelihood function is linear with respects to the parameters. Moreover, in biochemical reaction systems, the functional form for the transition rates is known (as they are the propensity functions).

Given a time interval $[a,b]$, we will perform a discretization in N subintervals of length $\triangle t = \frac{b-a}{N}$, such that at most one reaction (transition) can take place in any of the N subintervals.

Let $L(\Theta; X(a), X(b))$ denote the likelihood function for the biochemical reaction system in the time interval $[a,b]$, with full observations at $a = t_0$, $X(a) = X(t_0)$ and at $b = t_{J+1}$, $X(b) = X(t_{J+1})$. Assuming that the three hypothesis from [3] are fulfilled and that J reactions have occurred in the time interval $[a,b]$, with the corresponding times of the reactions (R_j, t_j), $j = 1, 2, ..., J$, the likelihood function can be represented as:

$$L(\Theta; X(a), X(b)) = \sum_{X \in S_N(a,b)} P(X^0) \left(\exp\left[-\sum_{j=0}^{J} (t_{j+1} - t_j) \cdot a_0(X(t_j), \Theta) \right] \cdot \prod_{j=1}^{J} a_j(X(t_j), \Theta) \right)$$

and its gradient is:

$$\frac{\partial L(\Theta; X(a), X(b))}{\partial \theta_r} = E\left[\sum_{j=0}^{J} \left(-(t_{j+1} - t_j) \frac{\partial a_0(X(t_j), \Theta)}{\partial \theta_r} \right) + \sum_{j=1}^{J} \frac{1}{a_j(X(t_j), \Theta)} \cdot \frac{\partial a_j(X(t_j), \Theta)}{\partial \theta_r} \right]$$

where the state $X(t_j)$ is the state of the system immediately after reaction number j, $a_0(X(t_j), \Theta)$ is the sum of all propensity functions for a given system

state and $E[\cdot]$ is the average over the different paths from $X^0 = X(a)$ to $X^N = X(b)$.

3. The algorithm (SAPEL)

The author of [3] proposed an algorithm (SAPEL – "**S**tochastic **A**lgorithm for **P**arameter **E**stimation with **L**ikelihood function") which outlays the steps necessary for implementing the estimation method. The algorithm consists of two main steps:

1. *Sampling step*
2. *Parameter estimation step*

The procedure requires the user to start with initial values for the parameters, which we will denote by $\hat{\theta}^{(0)} = (\theta_1, \theta_2, ..., \theta_n)$, and with known (observed) number of molecules for each species at $t = a$ and $t = b$ (in the fully observed case).

With these values, a number of system's trajectories is generated between $t = a$ and $t = b$ (sampling step). The user will choose a number N of subintervals for $[a,b]$, such that at most one reaction takes place in each subinterval (the choosing of N requires the user to have extensive knowledge of the biological phenomenon considered). The generation of some possible trajectories translates into determining the number of molecules from each species at the intermediate time points t_i, $i = \overline{1,N}$, taking into account the number of molecules at time $t = a$ and the number of molecules at time $t = b$. This step can be implemented by using Gillespie's First Reaction Method (see [2]).

After the trajectories are determined, the likelihood function and its gradient are calculated. The parameter estimation step basically consists of maximizing the likelihood function. Thus, the values of the parameters are updated according to the maximization process. This step can be implemented using a Gradient Ascent Method, as follows:

1. Determine the ascent direction: $d_k = \nabla L(\theta^k)$;
2. Update the parameter vector: $\theta^{k+1} = \theta^k + \lambda \cdot d_k$ (the step size $\lambda > 0$ is a small fixed value);
3. Test the stopping criterion: $|\theta^{k+1} - \theta^k| \le \varepsilon$, with a suitable value $\varepsilon > 0$ (this value must depend on the parameters range values).
4. If the stopping criterion is not satisfied, go back to the sampling step with the updated parameters values θ^{k+1}.

An interesting remark is that this algorithm also works in the case of partially observed data. This basically means that the number of molecules of a certain species (or more species) is unknown at the time moments $t=a$ and $t=b$. The mathematical framework allows parameter estimation in this case by simply declaring the unobserved species as unknown parameters (to be estimated).

4. Applications of the algorithm in biological models

In order to test the algorithm, we will consider two of the models presented in [1] and one of the examples provided in [3].

Example 1. The first model is a simple reaction model (see Table 1).

Table 1

Reaction model with 3 molecular species

Reactions	Intensity rates a_μ	State change vector $v_\mu(A,B,C)$
$R_1: \ A \xrightarrow{\theta_1} B$	$a_1=\theta_1 N_A$	$v_1=(-1,1,0)$
$R_2: \ B \xrightarrow{\theta_2} C$	$a_2=\theta_2 N_B$	$v_2=(0,-1,1)$

We denoted by N_A and N_B the number of molecules from species A and B, respectively. Through reaction R_1, one individual from species A transforms into an individual from species B and through the reaction R_2, one individual from species B transforms into an individual from species A.

The true values of the parameters, as found in [1], are $\theta_{true}=(0.04, 0.11)$. We consider the initial number of individuals at time $t=a$ as $\left(N_A^a, N_B^a, N_C^a\right)=(7,8,0)$ and at time $t=b$ as $\left(N_A^b, N_B^b, N_C^b\right)=(4,0,11)$, with $[a,b]=[0,23]$. For $\varepsilon=0.001$ (from the stopping criterion), we get:

Table 2

Estimated parameter values depending on the initial conditions, $\theta_{true}=(0.04, 0.11)$

Initial parameter values	Estimated parameter values
$(1,1)$	$(0.122, 0.568)$
$(0.5, 0.8)$	$(0.0703, 0.3231)$
$(0.1, 0.5)$	$(0.0445, 0.198)$

Example 2. The second example we consider is a viral infection model. As stated in [1], the viral infection process goes through the following stages: adsorption to the host cell and entry, the uncoating of the genome, transcription and translation, genome replication, assembly and release of the virus progeny. For this paper, we will consider a simple version of the model which corresponds to the early stages of the infection (see Table 3).

Table 3

Viral infection model

Reactions	Intensity rates a_μ	State change vector $v_\mu(V,G,M,P)$
$R_1: \ V \xrightarrow{\theta_1} G$	$a_1 = \theta_1 N_V$	$v_1 = (-1,1,0,0)$
$R_2: \ G \xrightarrow{\theta_2} G + M$	$a_2 = \theta_2 N_G$	$v_2 = (0,0,1,0)$
$R_3: \ G \xrightarrow{\theta_3} 2G$	$a_3 = \theta_3 N_G$	$v_3 = (0,2,0,0)$
$R_4: \ M \xrightarrow{\theta_4} M + P$	$a_4 = \theta_4 N_M$	$v_4 = (0,0,0,1)$

The molecular species are involved in the model are V (inactivated viral genome), G (activated viral genome), M (mRNA) and P (red fluorescent protein). For more information regarding the interaction of these molecules, see [1] We denoted by N_V, N_G and N_M the number of molecules from the respective species.

The inactivated viral genome V activates into G through reaction R_1. The transcription reaction R_2 creates mRNA, M, from the activated genome G. Reaction R_3 is responsible for the replication of the activated viral genome G. The translation reaction R_4 creates red fluorescent protein P from the mRNA M.

The true values of the parameters, as found in [1], are $\theta_{true} = (0.15, 0.02, 0.05, 1)$. We consider the initial number of individuals at time $t = a$ as $\left(N_V^a, N_G^a, N_M^a, N_P^a\right) = (10,0,0,0)$. For the values at time $t = b$, we generate a single trajectory using the true parameter values, and use the result as input for the algorithm. We consider the time interval $[a,b] = [0,30]$. For $\varepsilon = 0.001$ (from the stopping criterion), we get the parameter estimations found in Table 4.

Table 4

Estimated parameter values depending on the initial conditions $\theta_{true} = (0.15, 0.02, 0.05, 1)$

Initial parameter values	Estimated parameter values
$(1,1,1,2)$	$(0.3776, 0.152, 0.09, 1.301)$
$(0.3, 0.1, 0.1, 1.5)$	$(0.1721, 0.0332, 0.0505, 1.02)$
$(0.1, 0.05, 0.05, 0.8)$	$(0.1499, 0.02814, 0.0498, 1.031)$

Example 3. The third model we include is a stochastic version of a reversible decay-dimerization with conversion. The system contains 4 reactions, involving one species decay, a reversible dimerization and a conversion reaction (see Table 5).

Table 5

Decay-dimerization model

Reactions	Intensity rates a_μ	State change vector $v_\mu(N_{S_1}, N_{S_2}, N_{S_3})$
$R_1: \ S_1 \xrightarrow{\theta_1} *$	$a_1 = \theta_1 N_{S_1}$	$v_1 = (-1, 0, 0)$
$R_2: \ S_1 + S_1 \xrightarrow{\theta_2} S_2$	$a_2 = \theta_2 \dfrac{N_{S_1}(N_{S_1} - 1)}{2}$	$v_2 = (-2, 1, 0)$
$R_3: \ S_2 \xrightarrow{\theta_3} S_1 + S_1$	$a_3 = \theta_3 N_{S_2}$	$v_3 = (2, -1, 0)$
$R_4: \ S_2 \xrightarrow{\theta_4} S_3$	$a_4 = \theta_4 N_{S_2}$	$v_4 = (0, -1, 1)$

We denoted by $N_{S_1}, N_{S_2}, N_{S_3}$ the molecular species counts and assume that the reaction system follow the mass-action law.

The true values of the parameters, as found in [1], are $\theta_{true} = (0.2, 0.04, 0.5)$. We consider the initial number of individuals at time $t = a$ as $\left(N_{S_1}^a, N_{S_2}^a, N_{S_3}^a\right) = (1000, 10, 10)$.

For the values at time $t = b$, we generate a single trajectory using the true parameter values, and use the result as input for the algorithm. We consider the time interval $[a, b] = [0, 0.01]$. For $\varepsilon = 0.001$ (from the stopping criterion), we get:

Table 6

Estimated parameter values depending on the initial conditions, $\theta_{true} = (0.2, 0.04, 0.5)$

Initial parameter values	Estimated parameter values
$(1,1,1)$	$(0.219, 0.0556, 0.8798)$
$(0.5, 0.3, 0.8)$	$(0.2023, 0.0445, 0.565)$

5. Computational issues

The high computational cost is a known problem in the implementation of stochastic algorithms. Remark that, for the algorithm considered in this paper, a number (usually a fairly large number) of system trajectories are generated after each modification of the parameter values after the second step (if the stopping criterion was not met). For some simulation settings, it is not uncommon to expect a result after hours of waiting.

An important issue with this algorithm is that the method depends greatly on the initial conditions provided by the user. Note that for some initial conditions, the algorithm may not converge or it may converge towards inaccurate values.

Another aspect that requires special attention is the stopping criteria value ε. This value needs to be adjusted relative to the parameter range of values. If ε is very small (too small), then the computational cost will increase, but if it is too big, then the estimations may be amiss.

The computational cost also depends on the number of subintervals N which divide the time interval $[a,b]$. The user needs to take into account the implications of choosing this number. If the number of subintervals is too small, then more than one reaction may take place in any subinterval (which contradicts the theoretical results); if the number of subintervals is too big, then the computational cost will increase.

Regarding the sampling step, as stated before, every execution of this step requires the generation of many system trajectories that are consistent with the observed data. If the observed data consist of many time points, simulating a trajectory that passes through all of the data will be extremely unlikely, even when using the true parameter values (this is common when working with stochastic models). In order to bypass this problem, we can consider checking if the distance between the observed number of molecules and the simulated number (at each time point) is less than a user-defined threshold δ. We can use a normalized L_1 distance d and test if $d \le \delta$, where d has the form:

$$d = \sum_{N} \frac{\left| x_{sim}(t_i) - x_{obs}(t_i) \right|}{1 + x_{sim}(t_i)}$$

Note that, by doing this, we can also account for the possible observation errors.

6. Conclusions

Working with the true values of a system's parameters is a requirement in mathematical modelling. Having prior knowledge of these values is impossible in most cases. Thus, the development and implementation of parameter estimation methods is of outmost importance.

Due to the probabilistic nature of biological processes, mathematical modelling in biology usually implies the use of stochastic models. The purpose of this paper was to take an existing method for parameter estimation, to describe the associated algorithm and discuss the computational issues that arise.

The use of this algorithm demands that the user possesses extensive information regarding the process which is modelled. For example, the initial parameter values (which are imputed by the user) influence the estimated values, so prior knowledge of the interaction between the species is preferable.

A known issue of stochastic algorithms is the high computational cost. Moreover, the computation cost may increas relative to some values which are critical to the convergence of the algorithm. Some of the variables which influence the computation cost (but which also increase the accuracy of the estimation), are the stopping value ε, the learning value λ and the number of subintervals N for the trajectories. It is recommended that the user test different values for these variables in order to obtain the best computational cost versus accuracy ratio.

An upside to applying this method is that it also works in the case of partially observed data. The solution found in [3] is to consider the unobserved species as extra parameters. The problem in doing this is that the vector of parameters grows in length and the optimization problem increases in difficulty

Overall, excluding the high computation cost, the algorithm is general in nature and adaptable to different simulation scenarios and the estimations are close to the true parameter values (with appropriate initial values).

R E F E R E N C E S

[1]. *Ankur Gupta*, Parameter Estimation in Deterministic and Stochastic Models of Biological Systems, PhD Thesis, University of Wisconsin-Madison, 2013

[2]. *D.T. Gillespie*, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys. 22, (1976), 403-434

[3]. *Raluca Purnichescu-Purtan*, Mathematical Models in Biology, PhD Thesis, University Politehnica of Bucharest, 2013