ESTIMATION OF STOCHASTIC RATE CONSTANTS AND TRACKING OF SPECIES IN BIOCHEMICAL NETWORKS WITH SECOND-ORDER REACTIONS

Petar M. Djurić and Mónica F. Bugallo

Department of Electrical and Computer Engineering
Stony Brook University, Stony Brook, NY 11794, USA

e-mail: djuric@ece.sunysb.edu, monica@ece.sunysb.edu

ABSTRACT

In a recent work we applied particle filtering to simple biochemical networks composed of first-order reactions with the objective of estimating unknowns in the studied system that include stochastic rate constants and species with time-evolving numbers of molecules. In this paper we extend that effort to biochemical networks which have second-order reactions. We model the unknown stochastic rate constants by Gamma distributions and the number of reactions in a given time interval as Poisson random variables. The observations are nonlinear functions of some of the species in the system, and they are distorted by noises with known distributions. With these assumptions, we develop a particle filter that tracks the number of molecules of all the species in the network with time and estimates the unknown stochastic rate constants. We demonstrate the method on a reaction of importance in studying Ras regulation.

1. INTRODUCTION

The computational approaches to studying biological networks are either based on deterministic or stochastic methods. The former represent the biochemical reactions described by the network with differential equations and where the unknowns are the concentrations of the various species in the network. These equations are often numerically solved. In the stochastic framework, the central role is played by the chemical master equation and the unknowns are the probability distributions of the species. Of the two approaches, the stochastic one is more general, and it can improve on the modeling of the biochemical network. This is particularly obvious when the numbers of molecules of some of the species in the network are small and/or when studying signal transduction and gene expression [1], [2], [3]. Also, there are phenomena that cannot be explained with the deterministic methods, and they include multistability of the observed system when it is driven by random dynamic switching between stationary states [4].

For a given network, one can have two types of computational problems, usually referred to as forward and inverse problems [5]. The forward problem amounts to simulating a network from some starting sizes of molecular species and sequentially propagating them in time. This presumes that all the reaction rates of the network are known. The inverse problem is to estimate the unknowns of the network based on experimental time series measurements of some of the molecular species. The unknowns can be the reaction rates in the network or the numbers of some of the molecular species that cannot be measured.

Recently, we addressed the inverse problem by applying particle filtering to a set of measurements of some of the species in biochemical networks with first-order reactions [6]. In this paper, we extend this work to networks that include second-order reactions. We show that particle filtering can readily be used for accurate estimation of not only the dynamic variables in the system, like the time varying number of molecules of each of the species, but also of the stochastic rate constants. Due to the model assumptions, we can implement improved particle filtering where we only generate particles for the species in the system, whereas the unknown stochastic rate constants are integrated out. We can write the posteriors of the constants as mixture Gamma densities.

The paper is organized as follows. First, in Section 2 we state the problem as one of studying a discrete state-space model. In Section 3, we show how we implement the particle filter and track the constant and time varying unknowns. In Section 4, we demonstrate the performance of the proposed approach by computer simulations of a network that is important in modeling the regulation of the Ras protein. Finally in Section 5, we present brief conclusions.

2. THE PROBLEM

We consider biochemical networks that include both first- and second-order reactions. For example, consider the network described by

$$X_A + X_B \xrightarrow{c_1} X_C + X_D \underset{c_2}{\overset{c_3}{\rightleftharpoons}} X_E \tag{1}$$

where we have three reactions, two of second-order, and one of first-order. More specifically, the first reaction is given by

$$X_A + X_B \xrightarrow{c_1} X_C \tag{2}$$

the second by

$$X_C + X_D \xrightarrow{c_2} X_E \tag{3}$$

This work has been supported by the National Science Foundation under Award CCF-0515246 and the Office of Naval Research under Award N00014-06-1-0012. The work has been carried out while the first author held the Chair of Excellence of Universidad Carlos III de Madrid-Banco de Santander in 2008-2009.

© EURASIP, 2009

and the third by

$$X_E \xrightarrow{c_3} X_C + X_D \tag{4}$$

Besides the types of molecules that take part in the reactions, the latter are also defined by the reaction constants c_i. The stochastic nature of the processes that take place is expressed by random variables that describe how many reactions occur in a unit time interval. In our work, the number of reactions (whether first- or second-order) are given by random variables with truncated Poisson distributions. For example, in the time interval between t and $t+\Delta t$, and where at t the number of molecules of the species X_A and X_B is $x_{A,t}$ and $x_{B,t}$, respectively, the number of reactions r_1 defined by (2) is modeled by

$$p(r_1) \quad = \quad \frac{\lambda_1^{r_1}}{C\, r_1!} e^{-\lambda_1}, \quad r_1 = 0, 1, \cdots, \min(x_{A,t}, x_{B,t}) \tag{5}$$

where C is a normalizing constant, and $\lambda_1 = c_1 x_{A,t} x_{B,t} \Delta t$. It is instructive to view the previous expression also as $p(x_{A,t+\Delta t} \,|\, c_1, x_{A,t}, x_{B,t})$ and where $r_1 = x_{A,t} - x_{A,t+\Delta t}$.

We assume that we have $n-$time varying unknowns in the system denoted by $\mathbf{x}_t$, where $\mathbf{x}_t = [x_{1,t}\; x_{2,t} \cdots x_{n,t}]^\top$, and that they represent the number of molecules of the various species in the network. We also assume that each species can undergo only one reaction in the time interval Δt. Another set of unknowns in the network is the vector of stochastic rate constants $\mathbf{c}$. Given $\mathbf{c}$, we have that

$$\mathbf{x}_t \sim p(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{c}) \tag{6}$$

i.e., the process that describes the reactions in time is Markovian. Note that at the beginning, we either know the initial number of molecules or we model them by the prior $p(\mathbf{x}_0)$. For the vector of stochastic rate constants $\mathbf{c}$, we specify its prior, i.e.,

$$\mathbf{c} \quad \sim \quad p(\mathbf{c}). \tag{7}$$

We will assume that the prior $p(\mathbf{c})$ is a product of individual priors, i.e.,

$$p(\mathbf{c}) = \prod_{k=1}^{K} p_k(c_k) \tag{8}$$

where K is the number of reactions in the system, and where each individual prior is a Gamma distribution with parameters (α_k, β_k). Finally, we need to model the observations $\mathbf{y}_t$ in terms of the unknown states $\mathbf{x}_t$, that is,

$$\mathbf{y}_t \sim p(\mathbf{y}_t \,|\, \mathbf{x}_t). \tag{9}$$

Thus, given the distributions $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{c})$, $p(\mathbf{x}_0)$, $p(\mathbf{c})$, $p(\mathbf{y}_t \,|\, \mathbf{x}_t)$ and the observations $\mathbf{y}_{1:T}$, we want to estimate all the dynamic states $\mathbf{x}_t$ and the stochastic rate constants $\mathbf{c}$.

3. THE SOLUTION

Our solution is based, as in [6], on the use of particle filtering [7]. Here we briefly summarize the main points of the approach. First, the particle filter approximates the posterior of the unknowns with a discrete random measure, $\chi_t = \{(\mathbf{x}_t^{(m)}, \mathbf{c}^{(m)}), w_t^{(m)}\}_{m=1}^{M}$, where $(\mathbf{x}_t^{(m)}, \mathbf{c}^{(m)})$ is the $m-$th particle and $w_t^{(m)}$ is the associated weight of the particle. The constants $\mathbf{c}$ require special care in particle filtering because there is no dynamics involved with them. Ideally, we would like to integrate them out if at all possible. We can write

$$p(\mathbf{x}_t \,|\, \mathbf{x}_{1:t-1}) \quad = \quad \int p(\mathbf{x}_t \,|\, \mathbf{c}, \mathbf{x}_{t-1}) p(\mathbf{c} \,|\, \mathbf{x}_{0:t-1}) d\mathbf{c}. \tag{10}$$

The integrand $p(\mathbf{x}_t \,|\, \mathbf{c}, \mathbf{x}_{t-1})$ is a product of individual distributions, i.e.,

$$p(\mathbf{x}_t \,|\, \mathbf{c}, \mathbf{x}_{t-1}) = \prod_{k=1}^{K} p_k(\mathbf{x}_t \,|\, \mathbf{c}, \mathbf{x}_{t-1}) \tag{11}$$

where K is the number of reactions in the system, and $p_k(\mathbf{x}_t \,|\, \mathbf{c}, \mathbf{x}_{t-1})$ is the probability distribution that describes the $k-$th reaction, and its actual arguments are the reactants of the $k-$th reaction. The forms of $p_k(\cdot)$ are truncated Poisson distributions as shown by (5). Similarly, the posterior $p(\mathbf{c} \,|\, \mathbf{x}_{0:t-1})$ is a product of K factors, one for each constant c_k and represented by a Gamma distribution. For as long as the number of molecules that are reactants does not drop very low, we can approximate the truncated Poissons as regular Poisson distributions, in which case the K integrals from (10) can be analytically solved. For example, if we let without loss of generality $\Delta t = 1$, we can write for the first reaction (2) the following:

$$\begin{aligned} p_1(x_{A,t}|x_{A,0:t-1}, x_{B,0:t-1}) &= \frac{(x_{A,t-1} x_{B,t-1})^{x_{A,t-1} - x_{A,t}}}{(x_{A,t-1} - x_{A,t})!} \\ &\times \frac{(\beta + x_{A,0}x_{B,0} + \cdots + x_{A,t-2}x_{B,t-2})^{\alpha + x_{A,0} - x_{A,t-1}}}{\Gamma(\alpha + x_{A,0} - x_{A,t-1})} \\ &\times \frac{\Gamma(\alpha + x_{A,0} - x_{A,t})}{(\beta + x_{A,0}x_{B,0} + \cdots + x_{A,t-1}x_{B,t-1})^{\alpha + x_{A,0} - x_{A,t}}} \end{aligned} \tag{12}$$

where we recognize the Poisson-Gamma distribution [8]. Thus, in principle we can generate the particles of $x_{A,t}$ without having to generate particles for $\mathbf{c}$. We can also readily obtain the posteriors of c_k. They are all Gamma distributions with parameters $\alpha_{k,t}$ and $\beta_{k,t}$. For example, for $p(c_1|\mathbf{x}_{0:t})$ we have

$$p(c_1|\mathbf{x}_{0:t}) = \frac{\beta_{1,t}^{\alpha_{1,t}}}{\Gamma(\alpha_{1,t})} c_1^{\alpha_{1,t-1}} e^{\beta_{1,t} c_1} \tag{13}$$

where

$$\alpha_{1,t} \quad = \quad \alpha_1 + x_{A,0} - x_{A,t} \tag{14}$$

$$\beta_{1,t} \quad = \quad \beta_1 + \sum_{j=1}^{t} x_{A,j-1} x_{B,j-1}. \tag{15}$$

The method is implemented using at each time instant t the following steps:

1. Generate particles of all the species using Poisson–Gamma distributions of type (12).
2. Compute the weights of the particles

$$w_t^{(m)} \propto w_{t-1}^{(m)} p(\mathbf{y}_t \,|\, \mathbf{x}_t^{(m)}). \tag{16}$$

3. Resample if necessary.

As pointed, the overall posterior of the rate constants is a mixture of Gamma distributions, i.e., for example, the posterior of c_1 is approximated by

$$p(c_1 \,|\, \mathbf{y}_{1:t}) \simeq \sum_{m=1}^{M} w_t^{(m)} Ga(\alpha_{1,t}^{(m)}, \beta_{1,t}^{(m)}). \tag{17}$$

If necessary we can draw particles from the above distributions, or we can obtain some point estimates (for example the MSE estimate) without resorting to Monte Carlo sampling.

4. COMPUTER SIMULATIONS

We demonstrate the performance of our method on the following biochemical network:

$$X_A + X_B \xrightarrow{c_1} X_C + X_D \xrightarrow{c_2} X_B + X_E \tag{18}$$

This is a reaction scheme that can be used for describing the mechanism by which the regulation of Ras protein[1] (activation/inactivation) takes place [9], i.e., $\text{Ras}-\text{GDP}+\text{GEF} \xrightarrow{c_1} \text{Ras}-\text{GDP}-\text{GEF}+\text{GTP} \xrightarrow{c_2} \text{Ras}-\text{GTP}+\text{GEF}+\text{GDP}$. Namely, in that case the species X_A is the inactive conformation Ras-GDP, where GDP is a nucleotide, X_B represents the exchange factor GEF which induces disassociation of GDP from Ras by first forming the intermediate complex Ras-GDP-GEF, denoted by X_C. The species X_D is the GTP nucleotide that substitutes the GDP in the Ras-GDP-GEF conformation and thereby releases the exchange factor GEF. Thus, X_E is the activated Ras in the form of $\text{Ras}-\text{GTP}$. Note that for simplicity in (18) we dropped the species GDP (which results as a product of the second reaction) because it does not affect our study in the simulations. Once activated, Ras initiates a number of pathways [10].

We simulated the reaction (18) with the following setup: $X_{A,0} = 1{,}000$, $X_{B,0} = 800$, $X_{C,0} = 200$, $X_{D,0} = 300$, and $X_{E,0} = 0$; $c_1 = 10^{-5}$, and $c_2 = 8 \times 10^{-5}$. The observations were modeled according to

$$y_{1,t} = \text{round}(X_{A,t}^{3/2}) + v_{A,t} \tag{19}$$

$$y_{2,t} = \text{round}(X_{E,t}^{3/2}) + v_{E,t} \tag{20}$$

where round$(\cdot)$ stands for rounding operation, $v_{A,t} \sim \mathcal{N}(0, \sigma_v^2)$, and $v_{E,t} \sim \mathcal{N}(0, \sigma_v^2)$, with $\sigma_v = 200$. For the priors of c_1 and c_2 we used Gamma distributions with parameters $(\alpha_1 = 10, \beta_1 = 10^5)$, and $(\alpha_2 = 90, \beta_2 = 10^5)$. Thus, there were two observations per each time instant, one that was a measurement related to $X_{A,t}$ and the other related to $X_{E,t}$.

[1]The Ras molecules are proteins attached to the cell membranes and are key components for controlling various cell processes including cytoskeletal integrity, proliferation, adhesion, apoptosis, and migration.

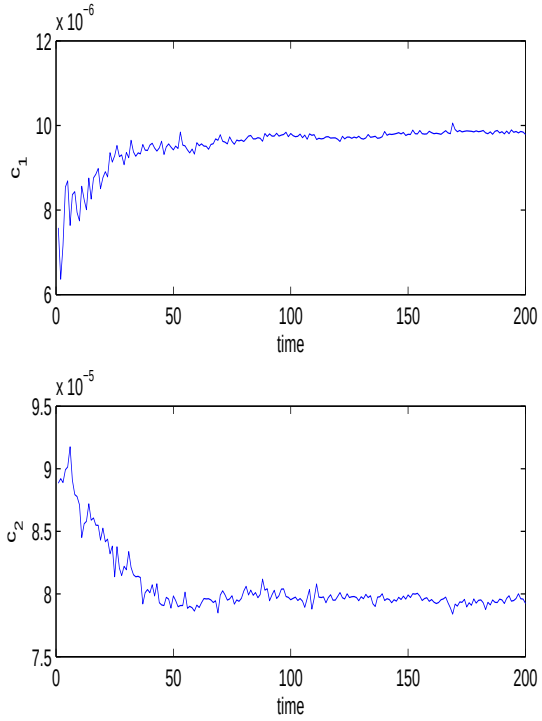

Figure 1: Estimates of the constants c_1 and c_2 as functions of time. The true values were $c_1 = 10^{-5}$, and $c_2 = 8 \times 10^{-5}$.

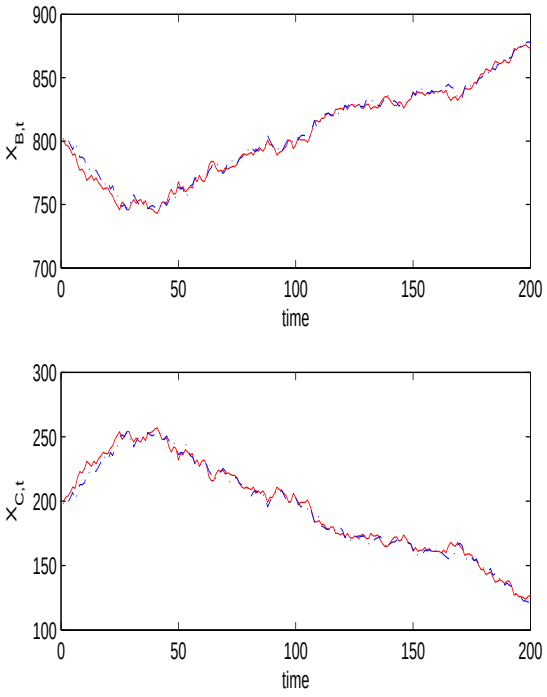

Figure 2: Estimates of the species $X_{B,t}$ and $X_{C,t}$ as functions of time. The solid (red) line represents the true values and the das-dot (blue) line the estimated values.

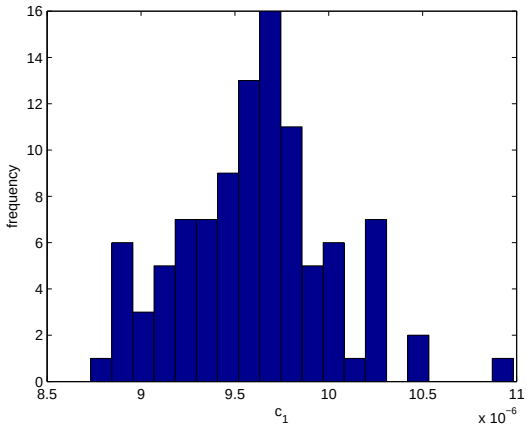

Figure 3: A histogram of estimated values of the constant c_1 after $T = 200$ observations. The true value of the constant was $c_1 = 10^{-5}$.

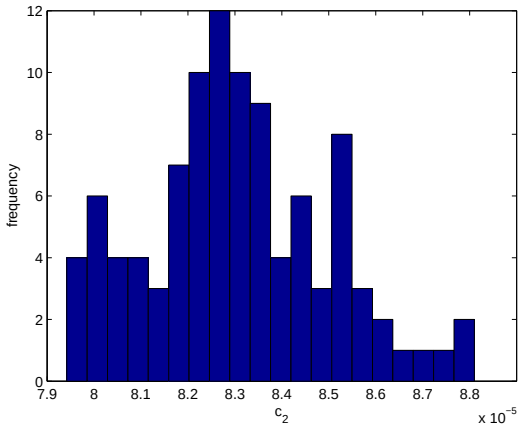

Figure 4: A histogram of estimated values of the constant c_2 after $T = 200$ observations. The true value of the constant was $c_2 = 8 \times 10^{-5}$.

Some of the results are shown in Figs. 1–4. In Fig. 1, we see the estimates of the constants c_1 and c_2 as functions of time for one realization of the biochemical process and the associated measurements. We can see that soon after the beginning, the values of the stochastic rate constants settled near their true values.

In Fig. 2, we observe the estimated evolution of the unobserved species $X_{B,t}$ and $X_{C,t}$ and how they compare with the true values of the species. Again, we can conclude that there is a good agreement between the true and the estimated values.

In Figs. 3 and 4, we plotted the histograms of the MSE estimates of the stochastic rate constants from 100 different realizations. From the histograms, we may conclude that these estimates are somewhat biased; on average c_1 was underestimated and c_2 was over estimated.

5. CONCLUSIONS

In this paper we presented a particle filtering method for tracking the unknown numbers of molecules in a biochemical network that may contain first- and second-order reactions. The priors of the unknown stochastic rate constants were modeled by Gamma distributions, and the distribution of the number of molecules of the species given their history, as Poisson-Gamma distributions. The performance of the particle filtering method was demonstrated on a biochemical network of interest in modeling Ras regulation.

Acknowledgment: The authors would like to acknowledge the group of Prof. Dafna Bar-Sagi at the Department of Biochemistry, New York University, for introducing them to the Ras regulation model, for guidance, and many fruitful discussions.

REFERENCES

[1] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda infected *escherichia coli* cells," *Genetics*, vol. 149, pp. 633–648, 1998.

[2] A. Deutsch, L. Brusch, H. Byrne, G. de Vries, and H. Herzel, Eds., *Mathematical Modeling of Biological Systems*, Birkhäuser, 2007.

[3] H. H. McAdams and A. Arkin, "It's a noisy business: Genetic regulation at the nanomolar scale," *Trends in Genetics*, vol. 15, pp. 65–69, 1999.

[4] M. Samoilov, S. Plyasunov, and A. P. Arkin, "Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations," *Proceedings of the National Academy of Sciences of the USA*, vol. 102, pp. 2310–2315, 2005.

[5]

[6] P. M. Djurić and M. F. Bugallo, "Particle filtering and the inverse problem of biochemical networks," in *the Proceedings of EUSIPCO*, Lausanne, Switzerland, 2008.

[7] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer, New York, 2001.

[8] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons, New York, 1994.

[9] C. Lenzen, R. H. Cool, H. Prinz, J. Kuhlmann, and A. Wittinghofer, "Kinetic analysis by fluorescence of the interaction between Ras and the catalytic domain of the guanine nucleotide exchange factor Cdc25^{mm}," *Biochemistry*, vol. 37, pp. 7420–7430, 1998.

[10] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky, and J. Darnell, *Molecular Cell Biology*, W. H. Freeman, San Francisco, CA, 2003.