A STOCHASTIC APPROACH TO SOLVING INVERSE PROBLEMS OF BIOCHEMICAL NETWORKS

Mónica F. Bugallo and Petar M. Djurić

Department of Electrical & Computer Engineering
Stony Brook University, Stony Brook, NY 11794-2350
email: {monica, djuric}@ece.sunysb.edu

ABSTRACT

Advances in the development of models that can satisfactorily describe biochemical networks are extremely valuable for understanding life processes. In order to get full description of such networks, one has to solve the *inverse* problem, that is, estimate unknowns (rates and populations of various species) or choose models from a set of hypothesized models using experimental data. In this paper we discuss signal processing techniques for resolving the inverse problem of biochemical networks using the stochastic approach based on Bayesian theory. The proposed methods are tested in simple scenarios and the results are promising and suggest application of these methods to more complex networks.

Index Terms— Inverse problem, biochemical networks, Bayesian theory, Monte Carlo methods, Cramér-Rao bounds.

1. INTRODUCTION

Biochemical networks describing intra- and inter-cellular processes (e.g. genetic networks, signal transduction networks, or metabolic pathways) can be studied using system theory [8]. These systems are represented by sets of coupled chemical reactions described by chemical species taking part in the reactions and the respective reaction rates [1]. Such models are used to understand the dynamics of the system, its behavior over time, its traffic patterns, why they emerge and how one can control them [8].

Quantitative studies of biochemical networks are typically carried out numerically by solving a set of coupled differential equations. These methods are known as deterministic and they have serious limitations. Under certain conditions, they represent the system under study inadequately and cannot be used for prediction of concentrations of biochemical species. This leads to instabilities particularly emphasized when some of the reactions in the biochemical network involve small number of molecules [9]. Stochastic methods, on the other hand, attempt to improve on the accuracy of the deterministic methods by employing Monte Carlo computations. One can use them to either simulate specific realizations of the studied processes (forward problem) or to obtain estimates of unknowns in the studied system or to choose models from a set of predefined models (inverse problem). In the recent past much work has been devoted to the forward problem and significant advances have been made [7]. The inverse problem has received much less attention.

In this paper we study the *inverse* problem of biochemical networks using the stochastic approach. This includes: (a) estimation of unknowns (usually number of some of the molecular species and reaction rates in the system) from limited experimental time series measurements, (b) derivation of posterior Cramér-Rao bounds (PCRBs) of the obtained estimates as functions of experimental design parameters, and (c) model selection. We propose to apply the Bayesian methodology in all of the tasks and use Monte Carlo-based methods for implementation.

The remainder of the paper is as follows. The next section describes the mathematical formulation of the system under the stochastic framework. In Section 3 the *inverse* problem is explained in detail as well as the proposed methodology. Section 4 provides with simple examples as proofs of concepts of the proposed methods and Section 5 finishes the paper with some concluding remarks.

2. PROBLEM STATEMENT

In general, we assume that the biochemical network is described by a set of species $X_1, X_2, \cdots, X_N$ which take part in reactions $R_1, R_2, \cdots, R_K$, with associated stochastic rate constants $c_1, c_2, \cdots, c_K$ [6, 14]. From a signal processing point of view, we represent the biochemical network as a discrete-time dynamic state-space model where the states of the system are composed of *number of molecules* of the various species in the network and *stochastic rate constants*[1]. The network is described by two probability distributions, the first representing the evolution of the network with time, and the second one, the measurements of some of the species in the network given the system states. We can formally express the state and observation equations of the network simply by

$$\text{state equation} \quad : \quad p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{1}$$

$$\text{observation equation} \quad : \quad p(\mathbf{y}_t|\boldsymbol{x}_t) \tag{2}$$

where $\boldsymbol{x}_t = [x_{1,t} \;\; x_{2,t} \;\; \cdots \;\; x_{N,t} \;\; c_1 \;\; \cdots \;\; c_K]^\top$ denotes the state vector at time instant $t\tau$, with $x_{n,t}$ being the number of molecules of species X_n, $t = 0, 1, 2, \cdots, T$ the discrete time index and τ the sampling time interval. Measurements of some features of the biochemical network, $\mathbf{y}_t$, are obtained[2] (for instance, by fluorescence spectroscopy [10]) for an interval of duration $T\tau$. Here we assume that we have uniform sampling of the data. However, the methods that we propose *do not* require uniform sampling.

The general objective is to estimate the unknown state vector $\boldsymbol{x}_{1:t}$ based on the discrete-time (and possibly) incomplete observations $\mathbf{y}_{1:t}$, where the notation, for example, $\boldsymbol{x}_{1:t}$ means

This work has been supported by the National Science Foundation under Awards CCF-0515246 and the Office of Naval Research under Award N00014-06-1-0012.

[1] Note that other parameters can also be included in the state of the system if necessary.

[2] In general, the obtained measurements will be described by models with nonlinearities [1].

$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t$. We note that $\boldsymbol{x}_0$ represents the initial value of the state vector, which is assumed known.[3]

3. THE *INVERSE* PROBLEM

The *inverse* problem of biochemical networks revolves around three interrelated topics. In the next subsections we explain in detail each of them as well as the methodology used for solving them.

3.1. Estimation of unknowns in the biochemical network

It is well known that the stochastic rate constants of biochemical reactions are often not accessible directly through experiments. Most of the population sizes of the species in the system and their variation with time are also not available. Therefore, the inverse problem of estimating them is quite challenging, specially in the context of stochastic models [11].

From point of view of Bayesian theory, all the information about the unknowns is quantitatively described by the posterior density of the unknowns. For example, at time instant $t\tau$, we have $p(\boldsymbol{x}_{1:t}|\mathbf{y}_{1:t}, \boldsymbol{x}_0)$, where $\mathbf{y}_{1:t}$ represents all the measurements up to $t\tau$. With this approach, we also use prior information which is quantified by the prior distributions of the unknowns. The posterior of the unknowns is formally written as

$$p(\boldsymbol{x}_{1:t}|\mathbf{y}_{1:t}, \boldsymbol{x}_0) \quad \propto \quad p(\mathbf{y}_{1:t}|\boldsymbol{x}_{0:t})p(\boldsymbol{x}_{1:t}|\boldsymbol{x}_0) \tag{3}$$

where $p(\mathbf{y}_{1:t}|\boldsymbol{x}_{0:t})$ is the likelihood function, and $p(\boldsymbol{x}_{1:t}|\boldsymbol{x}_0)$ is the prior distribution. The goal is to obtain the complete posterior distribution of $\boldsymbol{x}_{1:t}$.

From this posterior one can construct various types of point estimates, such as

$$\begin{aligned} \text{maximum a posteriori} \quad &: \quad \hat{\boldsymbol{x}}_t = \arg\max_{\boldsymbol{x}_t} p(\boldsymbol{x}_t|\mathbf{y}_{1:t}, \boldsymbol{x}_0) \\ \text{minimum mean} - \text{square} \quad &: \quad \hat{\boldsymbol{x}}_t = \int \boldsymbol{x}_t p(\boldsymbol{x}_t|\mathbf{y}_{1:t}, \boldsymbol{x}_0) d\boldsymbol{x}_t \end{aligned}$$

where $0 < t \leq T$. The posterior distribution also contains information about the uncertainty of our estimates, and there are various metrics that can be used to describe it. For example, from the posterior we can easily construct confidence intervals of the estimates.

3.2. Computation of the PCRBs

The PCRB serves as a benchmark and provides a metric of how far our estimator is from *optimal* performance. Also, the PCRB may yield information on how to choose parameters of an experiment so that the accuracy of the estimated unknowns is improved or on whether a parameter is identifiable from a given experiment design and the available measurements.

Suppose that the unknowns are the stochastic rate constants $\boldsymbol{c} = [c_1\ c_2,\ \cdots c_K]^\top$, and the observations are given by $\mathbf{y}_{1:T}$. Then we can write [13]

$$E\left(\hat{\boldsymbol{c}} - \boldsymbol{c}\right)\left(\hat{\boldsymbol{c}} - \boldsymbol{c}\right)^\top \quad \geq \quad \left(\mathbf{J}_D + \mathbf{J}_P\right)^{-1} \tag{4}$$

where $\hat{\boldsymbol{c}}$ is a function of all the observations $\mathbf{y}_{1:T}$, i.e., $\hat{\boldsymbol{c}}(\mathbf{y}_{1:T})$; $\mathbf{J}_D$ is the information matrix obtained from the *data* and $\mathbf{J}_P$ the information matrix obtained from the prior. The elements of these matrices are defined by [13]

$$\begin{aligned} \left[\mathbf{J}_D\right]_{ij} \quad &\triangleq \quad -\int \frac{\partial^2 \ln p(\mathbf{y}_{1:T}|\boldsymbol{c})}{\partial c_i \partial c_j} p(\mathbf{y}_{1:T}|\boldsymbol{c})\pi(\boldsymbol{c}) d\mathbf{y}_{1:T} d\boldsymbol{c} \\ \left[\mathbf{J}_P\right]_{ij} \quad &\triangleq \quad -\int \frac{\partial^2 \ln \pi(\boldsymbol{c})}{\partial c_i \partial c_j} \pi(\boldsymbol{c}) d\boldsymbol{c} \end{aligned} \tag{5}$$

where $\pi(\boldsymbol{c})$ is the prior of the stochastic rate constants $\boldsymbol{c}$. The diagonal elements of the inverse of $\mathbf{J}_D + \mathbf{J}_P$ represent the minimal mean-square estimates that can be achieved with *any* estimator, and the off-diagonal elements are the cross correlations.

In the case when the unknowns are dynamic and evolve with time, which is always the case when we estimate changing number of molecules of species, the number of unknowns increases with time. For deriving the PCRB in such cases, we use the theory presented in [12]. There, recursive equations for computing the PCRB of the states of a discrete-time nonlinear dynamic system are presented. Almost always, these bounds have to be computed numerically.

3.3. Model selection

The main problem in working with computational models of biochemical networks is uncertainty about the adopted model. How do we know that the model is correct? Or, when we have two or more competing models for the same phenomenon, which of the two models is better?

Let the different models be denoted by $\mathcal{M}_l$, $l = 1, 2, \cdots, L$. If all the observed data are given by $\mathbf{y}_{1:T}$ and all the models are a priori equally likely, according to the maximum a posterior (MAP) criterion of Bayes' theory, the best model is the one with the largest a posteriori probability $p(\mathcal{M}_l|\mathbf{y}_{1:T}, \boldsymbol{x}_0)$, i.e.,

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}_l} p(\mathcal{M}_l|\mathbf{y}_{1:T}, \boldsymbol{x}_0). \tag{6}$$

Since

$$p(\mathcal{M}_l|\mathbf{y}_{1:T}, \boldsymbol{x}_0) \quad = \quad \frac{p(\mathbf{y}_{1:T}|\boldsymbol{x}_0, \mathcal{M}_l)p(\mathcal{M}_l)}{p(\mathbf{y}_{1:T}|\boldsymbol{x}_0)} \tag{7}$$

and if the prior probabilities of the models are all equal, we deduce that for comparison we only need to use $p(\mathbf{y}_{1:T}|\boldsymbol{x}_0, \mathcal{M}_l)$, which is the *predictive* density of the observations given the model $\mathcal{M}_l$. In order to find this predictive density we need to marginalize all the state unknowns, or

$$p(\mathbf{y}_{1:T}|\boldsymbol{x}_0, \mathcal{M}_l) \quad \propto \quad \int p(\mathbf{y}_{1:T}|\boldsymbol{x}_{0:T}, \mathcal{M}_l)p(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0, \mathcal{M}_l) d\boldsymbol{x}_{1:T} \tag{8}$$

where $p(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0, \mathcal{M}_l)$ is the prior distribution of the states given model $\mathcal{M}_l$ and the initial value of the state vector $\boldsymbol{x}_0$. Obviously, analytical computation of (8) in any practically interesting situation is impossible, and therefore we will have to resort to computational methods based on Monte Carlo simulations.

It is worth pointing out that with this approach we alleviate the problem of overfitting the data. We only compare models based on how they predict *future* observed data, and not those that have already been used for estimation of the states of the model. Also, note that the MAP criterion asymptotically corresponds to the criterion known as minimum description length (MDL), which has been recently used for selection of models of biochemical pathways in [2].

[3] If the initial state vector is not known exactly, the a priori knowledge about it can be modeled by a probability distribution function.

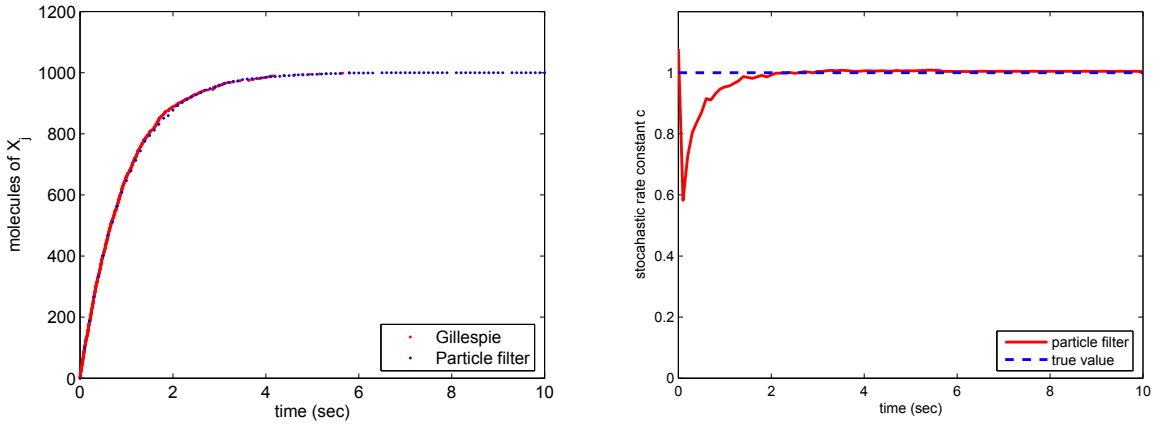

Fig. 1. Left: Tracking the evolution of species X_j. Right: Estimation of the stochastic rate constant c.

3.4. Methodology

The complexity of the models requires the use of Monte Carlo-based methods [4]. We use Bayesian computational methods and depending on the addressed problem, they will include Markov chain Monte Carlo (MCMC) simulation [5], population Monte Carlo (PMC) sampling [3], and particle filtering (PF) [4].

In the next section we show an example that addresses the three topics of the inverse problem and when necessary using particle filtering [4].

4. SIMULATION RESULTS

In this section, we provide basic examples related to the studied problem.

4.1. An example of estimation

We considered the reaction $X_i \xrightarrow{c} X_j$ for which we assumed that we had measurements of the number of molecules of species X_i taken with a sampling time interval τ. We let, as before, t be a discrete-time index, where $t = 0, 1, \cdots, T$, and we denoted the measurements of X_i by $x_{i,t}$. Based on the T measurements, we wanted to estimate the unknown rate constant c. For simplicity and derivation purposes, we assumed that the measurements are perfectly accurate.

The probability of a reaction of one molecule X_i, $\mathbb{P}(X_i \xrightarrow{c_k} X_j \text{ during } \tau)$, was given by

$$\mathbb{P}(X_i \xrightarrow{c} X_j \text{ during } \tau) \quad = \quad 1 - e^{-c\tau}. \tag{9}$$

When this probability was small, the likelihood of the number of molecules of species X_i converting to species X_j was modeled by a binomial distribution.

Given that the measurements of X_i were $x_{i,t}$, $t = 0, 1, \cdots, T$, and that we used a prior of c, $\pi(c)$, defined by the Gamma distribution with parameters α and β, that is,

$$\pi(c) \quad = \quad \frac{\beta^\alpha}{\Gamma(\alpha)} c^{\alpha-1} e^{-\beta c}, \quad c > 0 \tag{10}$$

we can obtain the MAP estimate of c readily by solving a nonlinear equation of c. This estimate is a batch estimate of the constant c and in general is not practical because we do not have access to accurate number of molecules, $x_{i,t}$, while the reaction takes place. A practical alternative is to have observations from which we can deduce the unknown number of molecules in the reaction. In other words, besides working with the state equation (1), we also have an observation equation as in (2).

One methodology for joint estimation of $x_{i,t}$ and c is PF. We carried out a simple simulation experiment to illustrate the performance of a particle filter that jointly tracks the evolution of the species and estimates the stochastic rate constant c. In the experiment, the species were observed with error. The transition of the state was modeled by $\mathbf{x}_{t+1} \sim p(\mathbf{x}_{t+1}|\mathbf{x}_t)$ where $p(\cdot)$ was a binomial distribution and the measurement of the number of molecules $x_{i,t}$ were modeled by

$$y_{t+1} = g(x_{i,t+1}) + v_{t+1}$$

where v_{t+1} was noise (or error), and $g(\cdot)$ was a function of the number of molecules (nonlinear measurements from fluorescence spectroscopy experiments were obtained [10]). The distribution of the noise was assumed known.

We considered a system whose initial number of molecules of X_i and X_j were set to 1000 and 0, respectively, and where the stochastic rate constant was $c = 1$. Figure 1 (left) shows the evolution of X_j in a single simulation run obtained by the exact method given by Gillespie [6] and the estimation obtained by the particle filter. It is apparent that the PF algorithm tracks the evolution of the species X_j very accurately and remains locked to the true value (the curve representing the true values and the curve depicting the estimates are almost indistinguishable). Figure 1 (right) depicts the MAP estimate of the stochastic constant rate by the particle filter.

4.2. An example of computation of a PCRB

Here we show the result of the derivation of the PCRB for the example from the previous subsection, where we had direct counts of the number of molecules of X_i uniformly in time and where the prior of c was the Gamma distribution defined by (10). We derived

$$E(\hat{c} - c)^2 \geq \frac{1}{\frac{\beta^2}{\alpha-2} + x_{i,0}\tau^2 \int_0^\infty f(c)\pi(c)dc} \tag{11}$$

where

$$f(c) = \frac{(1 - e^{-c\tau T})e^{-c\tau}}{(1 - e^{-c\tau})^2}.$$

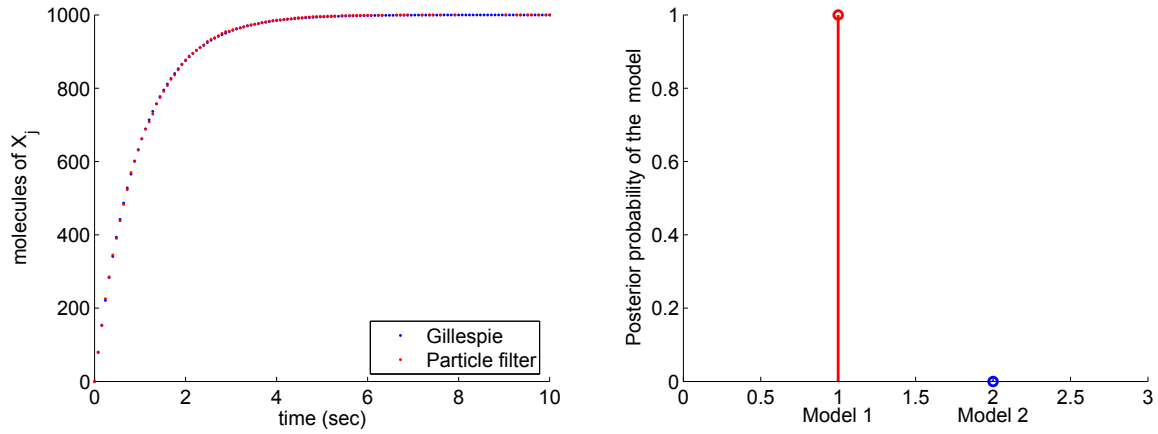

Fig. 2. Model selection problem – Model 1. Left: Tracking the evolution of species X_j. Right: Decision on model selection.

The PCRB for various values of α, β and $x_{i,0}$ can be obtained by numerical integrations.

4.3. An example of model selection

We also illustrate the performance of PF applied to model selection. We considered a scenario described by two possible models

$$\begin{array}{ll} \text{Model 1:} & X_i \xrightarrow{c_1} X_j \\ \text{Model 2:} & X_i \xrightarrow{c_2} X_j \\ & X_j \xrightarrow{c_3} X_i \end{array} .$$

The objective was to decide which model was the true one based on the observed data. To that end, we simulated a realization of the system according to Model 1 and we ran the PF on both models. The result is shown in Figure 2. It is evident that the particle filter using Model 1 tracks accurately the system of the state (left) and it provides the answer that Model 1 is the correct one with posterior probability of almost one.

We repeated the experiment but generating data from Model 2 and similar results as shown in Figure 2 were obtained. The posterior probability of the correct model was again almost one.

5. CONCLUSIONS

In this paper we present a stochastic approach to resolve the *inverse* problem posed in biochemical networks. We discuss methods for estimating unknowns, derivation of the posterior Cramér-Rao bounds of the obtained estimates, and methods for model selection. The applied methodology is Bayesian and its implementation is based on Monte Carlo techniques including particle filtering. Future lines of research include extending the methods for more complicated networks as well as application of the methods to real data.

6. REFERENCES

[1] J. M. Bower and H. Bolouri, Eds., *Computational Modeling of Genetic and Biochemical Networks*, MIT Press, 2004.

[2] R. Brause, "Model selection and adaptation for biochemical pathways," in *Biological and Medical Data Analysis*, pp. 439–449. Springer, 2004.

[3] O. Cappé, A. Guillin, and C. P. Robert, "Population Monte Carlo," *Journal of Computational and Graphical Statistics*, vol. 13, pp. 927–929, 2004.

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

[5] D. Gamerman, *Markov Chain Monte Carlo*, Chapman & Hall, 1997.

[6] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Physical Chemistry*, vol. 61, pp. 2340–2361, 1977.

[7] D. T. Gillespie and L.R. Petzold, "Improved leap-size selection for accelerated stochastic simulation," *The Journal of Chemical Physics*, vol. 119, pp. 8229–8234, 2003.

[8] H. Kitano, "Systems biology: A brief overview," *Science*, vol. 295, pp. 1662–1664, 2006.

[9] C. J. Morton-Firth and D. Bray, "Predicting temporal fluctuations in an intracellular signaling pathway," *Journal of Theoretical Biology*, vol. 192, pp. 117–128, 1998.

[10] A. Sharma and S. G. Schulman, *Introduction to Fluorescence Spectroscopy*, Wiley interscience, 1999.

[11] T. Tian, S. Xu, J. Gao, and K. Burrage, "Simulated maximum likelihood method for estimating kinetic rates in gene expression," *Bioinformatics*, vol. 23, pp. 84–91, 2007.

[12] P. Tichavský, C. H. Muravchik, and A. Nehorai, "Posterior Cramér-Rao bounds for discrete-time nonlinear filtering," *IEEE Transactions on Signal Processing*, vol. 46, pp. 1386–1396, 1998.

[13] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, John Wiley & Sons, 1968.

[14] D. J. Wilkinson, *Stochastic Modeling for System Biology*, Chapman & Hall/CRC, 2006.