Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Regularized Maximum Likelihood Estimation of Sparse Stochastic Monomolecular Biochemical Reaction Networks

Hong Jang*, Kwang-Ki K. Kim, Jay H. Lee* IFAC Fellow, Richard D. Braatz*** IFAC Fellow**

** Department of Biomolecular and Chemical Engineering, Korea Advanced Institute of Science and Technology,*
Daejeon, Korea (Tel: +82-42-350-3966; e-mail: jayhlee@kaist.ac.kr)
*** School of Electrical and Computer Engineering, Georgia Institute of Technology,*
Atlanta, GA 30332 USA
**** Massachusetts Institute of Technology, Cambridge, MA 02139 USA*

Abstract: A sparse parameter estimation method is proposed for identifying a stochastic monomolecular biochemical reaction network system. Identification of a reaction network can be achieved by estimating a sparse parameter matrix containing the reaction network structure and kinetics information. Stochastic dynamics of a biochemical reaction network system is usually modeled by a chemical master equation, which is composed of several ordinary differential equations describing the time evolution of probability distributions for all possible states. This paper considers closed monomolecular reaction systems for which an exact analytical solution of the corresponding chemical master equation is available. The estimation method presented in this paper incorporates the closed-form solution into a regularized maximum likelihood estimation (MLE) for which model complexity is penalized, whereas most of existing studies on sparse reaction network identification use deterministic models for regularized least-square estimation. A simulation result is provided to verify performance improvement of the presented regularized MLE over the least squares (LSE) based on a deterministic mass-average model in the case of a small population size. Improved reaction structure detection is achieved by adding a penalty term for ℓ_1 regularization to the exact maximum likelihood function.

Keywords: Sparse parameter estimation; Exact maximum likelihood estimation; Monomolecular biochemical reaction network; Chemical master equation; Stochastic simulation algorithm; Regularized maximum likelihood estimation

1. INTRODUCTION

Stochastic dynamics of bio- or nano-systems have lately received increased attention from researchers in the fields of biological and material engineering. In the past, such studies were greatly hampered by lack of measurements, but recent developments in sensing techniques that can provide real-time observations of stochastic dynamics at small length scales have motivated many scientific investigations with results published in prominent academic journals for the biological field (Raj et al., 2010; Taniguchi et al., 2010) and the nanotechnological field (Cognet et al., 2007; Jin et al., 2010).

One such sensing technique for bio-systems is bio-imaging using fluorescent proteins (Taniguchi et al., 2010). By grafting a fluorescent protein into the gene expression, the protein and mRNA expressions originating from targeted DNA can be detected quantitatively in real time. Specifically, membrane-localized yellow fluorescent protein (yfp) has been widely used for detecting changes with single-molecule sensitivity in individual live cells. A similarly effective development for nano-systems is near-infrared fluorescent carbon-nanotube (CNT) based nano-sensor arrays (Jin et al., 2010). CNT-based sensors can detect adsorption and desorption of a target molecule, such as hydrogen peroxide (H_2O_2), from changes in light emission. By monitoring step changes in the light intensity, the number of adsorbed target molecules on the surface of the sensor can be followed to single-molecule resolution. The monitoring of the adsorption behaviour in turn can provide information on concentrations of the target molecule in solution local to the CNT.

In the real-time data reported in the aforementioned papers, a strong stochastic behaviour is observed. For example, in the case of fluorescent protein expression, transcribed protein molecules bursts from the cell, controlled by an identical messenger RNA molecule, which have different copy numbers. In addition, the number of adsorbed molecules on the CNT-based nano-sensor under an exactly same experimental condition can exhibit significantly different time traces. In both cases, the total population number of molecules or species in the system within the detectable range is rather small, ranging from tens to hundreds of molecules. These two examples exhibit some common characteristics. First, the dynamics shows transient or non-equilibrium behaviour. Second, the system having a small-size population is best represented by a discrete state; however, the number of possible configurations can be quite large resulting in a large state space. Lastly, the experimental data are highly stochastic.

Stochastic dynamics of systems with discrete states can be modelled by the chemical master equation (CME) (Feinberg, 1979; Fichthorn and Weinberg, 1991),

$$\frac{\partial P(\boldsymbol{\sigma},t)}{\partial t}=\sum_{\sigma'}W(\boldsymbol{\sigma}',\boldsymbol{\sigma})P(\boldsymbol{\sigma}',t)-\sum_{\sigma'}W(\boldsymbol{\sigma},\boldsymbol{\sigma}')P(\boldsymbol{\sigma},t) \tag{1}$$

where $P(\boldsymbol{\sigma},t)$ is the probability of the system being in state $\boldsymbol{\sigma}$ at time *t*, and $W(\boldsymbol{\sigma}',\boldsymbol{\sigma})$ is the transition rate from state $\boldsymbol{\sigma}'$ to state $\boldsymbol{\sigma}$. The CME describes the time evolution of the probability distribution among all possible configurations. The CME (1) can be written as

$$\frac{d\boldsymbol{P}(t)}{dt}=\mathbf{A}(t;\boldsymbol{\beta})\boldsymbol{P}(t) \tag{2}$$

where $\boldsymbol{P}(t)$ is the state vector containing all the state probability variables and $\mathbf{A}(t;\boldsymbol{\beta})$ is a matrix containing all the transition rate constants, which have dependence on the model parameter vector $\boldsymbol{\beta}$.

Many numerical algorithms have been developed for solving the matrix ordinary differential equation (2), which can be divided into direct methods (MacNamara et al., 2008; Munsky and Khammash, 2006) and indirect methods (Gibson and Bruck, 2000; Gillespie, 1977). Direct methods attempt to evaluate the matrix exponential directly, such as the finite state projection (FSP) algorithm. In practice, given the large size of the state space, indirect methods that use stochastic simulation algorithms (SSA) to generate approximate probability distribution have been more popular.

Given experimental data, a parameter estimation method can be used to identify the parameters of a reaction network model, which in turn reveals its structure. Typically, the estimation is formulated to find parameter values minimizing the distance between the experimental data and their model predictions. Most of the literature has employed least-squares estimation (LSE) approaches, which fit stochastic data to a deterministic mass-average model (Golding et al., 2005). The LSE method, however, can provide poor parameter estimates for highly stochastic systems (Tian et al., 2007). Recently, many stochastic parameter estimation methods based on the stochastic differential equation of type (1) have been published (Munsky et al., 2012). These methods attempt to solve for the probability density functions (PDFs) of the CME and use them for estimation. Previous studies employed the moment-based method (Zechner et al., 2012; Munsky et al., 2009), the Bayesian method (Golightly and Wilkinson, 2011; Lillacci and Khammash, 2010; Boys et al., 2008), the maximum likelihood estimation (MLE) method (Neuert et al., 2013; Daigle et al., 2012; Tian et al., 2007), and the density function distance (DFD) method (Lillacci and Khammash, 2013; Poovathingal and Gunawan, 2010). However, these published methods are based on approximated PDFs rather than exact PDF solutions of the CME. In many cases, the PDFs are approximated using the SSA approach, which typically demands a very large number of simulations to be performed for an accurate estimation.

This paper considers closed monomolecular reaction systems, which enables the use of an exact analytical solution of the CME that is described by a multinomial distribution. The exact solution enables formulation of an exact MLE method. Improved performance over the LSE method is shown in a simulation study for an artificial reaction network system.

2. BIOCHEMICAL REACTION NETWORK SYSTEM

Some of the biochemical reaction network systems such as gene expression or metabolic pathway can be described by a combination of several monomolecular reactions. Possible monomolecular reactions can be categorized with conversion, inflow, and outflow reactions with a set of n different species denoted by S_i, $i=1,\dots,n$:

$$\begin{array}{ll} S_i \xrightarrow{k_{ij}} S_j & \text{Conversion } (i\neq j) \\ S_0 \xrightarrow{k_{0j}} S_j & \text{Inflow} \\ S_i \xrightarrow{k_{i0}} * & \text{Outflow} \end{array} \tag{3}$$

where S_0 is a pseudo-species outside the system and k_{ij} is nonnegative rate constant for the reaction from S_i to S_j for $i\neq j$ and can be time-varying. The monomolecular conversion reaction excludes catalytic or splitting reactions:

$$\begin{array}{ll} S_i \xrightarrow{k} S_i+S_j & \text{Catalytic } (i\neq j) \\ S_i \xrightarrow{k} S_j+S_r & \text{Splitting } (i\neq j\neq r) \end{array} \tag{4}$$

If the number of species in the system is sufficiently large, the dynamics of the system can be simply described by the deterministic ordinary differential equation

$$\frac{dC_i(t)}{dt}=k_{0j}(t)+\sum_{j\neq i}k_{ji}(t)C_j(t)-\sum_{j\neq i}k_{ij}(t)C_i(t) \tag{5}$$

where $C_i(t)$ is the population density or concentration of the species S_i and continuous variable. For a small number of species, the CME describes the stochastic dynamics of the system by (Gadgil et al, 2005; Jahnke and Huisinga, 2007)

$$\begin{aligned}\frac{\partial P(\boldsymbol{x},t)}{\partial t} &= \sum_{i=1}^{n}k_{0i}(t)\left(P(\boldsymbol{x}-\boldsymbol{e}_i,t)-P(\boldsymbol{x},t)\right)\\ &+\sum_{j=1}^{n}k_{j0}(t)\left((x_j+1)P(\boldsymbol{x}+\boldsymbol{e}_j,t)-x_jP(\boldsymbol{x},t)\right)\\ &+\sum_{j=1}^{n}\sum_{i=1}^{n}k_{ji}(t)\left((x_j+1)P(\boldsymbol{x}+\boldsymbol{e}_j-\boldsymbol{e}_i,t)-x_jP(\boldsymbol{x},t)\right)\end{aligned} \tag{6}$$

where $P(\boldsymbol{x},t)$ is a probability for the integer state vector $\boldsymbol{x}\in\mathbb{Z}^n$ with x_i as the population of the ith species and $\boldsymbol{e}_i\in\mathbb{R}^n$ denotes the standard basis vector (with 1 for the ith element and zero for the rest). In right-hand side of (6), the first, second, and third terms describe the influences of the inflow, outflow, and conversion reactions, respectively.

3. EXACT MAXIMUM LIKELIHOOD ESTIMATION

An exact solution of the CME (6) for the monomolecular reaction system can be obtained with mass-conserving or no-inflow assumptions. Proposition 1 of Jahnke and Huisinga (2007) indicates that the analytic PDF solution of the CME is defined by a multinomial distribution,

$$P(\boldsymbol{x},t)=\mathcal{M}\big(\boldsymbol{x},N,\boldsymbol{\lambda}(t)\big) \tag{7}$$

$$N=\sum_{i=1}^{n}x_i \tag{8}$$

where N is the total number of entities in the population in the system and $\boldsymbol{\lambda}(t)\in\mathbb{R}^n$ is probability parameter that is a vector of population fraction evolving according to the mass-average rate-reaction equation (5) which is defined by

$$\frac{d\boldsymbol{\lambda}(t)}{dt}=\boldsymbol{A}_k(t)\boldsymbol{\lambda}(0) \tag{9}$$

$$\boldsymbol{\lambda}(t)=\frac{\boldsymbol{C}(t)}{\sum_i C_i(t)} \tag{10}$$

$$A_{k_{ij}}(t)=k_{ji}(t),j\neq i \tag{11}$$

$$A_{k_{ii}}(t)=-\sum_{j\neq i}k_{ij}(t) \tag{12}$$

where $\boldsymbol{A}_k(t)\in\mathbb{R}^{n\times n}$ is the kinetic matrix, and the multinomial distribution, $\mathcal{M}\big(\boldsymbol{x},N,\boldsymbol{\lambda}(t)\big)$, is defined by

$$\begin{aligned}&\mathcal{M}\big(\boldsymbol{x},N,\boldsymbol{\lambda}(t)\big)\\&=\begin{cases}N!\dfrac{(1-\|\boldsymbol{\lambda}(t)\|_1)^{N-|x|}}{(N-\|\boldsymbol{x}\|_1)!}\displaystyle\prod_{i=1}^{n}\frac{\lambda_i^{x_i}(t)}{x_i!} & \text{if } \|\boldsymbol{x}\|_1\leq \mathrm{N}\\ 0 & \text{otherwise}\end{cases}\end{aligned} \tag{13}$$

A detailed proof is provided by Jahnke and Huisinga (2007). Simply, if the multinomial distribution (13) is substituted into both sides of the CME (6), both sides can be shown to be identical.

Another interpretation of Proposition 1 in Jahnke and Huisinga (2007) is that a multinomial distribution stays a multinomial distribution. That is, if the initial condition is defined by a multinomial distribution, then the probability distribution of the CME after several evolutions is still defined by a multinomial distribution. In practice, ordinary biochemical reaction systems rarely have a multinomial distribution as an initial condition, but can have arbitrary deterministic initial condition defined by the delta function,

$$P(\boldsymbol{x},t_0)=\delta_{\boldsymbol{\xi}}(x)=\begin{cases}1 & \text{if } \boldsymbol{x}=\boldsymbol{\xi}\\ 0 & \text{otherwise}\end{cases} \tag{14}$$

where $\delta_{\boldsymbol{\xi}}(x)$ is the Kronecker delta and $\boldsymbol{\xi}\in\mathbb{R}^n$ is a particular deterministic initial state. For this initial condition, Proposition 1 in Jahnke and Huisinga (2007) cannot be used for defining the exact solution of the CME at time t. However, when the multinomial distribution is defined with a probability parameter having initially full population fraction for the species S_i, the distribution has a delta function for the state,

$$\mathcal{M}(\boldsymbol{x},N,\boldsymbol{\lambda})=\delta_{N\boldsymbol{e}_i}(\boldsymbol{x}) \qquad \Leftrightarrow \qquad \boldsymbol{\lambda}=\boldsymbol{e}_i \tag{15}$$

With the assumption of the monomolecular reaction, the overall population can be divided into independent subsets for each species, $\boldsymbol{x}^{(i)}(t_k)\in\mathbb{R}^n$ for $i=1,\dots,n$, at time t_k,

$$\boldsymbol{x}(t_k)=\boldsymbol{x}^{(1)}(t_k)+\boldsymbol{x}^{(2)}(t_k)+\cdots+\boldsymbol{x}^{(n)}(t_k) \tag{16}$$

$$\begin{aligned}&\boldsymbol{x}^{(1)}(t_k)=\begin{bmatrix}x_1(t_k)\\0\\\vdots\\0\end{bmatrix},\ \boldsymbol{x}^{(2)}(t_k)=\begin{bmatrix}0\\x_2(t_k)\\\vdots\\0\end{bmatrix},\\&\boldsymbol{x}^{(n)}(t_k)=\begin{bmatrix}0\\0\\\vdots\\x_n(t_k)\end{bmatrix}\end{aligned} \tag{17}$$

The initial distribution of each subset, $\boldsymbol{x}^{(i)}(t_k)$, can be defined by the multinomial distribution based on (15) and can be evolved to the next time step t_{k+1} independently. Then the joint probability for all subset at time t_{k+1} is defined by convoluting every PDFs, which still remain as multinomial distributions. The exact solution, $\boldsymbol{P}(\cdot,t_{k+1})\in\mathbb{R}^{N^n}$ for all possible states (denoted by $\cdot$), of the CME for an arbitrary deterministic previous state, $\boldsymbol{x}(t_k)$, can be written as (Theorem 1 of Jahnke and Huisinga (2007)):

$$\begin{aligned}&\boldsymbol{P}(\cdot,t_{k+1})\\&=\mathcal{M}\left(\cdot,x_1(t_k),\boldsymbol{\lambda}^{(1)}(t_{k+1})\right)\\&*\mathcal{M}\left(\cdot,x_2(t_k),\boldsymbol{\lambda}^{(2)}(t_{k+1})\right)*\cdots\\&*\mathcal{M}\left(\cdot,x_n(t_k),\boldsymbol{\lambda}^{(n)}(t_{k+1})\right)\end{aligned} \tag{18}$$

where the asterisk, *, is the discrete convolution operator and the probability parameter, $\boldsymbol{\lambda}^{(j)}(t_{k+1})\in\mathbb{R}^n$ for the jth subset is defined by the Euler forward method applied to (9) with initially full state for the species S_i,

$$\boldsymbol{\lambda}^{(i)}(t_{k+1})=(\boldsymbol{I}+\delta t\boldsymbol{A}_k)\boldsymbol{e}_i \tag{19}$$

$$t_{k+1}=t_k+\delta t \tag{20}$$

where $\boldsymbol{I}\in\mathbb{R}^{n\times n}$ is identity matrix and δt is the size of the sampling time step.

The formulation of the parameter estimation method is based on the likelihood function, which is primarily defined by a conditional PDF for the reaction parameter matrix, $\boldsymbol{K}\in\mathbb{R}^{n\times n}$, assuming a uniform probability distribution, given the measurement data, $\hat{\boldsymbol{x}}(t_k)\in\mathbb{R}^n$ for the time index, $k=1,\dots,m$:

$$L\big(\boldsymbol{K}\big|\{\hat{\boldsymbol{x}}(t_1),\ \hat{\boldsymbol{x}}(t_2),\cdots,\hat{\boldsymbol{x}}(t_m)\}\big) \tag{21}$$

where K_{ij} is equal to k_{ij} and assumed to be time-invariant. The matrix, $\boldsymbol{K}$, is related to the kinetic matrix $\boldsymbol{A}_k$ in (9) given by

$$\boldsymbol{A}_k = \boldsymbol{K}^{\mathrm{T}} - \mathrm{diag}(\boldsymbol{K}\boldsymbol{e}) \tag{22}$$

where $\boldsymbol{e} \in \mathbb{R}^n$ is a vector of all ones. By using Bayes theorem with the Markov process assumption, the likelihood function can be defined by

$$L\left(\boldsymbol{K}\middle|\{\hat{\boldsymbol{x}}(t_1), \hat{\boldsymbol{x}}(t_2), \cdots, \hat{\boldsymbol{x}}(t_m)\}\right) \\ = P(\hat{\boldsymbol{x}}(t_1)|\boldsymbol{K}) \prod_{k=1}^{m-1} P(\hat{\boldsymbol{x}}(t_{k+1})|\hat{\boldsymbol{x}}(t_k), \boldsymbol{K}) \tag{23}$$

Each conditional PDF in the multiplication of (23) is exactly same as (18), which is the exact solution of the CME for the arbitrary deterministic initial condition,

$$L\left(\boldsymbol{K}\middle|\{\hat{\boldsymbol{x}}(t_1), \hat{\boldsymbol{x}}(t_2), \cdots, \hat{\boldsymbol{x}}(t_m)\}\right) \\ = \prod_{k=1}^{m-1} \begin{pmatrix} \mathcal{M}\left(\hat{\boldsymbol{x}}(t_{k+1}), \hat{x}_1(t_k), \boldsymbol{\lambda}^{(1)}(t_{k+1})\right) \\ * \mathcal{M}\left(\hat{\boldsymbol{x}}(t_{k+1}), \hat{x}_2(t_k), \boldsymbol{\lambda}^{(2)}(t_{k+1})\right) \\ \vdots \\ * \mathcal{M}\left(\hat{\boldsymbol{x}}(t_{k+1}), \hat{x}_n(t_k), \boldsymbol{\lambda}^{(n)}(t_{k+1})\right) \end{pmatrix} \tag{24}$$

Finally, the exact MLE that finds a parameter matrix having maximum value for the likelihood function is defined by

$$\max_{\boldsymbol{K}} L\left(\boldsymbol{K}\middle|\{\hat{\boldsymbol{x}}(t_1), \hat{\boldsymbol{x}}(t_2), \cdots, \hat{\boldsymbol{x}}(t_m)\}\right) \tag{25}$$

The parameter matrix can alternatively be solved by a more numerically convenient optimization

$$\min_{\boldsymbol{K}} -\log L\left(\boldsymbol{K}\middle|\{\hat{\boldsymbol{x}}(t_1), \hat{\boldsymbol{x}}(t_2), \cdots, \hat{\boldsymbol{x}}(t_m)\}\right) \tag{26}$$

by exploiting monotonicity of the logarithm.

For deducing the model complexity or the interactions of the system, a penalty term can be added to the exact MLE formulation (26) (August and Papachristodoulou, 2009). The regularized exact MLE is defined by

$$\min_{\boldsymbol{K}} -\log L\left(\boldsymbol{K}\middle|\{\hat{\boldsymbol{x}}(t_1), \hat{\boldsymbol{x}}(t_2), \cdots, \hat{\boldsymbol{x}}(t_m)\}\right) + \gamma\|\boldsymbol{K}\|_0 \tag{27}$$

where γ is a nonnegative weight specifying the tradeoff between the model complexity and prediction error. A common method for relaxation of the combinatorial optimization is ℓ_1 regularization (Hesterberg et al., 2008). In addition, constraints for the parameter matrix from prior knowledge can be included. The final constrained optimization is formulated by

$$\min_{\boldsymbol{K}} -\log L\left(\boldsymbol{K}\middle|\{\hat{\boldsymbol{x}}(t_1), \hat{\boldsymbol{x}}(t_2), \cdots, \hat{\boldsymbol{x}}(t_m)\}\right) \\ + \gamma\|\mathrm{vec}(\boldsymbol{K})\|_1 \tag{28}$$

s.t. $\boldsymbol{0} \le \boldsymbol{K} \le \boldsymbol{K}_{\max}$

where $\mathrm{vec}(\boldsymbol{K})$ is the concatenation of the elements of $\boldsymbol{K}$, $\boldsymbol{0} \in \mathbb{R}^{n\times n}$ is a matrix of all zeros, and $\boldsymbol{K}_{\max} \in \mathbb{R}^{n\times n}$ is a proper upper limit of the parameter matrix.

4. SIMULATION STUDY AND DISCUSSION

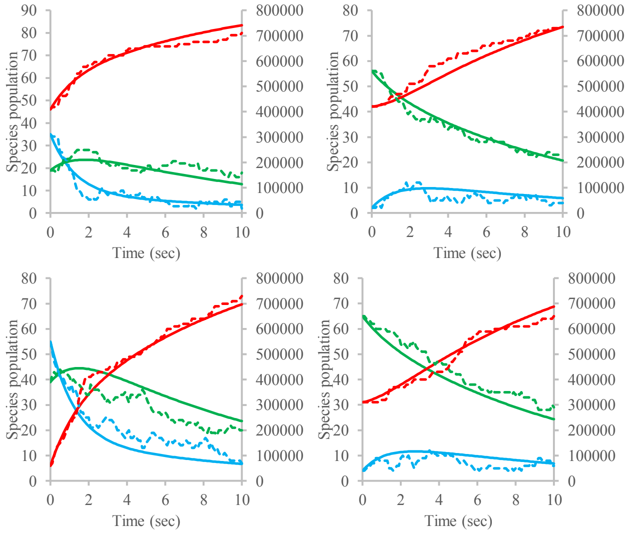

Fig. 1. Four representative sets of *in silico* experimental data generated from the SSA for 3 species (red: A, green: B, blue: C) in the reaction network system (solid line: a population of 10^6, dashed line: a population of 100).

The proposed parameter estimation formulation is applied to a reaction network system containing a population of 100 for 3 species. Potential trajectories are simulated by SSA and used in the proposed parameter estimation method. Species A and B has reversible reaction connectivity and species A and C has irreversible reaction connectivity.

$$B \leftrightarrows A \rightarrow C \tag{29}$$

According to the connectivity of the 3-species system, the true parameter matrix, $\boldsymbol{K}_{\text{true}} \in \mathbb{R}^{3\times 3}$, having zero and non-zero elements is defined by

$$\boldsymbol{K}_{\text{true}} = \begin{bmatrix} 0 & 0.2770 & 0.4 \\ 0.1667 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{30}$$

A total of 1000 *in silico* data sets are simulated for measurable species A, B, and C having different initial conditions. In Fig. 1, four representative data sets among the 1000 are shown for different populations. The data for a population of 10^6 are almost similar to the data from the deterministic simulation, as stochastic fluctuations are insignificant compared to the total population size. On the other hand, for a population of 100, the data exhibit significant stochastic fluctuations relative to total population.

In the parameter estimation, the actual connectivity of species A, B, and C is assumed to be unknown, in which case the reaction parameter matrix, $\boldsymbol{K}$, to be estimated is

$$\boldsymbol{K} = \begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{bmatrix} \rightarrow \begin{bmatrix} 0 & k_{12} & k_{13} \\ k_{21} & 0 & k_{23} \\ k_{31} & k_{32} & 0 \end{bmatrix} \tag{31}$$

The diagonal elements of the matrix, k_{ii}, are set to zero since those elements correspond to non-reactions that do not affect any of the governing reaction equations. With the generated data, the objective is to determine the connectivity of the reactions and associated reaction rate constant using the ℓ_1 regularized exact MLE method (28). The performance of the exact MLE method is compared with the LSE method as a typical reference method for the system identification. The LSE method simply finds the parameter matrix, $\boldsymbol{K}$, minimizing the sum of the squared errors between the stochastic data and predictions based on the deterministic model (Fig. 2),

$$\min_{\boldsymbol{K}} \sum_{i=1}^{m-1} \left(\hat{\boldsymbol{x}}(t_{i+1}) - \left(\hat{\boldsymbol{x}}(t_i) + \delta t \boldsymbol{A}_k \hat{\boldsymbol{x}}(t_i)\right)\right)^2 \qquad (32)$$

The prediction of current state can be obtained with previous measurement data and kinetic matrix, $\boldsymbol{A}_k$, by using the Euler forward method.

The constrained optimization for the exact MLE (28) and the LSE (32) with the same inequality constraints in (28) were implemented in MATLAB. The interior-point algorithm in `fmincon` in the Optimization Toolbox was used for the optimizations. For a population of 10^6, the estimated parameter matrix from the LSE, averaged over the 1000 mock datasets, with $\gamma = 0$ were

$$\widehat{\boldsymbol{K}}_{LSE,Avg} = \begin{bmatrix} 0 & 0.2945 & 0.3448 \\ 0.1447 & 0 & 0.0501 \\ 0.0019 & 0.0084 & 0 \end{bmatrix} \qquad (33)$$

which is reasonably close to the true parameter matrix. On the other hand, for a population of 100, the estimated parameter matrix from the LSE, averaged over the 1000 datasets, with $\gamma = 0$ were

$$\widehat{\boldsymbol{K}}_{LSE,Avg} = \begin{bmatrix} 0 & 0.4754 & 0.4610 \\ 0.1535 & 0 & 0.6420 \\ 0.0623 & 0.2509 & 0 \end{bmatrix} \qquad (34)$$

which is significantly different from the true parameter matrix. The averaged parameter matrix estimated from exact MLE with the same datasets with $\gamma = 0$ was

$$\widehat{\boldsymbol{K}}_{MLE,Avg} = \begin{bmatrix} 0 & 0.2931 & 0.3924 \\ 0.1333 & 0 & 0.1733 \\ 0.0544 & 0.0471 & 0 \end{bmatrix} \qquad (35)$$

which is much closer to the true parameter matrices. This indicates the importance of using the stochastic model in parameter estimation in the case of a small population size.

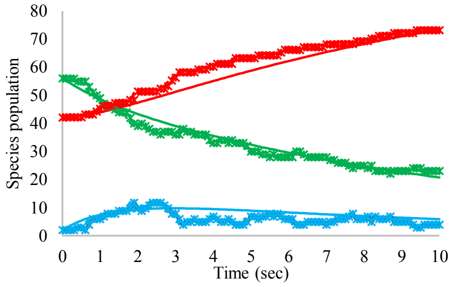

Fig. 2. A representative stochastic realization for populations of 3 species (red: A, green: B, blue: C) and the model predictions obtained by the LSE method (solid line: deterministic, dotted line: stochastic with a population of 100), star mark: model predictions).

In the case of a large population size (e.g., 10^6), use of the deterministic model appears justified.

More specifically, the key performance indices in comparison of the results from the LSE and exact MLE are detection of the sparsity of the parameter matrix, $\widehat{\boldsymbol{K}}$, and estimation of the parameter values, K_{ij} for the connected reactions.

The zero elements in the sparse matrix can be distinguished by using a tolerance criterion which defines an upper limit, μ, for choosing zero elements. For example, if the tolerance $\mu = 0.05$ is set for elements in the matrix results (34) and (35), the matrix (34) has no zero elements and the matrix (35) has one zero element, K_{32}. In the same way, the matrix (34) for $\mu = 0.1$ has one zero element, K_{31}, and the matrix (35) has two zero elements, K_{31} and K_{32}. The exact MLE method detected the sparsity better than the LSE method.

Alternatively, the sparsity detection and accuracy of the constant estimation can be quantified by the mean squared error (MSE) for the zero and non-zero elements of the parameter matrix,

$$P_{\text{Sparsity}} = \frac{|\hat{k}_{31}| + |\hat{k}_{32}| + |\hat{k}_{23}|}{3} \qquad (36)$$

$$P_{\text{Accuracy}} = \frac{(k_{21} - \hat{k}_{21})^2 + (k_{12} - \hat{k}_{12})^2 + (k_{13} - \hat{k}_{13})^2}{3} \qquad (37)$$

In Table 1, the exact MLE shows lower values for both performance indexes calculated by (36) and (37). That is, the proposed method shows better detection of the sparsity and more accurate estimation for the reaction rate constants.

Now consider the performance indices for increased γ in the ℓ_1 regularized exact MLE (28). Although the averaged results from exact MLE (35) shows better performance, it is still not clear that the element k_{23} has no connectivity. With appropriate value of $\gamma > 0$, some of the interconnections is expected to disappear without significantly changing the estimated error in model predictions. Fig. 3 shows the trade-off of the γ value between the sparsity and accuracy. Values of γ ranging from 0 to 20 have minimal reduction in the accuracy of estimates. On the other hand, values of γ over 50 can be used for more accurately determining the reaction connectivity. For example, the averaged parameter matrix for $\gamma = 100$ is

$$\widehat{\boldsymbol{K}}_{MLE,Avg} = \begin{bmatrix} 0 & 0.2753 & 0.0424 \\ 0.0253 & 0 & 0.0020 \\ 0.0213 & 0.0017 & 0 \end{bmatrix}. \qquad (38)$$

Table 1. The sparsity of the matrix, $\boldsymbol{K}$, and the accuracy of the parameter estimates for the LSE and exact MLE methods.

Method	P_{Sparsity}	P_{Accuracy}
LSE	0.3184	0.0144
Exact MLE	0.0917	0.0005

Compared with the results (35), the element k_{23} is further detected as a zero element with the tolerance $\mu = 0.005$. After determining the candidates of the zero elements with a criterion, the remained non-zero parameter can be estimated for obtaining best values for a particular sparsity structure.

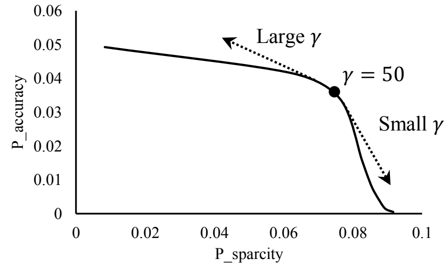

Fig. 3. Parity plot showing the dependency of the sparsity and accuracy of the parameter estimates for ℓ_1 regularized exact MLE for a wide range of γ.

5. CONCLUSIONS

A regularized exact maximum likelihood estimation (MLE) method is presented for determining the interaction topology of a biochemical reaction network system. The regularized exact MLE method is formulated with a closed-form solution of chemical master equations that describe stochastic monomolecular biochemical reaction systems. Improved performance of the exact MLE method is exhibited by using stochastic simulation data for a simple network system and comparing the results with least-squares estimation. The proposed method showed an improved ability to identify a sparse structure of the parameter matrix and estimate the associated reaction rate constants. The proposed method can potentially be used for robust reaction network identification problems such as those found in metabolic pathway or gene regulatory network.

ACKNOWLEDGEMENT

This work was supported by the Advanced Biomass R&D Center (ABC) of Global Frontier Project funded by the Ministry of Science, ICT and Future Planning (2011-0031354).

REFERENCES

August, E., and Papachristodoulou, A. (2009). Efficient, Sparse Biological Network Determination. *BMC Syst. Biol.*, 3, 25.

Boys, R.J., Wilkinson, D.J., and Kirkwood, T.B.L. (2008). Bayesian Inference for a Discretely Observed Stochastic Kinetic Model. *Stat. Comput.*, 18, 125–135.

Cognet, L., Tsyboulski, D.A., Rocha, J.-D.R., Doyle, C.D., Tour, J.M., and Weisman, R.B. (2007). Stepwise Quenching of Exciton Fluorescence in Carbon Nanotubes by Single-Molecule Reactions. *Science*, 316, 1465–1468.

Daigle, B.D., Roh, M.K., Petzold, L.R., and Niemi, J. (2012). Accelerated Maximum Likelihood Parameter Estimation for Stochastic Biochemical systems. *BMC Bioinformatics*, 13, 68.

Feinberg, M. (1979). *Lectures on Chemical Reaction Networks*, Notes of lectures given at the Mathematics Research Center, University of Wisconsin, Madison.

Fichthorn, K.A., and Weinberg, W.H. (1991). Theoretical Foundations of Dynamical Monte Carlo Simulations. *J. Chem. Phys.*, 95(2), 1090–1096.

Gadgil, C., Lee, C.H., and Othmer, H.G. (2005). A Stochastic Analysis of First-order Reaction Networks. *Bull. Math. Biol.*, 67, 901–946.

Gibson, M.A., and Bruck, J. (2000). Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J. Phys. Chem. A.*, 104, 1876–1889.

Gillespie, D.T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.*, 81(25), 2340–2361.

Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C. (2005). Real-time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123, 1025–1036.

Golightly, A., and Wilkinson, D.J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1, 807-820.

Hesterberg, T., Choi, N., Meier, L., and Fraley, C. (2008). Least Angle and ℓ_1 Penalized Regression: A Review. *Stat. Surv.*, 2, 61–93.

Jahnke, T., and Huisinga, W. (2007). Solving the Chemical Master Equation for Monomolecular Reaction Systems Analytically. *J. Math. Biol.*, 54, 1–26.

Jin, H., Heller, D.A., Kalbacova, M., Kim, J.-H., Zhang, J., Boghossian, A.A., Maheshri, N., and Strano, M.S. (2010). Detection of Single-Molecule H_2O_2 Signalling from Epidermal Growth Factor Receptor Using Fluorescent Single-Walled Carbon Nanotubes. *Nature Nanotech.*, 5, 302–309.

Lillacci, G., and Khammash, M. (2010). Parameter Estimation and Model Selection in Computational Biology. *PLoS Computational Biology*, 6.

Lillacci, G., and Khammash, M. (2013). The Signal Within the Noise: Efficient Inference of Stochastic Gene Regulation Models using Fluorescence Histograms and Stochastic Simulations. *Bioinformatics*, 29, 2311-2319.

MacNamara, S., Bersani, A.M., Burrage, K., and Sidje, R.B. (2008). Stochastic Chemical Kinetics and the Total Quasi-steady-state Assumption: Application to the Stochastic Simulation Algorithm and Chemical Master Equation. *J. Chem. Phys.*, 129, art. no. 095105.

Munsky, B., and Khammash, M. (2006). The Finite State Projection Algorithm for the Solution of the Chemical Master Equation. *J. Chem. Phys.*, 124, art. no. 044104.

Munsky, B., Neuert, G. and van Oudenaarden, A. (2009). Using Gene Expression Noise to Understand Gene Regulation. *Science*, 336, 183-187.

Munsky, B., Trinh, B. and Khammash, M. Listening to the Noise: Random Fluctuations Reveal Gene Network Parameters. *Molecular Systems Biology*, 5.

Neuert, G., Munsky, B., Tan, R.Z., Teytelman, L., Khammash, M., and van Oudenaarden, A. (2013). Systematic Identification of Signal-Activated Stochastic Gene Regulation. *Science*, 339, 584-587.

Poovathingal, S.K., and Gunawan, R. (2010). Global Parameter Estimation Methods for Stochastic Biochemical Systems. *BMC Bioinformatics*, 11, 414.

Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in Gene Expression Underlies Incomplete Penetrance. *Nature*, 463, 913–918.

Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, S. (2010). Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science*, 329, 533–538.

Tian, T., Xu, S., Gao, J., and Burrage, K. (2007). Simulated Maximum Likelihood Method for Estimating Kinetic Rates in Gene Expression. *Bioinformatics*, 23, 84–91.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., and Koeppl, H. Moment-based Inference Predicts Bimodality in Transient Gene Expression. *Proc. Nat. Acad. Sci.*, 109, 8340–8345.