International Journal of Pure and Applied Mathematics
Volume 112 No. 2 2017, 381-416
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: 10.12732/ijpam.v112i2.15
ijpam.eu

DEVELOPMENT OF THE MCR METHOD FOR ESTIMATION OF PARAMETERS IN CONTINUOUS TIME MARKOV CHAIN MODELS

Michele L. Joyner[1] §, Thomas Robacker[2]
[1]Department of Mathematics and Statistics
East Tennessee State University
Johnson City, USA
[2]Department of Mathematics
Warren Wilson College
Asheville, USA

Abstract: Parameter estimation techniques have been successfully and extensively applied to deterministic models but are in early development for stochastic models. In this paper, we introduce a new method, the minimum cost realization method or MCR method, for approximating parameters for a continuous-time Markov chain (CTMC) model. This method is an adaption of well-established techniques used in parameter estimation for deterministic systems to account for the variability inherent in stochastic systems. Comparing this method to an established method, the MCR method provides significantly better estimates for parameter values on the two example models considered.

AMS Subject Classification: 60H35, 60J22, 60J27, 65C20, 45Q05, 49N45
Key Words: MCR method, minimum cost realization, parameter estimation, Markov chain model, stochastic model, inverse problem

Received: November 30, 2016
Revised: December 6, 2016
Published: February 1, 2017

§Correspondence author

1. Introduction

When developing a mathematical model, one needs to determine whether the physical system can be modeled by a deterministic model or requires a stochastic model. A deterministic model incorporates no randomness, i.e., for a given input, the model always produces the same output. In contrast, a stochastic model is one in which the outcome is uncertain. For a given input, there could be multiple outputs. Although all physical systems exhibit a degree of randomness, Kurtz limit theory indicates that for a significantly large population, one can approximate a stochastic system with a deterministic one [3, 8, 13]. However, for small populations, even small perturbations can greatly affect the outcome; therefore, in this situation, a stochastic model is necessary to capture the dynamics of the system. For example, consider a simple predator-prey model as seen in any standard ordinary differential equations course

$$\begin{aligned} \frac{dx}{dt} &= \alpha x - \beta xy \\ \frac{dy}{dt} &= \delta xy - \gamma y. \end{aligned}$$

where x is the prey population, y is the predator population and α, β, δ and γ are given constants.

In Figure 1, we compare this deterministic predator-prey model with ten realizations from a corresponding continuous-time Markov chain model (given in Section 2.1.2) for a small population. We note that in the deterministic model, both the predator and prey population exhibit an oscillatory behavior which continues indefinitely. However, in each of the realizations of the stochastic model, either the predator or the prey population goes extinct. This is due to small perturbations which can occur when the animal population is low (illustrated as a 'valley' in the deterministic solution). This example illustrates how the outcomes can be completely different for the deterministic model as compared to the stochastic model. Therefore, in some instances, it is necessary to use a stochastic model to incorporate the varying dynamics within the system.

Regardless of the type of mathematical model one chooses, parameter estimation is a vital step in the development of the model. The validation of a mathematical model with empirical data allows one to use the model to gain insights into the processes inherent in the system as well as investigate the potential effect of perturbations on or within the system. Parameter estimation techniques for deterministic models have been developed extensively (see [3] and the many references therein); however, techniques for parameter estimation in

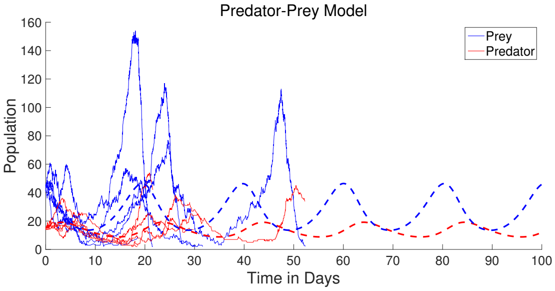

Figure 1: Comparison of a deterministic predator-prey model (dashed line) with ten realizations from a corresponding stochastic model (solid line) for a small population.

stochastic models is still relatively in its infancy [22]. Some techniques involve the likelihood-based methods [17, 19, 20, 21, 22], likelihood-free or bayesian methods [4, 6, 10, 11, 12, 16, 7, 14], and approximation methods using deterministic systems [3, 18].

In this paper, we focus on the development of a parameter estimation method for continuous-time Markov chain (CTMC) models which is an adaptation of the deterministic approximation method first developed by Ortiz et. al. [18]. Our method, the minimum cost realization method or MCR method, allows us to use some of the optimization strategies already in place for deterministic systems; however, it accounts for the limitations imposed by assuming the stochastic model can be well approximated by the deterministic model. As shown in the predator-prey example above, the dynamics captured by a deterministic model will not always match the dynamics of a corresponding stochastic model. In Section 2, we give a summary of the method developed by Ortiz et. al., discuss the two example models along with the synthetic data sets we use to test the methodology and finally give results of the parameter estimation problem using this deterministic method. We then adapt the deterministic method to construct a new method, the MCR method, in Section 3 which utilizes the dynamics inherent in the continuous-time Markov chain. We give results of the MCR method on the same example models and discuss the results and comparison of results to the deterministic method. In Section 4, we take a closer look at the MCR method and discuss potential variation in parameter estimates. Finally, in Section 5, we make some concluding remarks and discuss future work.

2. Deterministic Approach for Parameter Estimation

The ability to utilize deterministic approaches for parameter estimation in CTMC models allows one to tap into an area containing a vast amount of research. In this section, we summarize a parameter estimation methodology for CTMC models which couples the inverse problem methodology developed for deterministic models with Kurtz's limit theory for CTMC models. For a more detailed description of this methodology, we refer the reader to [3, 18] and the references therein.

Kurtz limit theorem, as it is denoted in [3] and originally developed in [8, 13], is given by the following theorem.

Theorem 1 (Kurtz Limit Theorem)**.** *Let $\boldsymbol{C}^N(t)$ be a continuous-time Markov chain. Suppose that $\lim_{M\to\infty}\boldsymbol{C}^N(0) = \boldsymbol{c}_0$ and for any compact set $\Omega \in \mathbb{R}^n$ there exists a positive constant η_Ω such that*

$$|\boldsymbol{g}(\boldsymbol{c}) - \boldsymbol{g}(\hat{\boldsymbol{c}})| \leq \eta_\Omega |\boldsymbol{c} - \hat{\boldsymbol{c}}|,$$

for $\boldsymbol{c}, \hat{\boldsymbol{c}} \in \Omega$. Then we have

$$\lim_{N\to\infty} \sup_{t\leq t_{\mathrm{f}}} |\boldsymbol{C}^N(t) - \boldsymbol{c}(t)| = 0 \tag{1}$$

almost surely for all $t_f > 0$, where $\boldsymbol{c}$ denotes the unique solution to the system of ordinary differential equations given by

$$\dot{\boldsymbol{c}}(t) = \boldsymbol{g}(\boldsymbol{c}), \quad \boldsymbol{c}(0) = \boldsymbol{c}_0.$$

This theorem warrants the approximation of a stochastic system by a corresponding deterministic one if the population size N is sufficiently large. In this scenario, the parameters for the CTMC model can be estimated by first approximating the model with its deterministic counterpart and then applying parameter estimation procedures for a deterministic system.

Therefore, we consider the parameter estimation problem for a deterministic system and proceed as in [3] to estimate parameters of a parameterized dynamical system

$$\begin{aligned} \frac{d\boldsymbol{c}(t)}{dt} &= \boldsymbol{g}(t, \boldsymbol{c}(t), \boldsymbol{\theta}), \\ \boldsymbol{c}(t_0) &= \boldsymbol{c}_0, \end{aligned} \tag{2}$$

where $\boldsymbol{g}$ is the right hand side of the deterministic system, $\boldsymbol{c}$ is the state vector, and $\boldsymbol{\theta}$ the vector of parameters.

One possible statistical model for the observation process (and the only statistical model we consider in this paper) is of the form

$$\boldsymbol{C}_j = \boldsymbol{f}(t_j; \boldsymbol{\theta}_0) + \boldsymbol{\mathcal{E}}_j, \quad j = 1, \ldots, n, \tag{3}$$

where $\boldsymbol{c}(t) = \boldsymbol{f}(t; \boldsymbol{\theta})$ is the solution of Equation (2) and $\boldsymbol{\mathcal{E}}_j$ is assumed to be normally distributed with unknown variance. This is the familiar ordinary least squares formulation. In other words, we assume that there exists an optimal vector of parameters $\boldsymbol{\theta}_0$ such that the random variable $\{\boldsymbol{C}_j\}_{j=1}^n$ (from which our data $\{\boldsymbol{c}_j\}_{j=1}^n$ is just one realization) can be written as the solution of the deterministic system $\boldsymbol{f}(t_j; \boldsymbol{\theta}_0)$ with exact parameter values $\boldsymbol{\theta_0}$ plus some measurement error $\boldsymbol{\mathcal{E}}_j$. For the statistical model given by Eq. (3), we define the vector of optimal parameter values as

$$\boldsymbol{\theta}_{OLS} = \arg\min_{\theta} J(\boldsymbol{\theta}) \tag{4}$$

where

$$J(\boldsymbol{\theta}) = \sum_{j=1}^{n} [\boldsymbol{C}_{\mathsf{j}} - \boldsymbol{f}(t_j; \boldsymbol{\theta})]^2 \tag{5}$$

denotes the cost function. We note that $\boldsymbol{\theta}_{OLS}$ is a random vector; hence if our data set $\{\boldsymbol{c}_j\}_{j=1}^n$ is one realization of the random variable $\{\boldsymbol{C}_j\}_{j=1}^n$, then solving

$$\hat{\boldsymbol{\theta}}_{OLS} = \arg\min_{\theta} \sum_{j=1}^{n} [\boldsymbol{c}_{\mathsf{j}} - \boldsymbol{f}(t_j; \boldsymbol{\theta})]^2 \tag{6}$$

provides a realization for $\boldsymbol{\theta}_{OLS}$. Throughout the paper, we will often drop the subscript OLS for the estimates when the context is clear and simply use $\hat{\boldsymbol{\theta}}$. Ordinary least squares is a commonly used method for parameter estimation in deterministic systems. We refer the reader to [3] for a generalized discussion of other possible statistical observation models with their corresponding optimization techniques.

To summarize, the *deterministic approach* for estimating parameters of a CTMC model assumes the appropriate model for the given data is truly stochastic in nature and should be modeled with a CTMC model. However, the CTMC model is approximated by an appropriate deterministic model. Then, the data is ‘compared’, in the least squares sense, to the solution of the approximate deterministic model in order to estimate the parameters of the system. This is done using applicable optimization strategies for deterministic systems. We now discuss the implementation and results of this approach for two CTMC models.

2.1. Example models

We have chosen two simple example models on which to test both the deterministic approach and MCR method for parameter estimation in a CTMC model. The first model is the SIS model which is typically used to model the spread of disease. A SIS epidemic model is based on the assumption that the infected individuals lose their immunity after some time. This model has been applied to diseases such as influenza or the common cold as well as some sexually transmitted diseases [2].

The second model we will consider in this paper is the Lotka-Volterra Predator-Prey model which was discussed briefly in the introduction. The simplest model of predator and prey interaction includes only natural growth or decay and the predator-prey interaction. The deterministic model can be developed from first principles as in [5] and many other elementary texts on differential equations.

2.1.1. The SIS Model

In the SIS epidemic model, susceptible individuals (S) become infected (I) but do not develop immunity after they recover. Individuals that become infected are also infectious. It is assumed that $1/\gamma$ is the average length of the infectious period. During this time, infectious individuals can transmit the infection to others; we let the parameter β represent the transmission rate, i.e. the effective number of contacts per unit time that result in an infection of a susceptible individual. We assume a fixed, homogeneously mixing population consisting of N individuals. Given these assumptions, the deterministic SIS epidemic model is given by

$$\begin{aligned} \frac{dS}{dt} &= \gamma I - \tfrac{\beta}{N} SI \\ \frac{dI}{dt} &= \tfrac{\beta}{N} SI - \gamma I \end{aligned}$$

with $S(0) + I(0) = N$. Since $S + I = N$ is fixed, we need to only consider one population; we choose the infected population. Therefore, the model reduces to

$$\frac{dI}{dt} = \frac{\beta}{N}(N - I)I - \gamma I. \tag{7}$$

In this paper, however, we assume the transmission of the disease can be more accurately represented by a stochastic model, namely a continuous time Markov chain (CTMC) model. In this model, transitions no longer occur with certainty; instead, we consider the probability of a transition during a small interval of time Δt. Let $I(t)$ denote the random variable for the number of infected individuals at time t. As derived in [1], the CTMC model is then given by

$$\text{Prob}\,\{\Delta I(t) = j \quad | \quad I(t) = i\} = \begin{array}{ll} \dfrac{\beta}{N} i(N-i)\Delta t + o(\Delta t), & j = 1 \\ \gamma i \Delta t + o(\Delta t), & j = -1 \\ 1 - [\gamma i + \beta i(N-i)]\,\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq 0, 1, -1 \end{array} . \tag{8}$$

where $\Delta I(t) = I(t + \Delta t) - I(t)$ and $i \in \{0, 1, \ldots, N\}$. In Figure 2, we plot ten stochastic realizations of the SIS CTMC model for $N = 1250$, $I_0 = 0.04N$, and parameters $\beta = 0.125$ and $\gamma = 0.1$. The dashed curves show the corresponding deterministic solution.

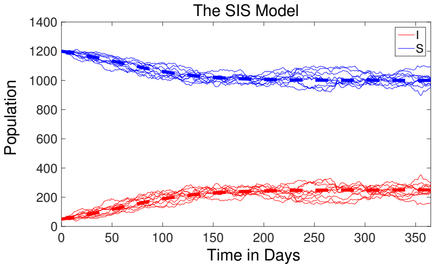

Figure 2: Ten stochastic realizations of the SIS model with $N = 1250$ and $I = 0.04N$ with parameters $\beta = 0.125$, $\gamma = 0.1$. The blue curves are the susceptible population, and the red curves are the infected individuals. The dashed lines give the deterministic solution.

Throughout the paper, we will seek to estimate the parameters $\beta = 0.125$ and $\gamma = 0.1$ for the SIS CTMC model given by Equation (8). We consider populations of size $N = 125$, 1250, and 12500. For $N = 125$ and $N = 1250$, we will look at three different synthetic data sets and one data set for $N = 12500$ for illustrative purposes. All synthetic data sets are realizations of the model in Equation (8) (with exact parameter values) simulated using the Gillespie algorithm [9]. In this section, we assume the data contains no noise (we investigate the addition of noise in the data in Section 4.3). The proportion of initial infective to susceptible individuals will remain the same for each population size at $I_0 = 0.04N$. Figure 3 shows the three data sets for population size $N = 125$;

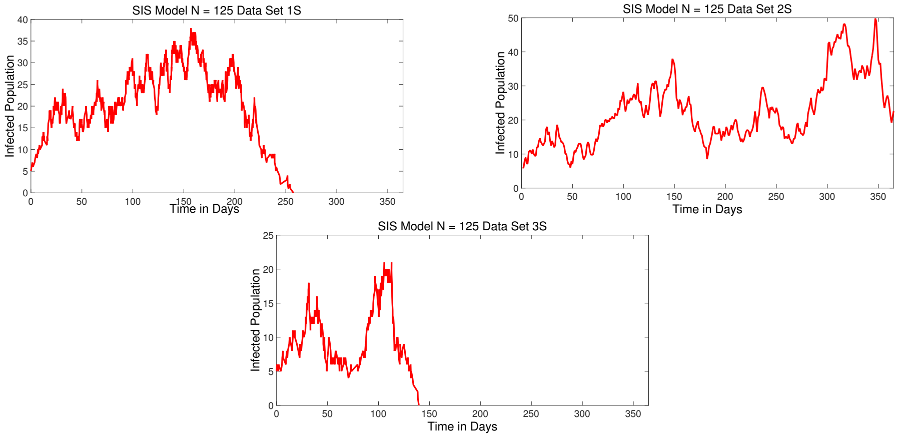

Figure 3: Three data sets for the SIS model labeled data sets 1S, 2S, and 3S for population size $N = 125$.

Figure 4 gives the three data sets for the SIS CTMC model with size $N = 1250$, and Figure 5 illustrates the data set for the SIS CTMC model with population size $N = 12500$.

We note that these realizations were chosen due to the dynamics of each of the realizations. For the small population size, $N = 125$, we have chosen a data set (data set 2S in Figure 3) in which the infected population remains for the entire length of the study, 365 days. On the other hand, in data set 1S, the infected population dies off shortly after 250 days. In data set 3S, the infection lasts for the shortest period of time, a little less than 150 days. We note that the dynamics in all of these data sets are different from those inherent in the deterministic solution for which the solution continues for the entire period and has a limiting behavior. The behavior of the solution to the deterministic model, although scaled for the population size, is illustrated in Figure 2. For the large population size, $N = 1250$, more of the limiting behavior is captured in the CTMC model data, especially in data set 1L (see Figure 4). We note that in all three data sets for the large population, the infection remains for the entire 365 days; however, in data set 3L, there is a noticeable decrease in the number of the infected population between about 200 and 300 days. As expected from Kurtz limit theorem, the synthetic data from the CTMC model with a very large population of $N = 12500$ (Figure 5) behaves most similarly to the dynamics of the deterministic system in which the stochastic effects become less important in the overall dynamics of the solution.

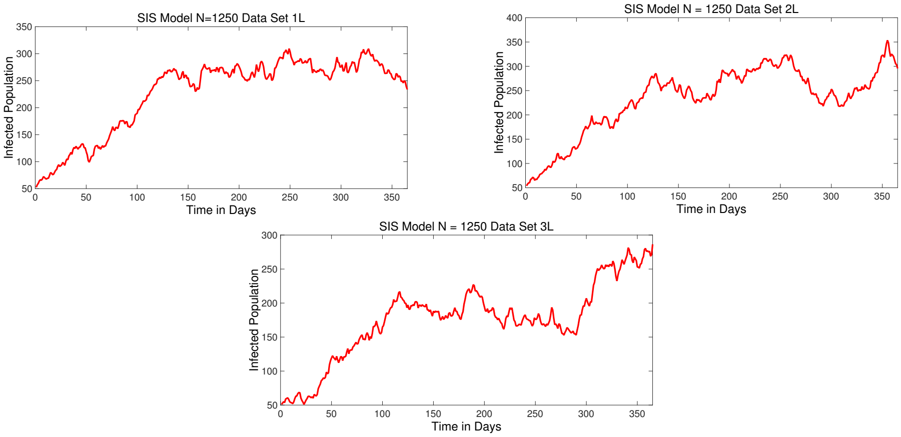

Figure 4: Three data sets for the SIS model labeled data sets 1L, 2L, and 3L for population size $N = 1250$.

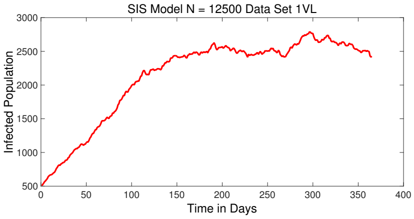

Figure 5: A data set for the SIS model labeled data set 1VL for population size $N = 12500$.

2.1.2. The Lotka-Volterra Predator-Prey Model

In this section we consider our second example model, the Lotka-Volterra predator-prey model. Let $x(t)$ and $y(t)$ denote the population sizes for the prey and predator at time t, respectively. The deterministic Lotka-Volterra predator-prey model we consider in this paper is given by the system of ODEs

$$\begin{aligned} \tfrac{dx}{dt} &= x\left(a_{10} - \frac{a_{12}}{N}y\right) \\ \tfrac{dy}{dt} &= y\left(\frac{a_{21}}{N}x - a_{20}\right) \end{aligned} \tag{9}$$

where the parameters $a_{ij} > 0,\ x(0) > 0$,and $y(0) > 0$. We include N in the equation in order to be able to scale the model to have similar dynamics for differing initial population sizes N. The parameter a_{10} represents the combination of the natural birth and death rate of the prey. The parameter a_{12} represents a death rate in the prey due to interaction with predators, and a_{21} represents a birth rate for the predator due to the same interaction with the prey. Finally, the parameter a_{20} represents the combination of the natural birth and death rate of the predator.

For the CTMC model, let $X(t)$ and $Y(t)$ denote random variables for the population sizes of the prey and predator at time t, respectively. As developed in [1], the CTMC model for this system is given by

$$\begin{aligned} &\mathrm{Prob}\left\{\Delta X(t) = i, \Delta Y(t) = j | X(t) = x, Y(t) = y\right\} \\ &= \begin{cases} a_{10}x\Delta t + o(\Delta t), & (i,j) = (1,0) \\ \dfrac{a_{12}}{N}xy\Delta t + o(\Delta t), & (i,j) = (0,1) \\ \dfrac{a_{21}}{N}xy\Delta t + o(\Delta t), & (i,j) = (-1,0) \\ a_{20}y\Delta t + o(\Delta t), & (i,j) = (0,-1) \\ 1 - x[a_{10} + a_{21}y]\Delta t & \\ -y[a_{20} + a_{21}x]\Delta t + o(\Delta t), & (i,j) = (0,0) \\ o(\Delta t), & \text{otherwise.} \end{cases} \end{aligned}$$

where $\Delta X(t) = X(t + \Delta t) - X(t)$ and $\Delta Y(t) = Y(t + \Delta t) - Y(t)$. Figure 1, given in the introduction, shows ten stochastic realizations of the CTMC along with the deterministic solution with $N = 60$, $X(0) = 0.75N$ with parameters $a_{10} = 0.50,\ a_{12} = 0.05,\ a_{12} = 0.01$, and $a_{20} = 0.20$.

Throughout this paper, we will use these parameters for the predator-prey model. We consider initial populations sizes of $N = 60$ and 600. For $N = 60$, we examine three different synthetic data sets and two data sets for $N = 600$. Analogous to the SIS model, we assume that only the predator population $Y(t)$

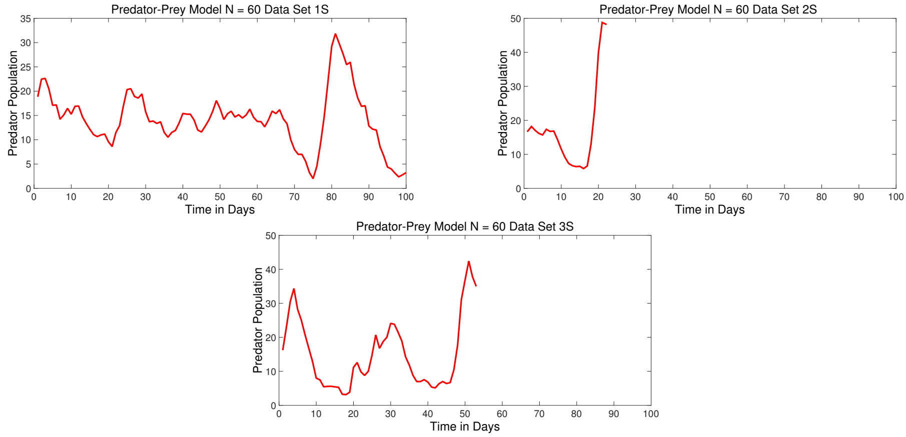

Figure 6: Three data sets for the Predator-Prey model labeled data set 1S, 2S, and 3S for population size $N = 60$.

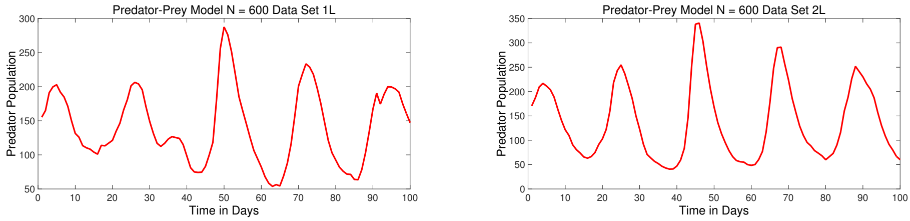

Figure 7: Two data sets for the Predator-Prey model labeled data set 1L and 2L for population size $N = 600$.

can be tracked; therefore, each synthetic data set is a realization of the CTMC model given by Equation (10) for the predator population. The proportion of initial predators will remain constant for each population size at $Y_0 = 0.25N$. Figure 6 illustrates the three data sets for the predator-prey CTMC model with population size $N = 60$, and Figure 7 gives the two data sets for population size $N = 600$.

In examining the small initial population synthetic data sets given in Figure 6 ($N = 60$), we note that in data set 1S, both populations remain for the entire period of the study, 100 days. However, in data sets 2S and 3S, the prey population dies off, leaving only the predator population. Once the prey population dies off, we assume it is no longer necessary to track the predator population. Therefore, for these scenarios, we consider the time at which the prey population dies off to be the end of the data set. For both of the large data sets (Figure 7), both populations remain for the entire time period of study and also exhibit more of an oscillatory behavior as seen in the deterministic model (Figure 1).

2.2. Results for parameter estimation: deterministic approach

In the *deterministic approach*, we assume we want to estimate parameters $\boldsymbol{\theta}$ in a CTMC model. For the SIS CTMC model given by Equation (8), the parameters are given by $\boldsymbol{\theta} = [\beta, \gamma]$. For the predator-prey CTMC model given by Equation (10), the parameters are given by $\boldsymbol{\theta} = [a_{10}, a_{12}, a_{21}, a_{20}]$. In doing the parameter estimation in this paper, we assume that the given CTMC model has already been chosen as the most appropriate model to capture the dynamics of the given system. Model selection criteria for choosing one CTMC model (e.g. SIS) over another (e.g. SEIS) would need to be explored in future research. We further assume we have data $\{\mathbf{c}_j\}_{j=1}^{n}$ from the physical system. In our simulations, we are using the synthetic data sets discussed in Sections 2.1.1 and 2.1.2 and shown in Figures 3-7. The steps for estimating the parameters are given in Algorithm 1.

Algorithm 1 Deterministic Approach Algorithm

Step 1: Derive an appropriate deterministic approximation for the CTMC model,

$$\frac{d\boldsymbol{c}(t)}{dt} = \boldsymbol{g}(t, \boldsymbol{c}(t), \boldsymbol{\theta}).$$

See [3, 18] for details on this derivation. (For example, the appropriate deterministic approximation for the SIS CTMC model is given by Equation (7), and Equation (9) gives an appropriate deterministic approximation for the predator-prey CTMC model.)

Step 2: Generate an initial educated guess for parameter values. (In the estimation of parameters for both models, we generated a normally distributed guess about the true values with 10% relative noise in the parameter.)

Step 3: Use an appropriate optimization method for minimizing the cost function

$$J(\boldsymbol{\theta}) = \sum_{j=1}^{n} [\boldsymbol{c}_{\mathrm{j}} - \boldsymbol{f}(t_j; \boldsymbol{\theta})]^2$$

given the initial guess in Step 2, where $f(t, \boldsymbol{\theta})$ is the solution of the deterministic approximation from Step 1. (In this paper, we used the program *fminsearch* in Matlab [15] which utilizes a Nelder-Mead simplex method. Other strategies such as genetic algorithms, Bayesian techniques, etc. can be implemented.)

Each realization of a CTMC model is different. The seven data sets given in Section 2.1.1 for the SIS CTMC model and five data sets given in Section 2.1.2 for the predator-prey model are just a handful of the many different data sets one might encounter for the *same* CTMC model with the *same* parameter values. Therefore, for each population size N, we generated 95% confidence intervals for estimating parameters for a model given any data set where the true values are given by

$$\theta_0 = [\beta, \gamma] = [0.125, 0.1] \tag{10}$$

in the SIS CTMC model and

$$\theta_0 = [a_{10}, a_{12}, a_{21}, a_{20}] = [0.5, 0.05, 0.01, 0.20] \tag{11}$$

in the predator-prey CTMC model. Algorithm 2 was implemented to estimate confidence intervals using the deterministic approach.

Algorithm 2 Algorithm for Confidence Intervals for Deterministic Approach

```
Initialize the count $k = 1$
**while** $k < K$(We use K=1000 in this paper) **do**
    • Generate a synthetic data set, $\{\mathbf{c}_j^k\}_{j=1}^n$:
        Use the Gillespie algorithm to simulate one realization of the CTMC model
        using exact parameter values. (Note: we ensured no two realizations were
        identical by fixing the initial seed value for the random number generator
        in Matlab to a different value for each data set.)
    • Follow the deterministic approach algorithm given above for obtaining parameter
      estimate $\boldsymbol{\theta}_k$ given data set $\{\mathbf{c}_j^k\}_{j=1}^n$
    • Update $k = k + 1$
**end while**
Construct the confidence interval which is given by
    $(\hat{\boldsymbol{\theta}} - z^* \frac{\boldsymbol{s}}{\sqrt{K}}, \hat{\boldsymbol{\theta}} + z^* \frac{\boldsymbol{s}}{\sqrt{K}})$
where $\hat{\boldsymbol{\theta}}$ is the mean of $\{\hat{\boldsymbol{\theta}}_k\}_{k=1}^K$, $z^*$ is the critical value, and $\boldsymbol{s}$ is the vector sample
standard deviation.
```

2.2.1. SIS Model Results

In this section we discuss the results of using the deterministic approach for parameter estimation in the CTMC SIS model using the synthetic data sets in

Table 1: Estimated parameter values for the very large $N = 12500$ SIS model.

	β			γ		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1 VL	0.125	0.125	.01 %	0.1	0.0991	0.86 %

Table 2: Estimated parameter values for the large $N = 1250$ SIS model.

	β			γ		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1L	0.125	0.1150	8.02 %	0.1	0.0894	10.59 %
2L	0.125	0.1376	10.07 %	0.1	0.1074	7.42 %
3L	0.125	0.1523	21.82 %	0.1	0.1269	26.91 %

Figures 3-5. Tables 1-3 give the results of the parameter estimation for the very large, large, and small SIS models with $N = 12500$, $N = 1250$, and $N = 125$, respectively.

From these results we can see that the deterministic approach worked very well for the very large population with $N = 12500$, mediocre for the large model with $N = 1250$, and unacceptably poor for two cases regarding the small model with $N = 125$. The results can be understood intuitively by re-examining Figures 3 - 5. The worst estimates result from the realizations whose infected populations vanish early, i.e. small data sets 1S and 3S; whereas, we have better estimations for the realizations that persist through 365 days and, more precisely, for those in which the data set most resembles the trend inherent in the deterministic system. The CTMC model most resembles the deterministic model for the very large population as we would expect from Kurtz Limit Theorem; in this case, the deterministic approach for parameter estimation provides an extremely accurate estimate.

To determine whether the accuracy (or lack of accuracy) in parameter es-

Table 3: Estimated parameter values for the small $N = 125$ SIS model.

	β			γ		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1S	0.125	2.2483	1698.64 %	0.1	1.9894	1889.70 %
2S	0.125	0.1422	13.76 %	0.1	0.1127	12.70 %
3S	0.125	2.8421e-16	100.00 %	0.1	0.0038	96.20 %

Table 4: Confidence intervals for parameter estimates in the SIS CTMC model using deterministic approach.

$N = 12500$			
Parameter	True Value	Confidence Interval	Max Rel. Error
β	0.125	(0.1241, 0.1272)	1.90 %
γ	0.1	(0.0993, 0.1019)	1.76 %

$N = 1250$			
Parameter	True Value	Confidence Interval	Max Rel. Error
β	0.125	(0.1255, 0.1368)	9.44 %
γ	0.1	(0.1008, 0.1101)	10.10 %

$N = 125$			
Parameter	True Value	Confidence Interval	Max Rel. Error
β	0.125	(0.3875, 1.0049)	703.92 %
γ	0.1	(0.3782, 0.9834)	883.40 %

timates was inherent to these specific data sets or if the trend was generalized for given population sizes, we followed the algorithm given in Algorithm 2 to generate 95% confidence intervals for the SIS model for each population size. As such, we considered 1000 different synthetic data sets from Equation (8) with parameters in Equation (10) and sought to estimate these parameter values using the deterministic approach. Table 4 gives the confidence intervals along with a column labeled maximum relative error. This value indicates that there is a 95% confidence of the calculated parameter estimate having less than this maximum relative error when compared to the exact value. As exhibited in the specific data sets, we see that the estimation in general is robust for a very large population $N = 12500$. For this population size, we expect 95% of the estimated parameters to have at most 1.90% and 1.76% relative error for β and γ, respectively. The deterministic approach also provided fairly accurate estimation results for the large population as well. The small population in general, however, produces unacceptable estimates. Clearly, the deterministic method for parameter estimation failed for the small population and succeeded for the large populations which we expect by Kurtz Limit Theorem.

2.2.2. Predator-Prey Model Results

In this section, we apply the deterministic approach for parameter estimation on the second model, the predator-prey CTMC model in Equation (10) using

Table 5: Estimated parameter values for the large $N = 600$ Predator-Prey model.

	a_{10}			a_{12}		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1L	0.50	0.4252	14.96%	0.05	0.0414	17.18 %
2L	0.50	0.6167	23.34%	0.05	0.0629	25.81 %

	a_{21}			a_{20}		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1L	0.01	0.0112	11.53%	0.20	0.1954	2.31 %
2L	0.01	0.0164	63.76%	0.20	.1789	10.66 %

the synthetic data sets in Figures 6 and 7. Tables 5 and 6 give the results of the parameter estimation for the large ($N = 600$) and small ($N = 60$) initial population models, respectively. If we examined only the results for these specific data sets, they seem to indicate that very few parameters can be estimated well for either initial population size. Depending on the data set, the results can vary drastically. The results are most likely due to the inability to use the deterministic model to accurately approximate the CTMC model for these particular data sets. However, the given data sets are only a few of the infinitely many different possible data sets which can be generated from the model. As we did in the previous section, we generate 1000 different synthetic data sets from Equation (10) with parameters in Equation (11) and create confidence intervals using Algorithm 2. The results are given in Table 7 for initial population size $N = 600$ and in Table 8 for initial population size $N = 60$. We note that, in general, the deterministic approach did well in estimating the parameters for the larger initial population size with less than 10% maximum relative error for all the parameters when looking at the 95% confidence interval (see Table 7); however, the results were unreliable in the small initial population model (Table 8). Therefore, in both example models, the deterministic approach is a viable approach if the population is ‘large enough’; however, it does not work for models with small population sizes. In the next section, we develop the MCR method for parameter estimation which still can utilize the many techniques available for a least squares optimization problem; however, it also accounts for the variation inherent in the realizations of a CTMC model, thus becoming a reliable parameter estimation method for both large and small population models.

Table 6: Estimated parameter values for the small $N = 60$ Predator-Prey model.

	a_{10}			a_{12}		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1S	0.50	0.6190	23.8%	0.05	0.0584	16.80 %
2S	0.50	2.0583	311.66%	0.05	0.1687	237.40 %
3S	0.50	.6819	36.38%	0.05	0.0494	1.20 %
	a_{21}			a_{20}		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1S	0.01	0.0081	19.00%	0.20	0.1676	16.20 %
2S	0.01	0.027	170.00%	0.20	0.1229	38.55 %
3S	0.01	.0144	44.00%	0.20	0.1510	24.50 %

Table 7: Confidence intervals for the $N = 600$ Predator-Prey model.

Parameter	True Value	Confidence Interval	Max Rel. Error
a_{10}	0.50	(0.5010, 0.5359)	7.18 %
a_{12}	0.05	(0.0506, 0.0546)	9.20 %
a_{21}	0.01	(0.0106, 0.0106)	6.00 %
a_{20}	0.20	(0.2003, 0.2163)	8.15 %

Table 8: Confidence intervals for the $N = 60$ Predator-Prey model.

Parameter	True Value	Confidence Interval	Max Rel. Error
a_{10}	0.50	(0.4642, 0.5817)	16.34 %
a_{12}	0.05	(0.0348, 0.0438)	30.40 %
a_{21}	0.01	(1988.7811, 12155.4461)	12.15e7 %
a_{20}	0.20	(1491.9326, 9116.9226)	45.585e5 %

3. MCR Method for Parameter Estimation

In the development of the MCR method for parameter estimation in CTMC models, it is helpful to visualize why the deterministic approach did not work well for these small population models, models in which the random variation is most pronounced. Recall that in the deterministic approach, we are minimizing Equation (5) given by

$$J(\boldsymbol{\theta}) = \sum_{j=1}^{n} [\boldsymbol{c}_{\mathrm{j}} - \boldsymbol{f}(t_j; \boldsymbol{\theta})]^2$$

where $\{\boldsymbol{c}_j\}_{j=1}^{n}$ is the data and $\boldsymbol{f}(t_j, \boldsymbol{\theta})$ is the solution of the deterministic system at time t_j given parameter $\boldsymbol{\theta}$. Therefore, in this method, one wants to minimize the sum of squared differences between the data and deterministic solution at each time step. Figure 8 gives an illustration of why the minimization produced the given optimal parameters as opposed to parameters closer to the exact value for the SIS CTMC synthetic data set 3S. The solid black line gives the solution to the deterministic SIS model with the exact parameters, $\boldsymbol{\theta}_0 = [\beta, \gamma] = [0.125, 0.1]$. The dashed red line is the solution of the deterministic model with the estimated optimal parameters, $\hat{\boldsymbol{\theta}} \approx [0, 0.0038]$. It is easy to visualize why the cost function J was smaller with the estimated parameters than for the exact parameters; the solution with the estimated parameters averages out the trend inherent in the data. Nonetheless, both deterministic solutions are smooth trends with no variation; therefore, we do not expect either of these solutions to follow the exact trend of the data like a realization of the CTMC model might. This is the basis for the MCR method. Instead of minimizing the squared difference between the deterministic solution, which is smooth, and the data, which contains a lot of variation, especially when considering models with small populations, we would like to be able to compare the data to simulations of the CTMC model, $\boldsymbol{h}(t_j, \boldsymbol{\theta})$, which are inherently more like the data than solutions to the deterministic system.

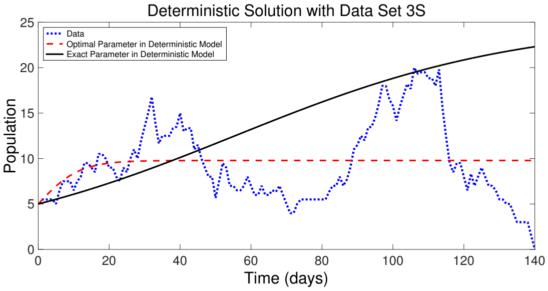

Figure 8: Synthetic SIS CTMC data set 3S is shown (dotted line) together with the solution to the deterministic model (Equation (7)) with both the exact values from Equation (10) (shown as the black solid line) and estimated values from Table 3 (shown as the red dashed line).

The problem in 'comparing' realizations of the CTMC model with the data is that, unlike the deterministic model in which there is a unique solution with a given initial condition and parameter values, there are infinitely many different realizations possible for a CTMC model with the same initial condition and same parameter values. One might consider averaging the solutions of the CTMC model and then using this averaged solution in the minimization problem; however, although there is more variation in this averaged solution of the CTMC model than in the deterministic model solution, there is still not as much variation as with a single realization. In fact, if the average trend is drastically different than a single realization, the results of the parameter estimation problem are no better than in the deterministic approach. Therefore, instead of averaging the realizations of the CTMC model, the MCR method 'compares' a given number, say M, realizations to the data and 'picks' the one most resembling the data. More precisely, for each of the M realizations,

$$J_m(\boldsymbol{\theta}) = \sum_{j=1}^{n} [\boldsymbol{c}_{\mathsf{j}} - \boldsymbol{h}_m(t_j; \boldsymbol{\theta})]^2 \tag{12}$$

is computed where $\{\boldsymbol{c}\}_{j=1}^{n}$ is the data and $\boldsymbol{h}_m(t; \boldsymbol{\theta})$ is a realization of the CTMC model. The realization in which J_m is smallest, $m = 1, ..., M$ is considered the best realization, or the realization from the CTMC model which most resembles the data in the sum of least squares sense. Figure 9 illustrates this concept in which the data is plotted with seven realizations from the SIS CTMC model. The solid red line is the realization which most resembles the data in the least squares sense, i.e., the realization in which J_m is smallest for the chosen optimal parameter value $\boldsymbol{\theta}$. The complete MCR algorithm is given by Algorithm 3.

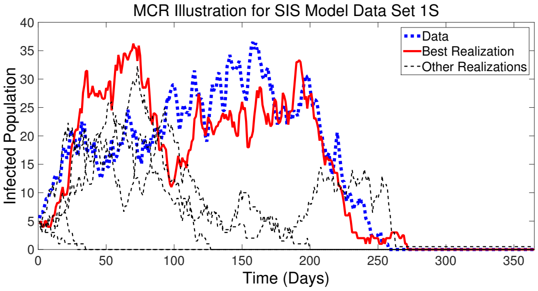

Figure 9: Illustration of the MCR Method. The blue curve represents data set 1S from the SIS CTMC model. The red curve is the realization that best fits the data set. The black curves are several other realizations that were not the best fit.

The first step in estimating parameters of a CTMC model using the MCR method is to choose the number of realizations M of the CTMC model for which the sum of squared differences between the realization and the data will be computed. We discuss the effect of the choice of M on the resulting accuracy of the parameter estimation in Section 4.2. Next, it is vital to keep the trend in the realization the same while only varying the parameter $\boldsymbol{\theta}$ in the optimization problem. In other words, each time a new $\boldsymbol{\theta}$ is provided by the optimization program, it is an input into the *same* M realizations as in the previous iteration. For example, in Figure 9, these same seven realizations would be used each time with only a varying parameter value $\boldsymbol{\theta}$. To accomplish this, we utilize Matlab and set the initial seed for the random number generator to assure the same realizations are computed each time in the minimization algorithm. (See [15] for a precise explanation of setting the seed for a random number generator.) Next, a priori information should be used to provide an initial guess for the optimization algorithm to estimate

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} J_{MCR}(\boldsymbol{\theta})$$

where

$$J_{MCR}(\boldsymbol{\theta}) = \min_{m \in 1,2,\ldots,M} J_m(\boldsymbol{\theta})$$

and $J_m(\boldsymbol{\theta})$ is given in Equation (12). In this paper, the optimization algorithm we implemented was the Nelder-Mead simplex method via *fminsearch* in Matlab. One subtle step in the process involves evaluating the CTMC model at the time points t_j, $j = 1, ..., n$. The Gillespie algorithm is used to simulate realizations of the CTMC model in which the time until the next transition is drawn from an exponential distribution; therefore, we cannot specify the time steps at which we want the state space of the CTMC model. Therefore, we use a binning algorithm to estimate the CTMC model at the given data time points. In both of the example models in this paper, we assume data is collected daily. Therefore, the binning algorithm is constructed to estimate the state of the CTMC model for each day. The steps are given in Algorithm 4.

3.1. Results for parameter estimation: MCR method

In this section, we analyze the results of using the MCR method to estimate the parameters of the two example CTMC models given the individual data sets in Sections 2.1.1 and 2.1.2. Furthermore, we examine the ability to estimate the parameters for the model in general given any synthetic data set from the model. We note that unlike in the deterministic approach, the result of the MCR method greatly depends on the choice of M realizations used in the cost function J_{MCR}. In this section, we keep M fixed at $M = 10$, i.e., the synthetic data is compared to 10 randomly chosen realizations of the CTMC model in the cost function. In Section 4.2, we examine the accuracy of the parameter estimation problem with varying values of M. Furthermore, in this section we give the results of the parameter estimation for only one set of $M = 10$ randomly chosen realizations; in Section 4.1, we explore how the choice of *which* M realizations are used effects the result as well.

3.1.1. The SIS Model

We first give the results for estimating β and γ in the CTMC SIS model given by Equation (8). Tables 9 and 10 give the results of the parameter estimation for the large and small SIS models with population sizes $N = 1250$ and $N = 125$, respectively. Recall that the deterministic approach provided good estimates for both parameters in the large population CTMC SIS model using all three individual synthetic data sets (Table 2). The parameter estimates obtained using the MCR method (Table 9) were only slightly more accurate than those obtained using the deterministic approach for the large population with a total average increase in accuracy of 4% across the 5 estimates in which there was an improvement (the estimate for γ was not improved for data set 1L). However, for the small SIS population data set 1S, the estimate for β has a relative error of only about 3% (Table 10) compared to over 1600% using the deterministic

Algorithm 3 MCR Method for Parameter Estimation

```
Step 1: Initialize $M$, the number of realizations of the CTMC model which will be 'compared' to the data in Step 4. (We use $M = 10$ in the initial results and explore the effect of the choice of $M$ in Section 4.)
Step 2: Randomly choose $M$ realizations of the CTMC model. (In this paper, we do this by randomly selecting $M$ initial seeds for the random number generator and using the same initial seed value each time in Step 4.1. See [15] for setting an initial seed for the random number generator in Matlab.)
Step 3: Generate an initial educated guess for parameter values. (In the estimation of parameters for both models, we generated a normally distributed guess about the true values with 10% variance.)
Step 4: Use an appropriate optimization method for estimating

    $$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} J_{MCR}(\boldsymbol{\theta})$$

given the initial guess in Step 3. (In this paper, we used the program fminsearch in Matlab [15] which utilizes a Nelder-Mead simplex method. Other strategies such as genetic algorithms, Bayesian techniques, etc. can be implemented.)
    To calculate $J_{MCR}(\theta)$ for each $\theta$, use the following steps:
    Step 4.1: Simulate the $M$ realizations from the CTMC model using parameter values given by $\boldsymbol{\theta}$. Recall: we use the same $M$ stochastic realizations, so only the parameter values are changing, not the overall trend.
    Step 4.2: Bin the $M$ data sets to match the time steps of $\{\boldsymbol{c}_j\}_{j=1}^{n}$ using Algorithm 4.
    Step 4.3: Calculate $J_m(\boldsymbol{\theta})$ for $m = 1, 2, \ldots M$:

        $$J_m(\boldsymbol{\theta}) = \sum_{j=1}^{n} [\boldsymbol{c}_{\mathrm{j}} - \boldsymbol{h}_m(t_j; \boldsymbol{\theta})]^2$$

        where
    Step 4.4: Set cost function as

        $$J_{MCR}(\boldsymbol{\theta}) = \min_{m \in 1,2,\ldots,M} J_m(\boldsymbol{\theta}).$$
```

Algorithm 4 Binning Algorithm

```
Initialize $\boldsymbol{s}_{bin}(1) = \boldsymbol{s}_{CTMC}(1)$ where $\boldsymbol{s}_{bin}$ is the new state vector for the binned data and $s_{CTMC}$ is the original state vector from the Gillespie algorithm
for $day = 1$ to $n$ do
    index=find($day - 1 \leq t_{CTMC} < day$) where $t_{CTMC}$ is the vector of original time steps from the Gillespie algorithm
    if index is not empty then
        $s_{bin}(day) = \text{mean}(s_{CTMC}(index))$
    else
        $s_{bin}(day) = s_{CTMC}(day - 1)$
    end if
end for
```

Table 9: MCR estimated parameter values for the large $N = 1250$ SIS CTMC model using $M = 10$ in J_{MCR}.

	β			γ		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1L	0.125	0.1153	7.76%	0.1	0.0893	10.70 %
2L	0.125	0.1315	5.20 %	0.1	0.1028	2.80 %
3L	0.125	0.1023	18.16 %	0.1	0.0815	18.50 %

method (Table 3). There is a similar improvement in the estimate of γ with the slightly more than 6% relative error using the MCR method compared to more than 1800% relative error using the deterministic approach. Similar results were found for the other small population data sets. Typically, one would like less than 10% relative error. We note that the relative error in the estimate for β given data set 3S is slightly over 18%; however, this estimate is still a remarkable improvement over the estimate using the deterministic approach in which the estimate for β was approximately $\hat{\beta} = 0$ with 100% relative error (see Table 3). These results are for three specific individual data sets for each population size out of the many different synthetic data sets one might have given the same model and parameters.

We use Algorithm 2 to estimate parameters for 1000 different data sets and calculate 95% confidence intervals for the parameter values in both the large $N = 1250$ population CTMC model and small population, $N = 125$ CTMC model; the results are given in Table 11. Although the results for the individual data sets indicate that the MCR method did not provide a significant improvement in estimates for the large population model, if one examines the confidence intervals obtained using the MCR method (Table 11) and those

Table 10: MCR estimated parameter values for the small $N = 125$ SIS CTMC model using $M = 10$ in J_{MCR}.

	β			γ		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1S	0.125	0.1212	3.04%	0.1	0.1064	6.40 %
2S	0.125	0.1362	8.96 %	0.1	0.1061	6.10 %
3S	0.125	0.1034	17.28 %	0.1	0.1033	3.3 %

Table 11: Confidence intervals for parameter estimates in the SIS CTMC model using MCR Method.

$N = 1250$			
Parameter	True Value	Confidence Interval	Max Rel. Error
β	0.125	(0.1207, 0.1223)	3.44 %
γ	0.1	0.0968, 0.0981)	3.20%

$N = 125$			
Parameter	True Value	Confidence Interval	Max Rel. Error
β	0.125	(0.1143, 0.1170)	8.58 %
γ	0.1	(0.1033, 0.1087)	8.73 %

using the deterministic approach (Table 4), we notice that in general, the MCR method provides an improvement of about 6% in parameter estimates over the deterministic approach. The improvement is much more pronounced when estimating parameters in the small population model. We notice that for the small SIS CTMC model, there is a 95% confidence of the exact parameters having less than 9% relative error. This is compared to the results in Table 4 using the deterministic approach in which the maximum relative error in the parameter values is more than 700%.

3.1.2. The Lotka-Volterra Predator-Prey Model

We perform the same analysis as in the previous section, using the MCR method to estimate parameters a_{10}, a_{12}, a_{21}, and a_{20} in the predator-prey CTMC models given by Equation (10). Tables 12 and Table 13 give the results of the parameter estimation for the initial large ($N = 600$) and small ($N = 60$) population models, respectively, still using $M = 10$ in evaluating J_{MCR}. In the large initial population model (Table 12), the MCR method produces more accurate estimates than the deterministic approach (Table 5) for all the parameter val-

Table 12: MCR Estimated parameter values for the $N = 600$ Predator-Prey CTMC model using $M = 10$ in J_{MCR}.

	a_{10}			a_{12}		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1L	0.50	0.4862	2.64%	0.05	0.0533	6.60 %
2L	0.50	0.4863	2.76 %	0.05	0.0533	6.60 %

	a_{21}			a_{20}		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1L	0.01	0.0106	6.00%	0.20	0.1614	19.3 %
2L	0.01	0.0107	7.00 %	0.20	0.1586	20.70 %

Table 13: MCR Estimated parameter values for the $N = 60$ Predator-Prey CTMC model using $M = 10$ in J_{MCR}.

	a_{10}			a_{12}		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1S	0.50	0.4640	7.2 %	0.05	0.0520	4.00 %
2S	0.50	0.4780	4.4 %	0.05	0.0536	7.20 %
3S	0.50	0.4669	6.62 %	0.05	0.0492	1.60 %

	a_{21}			a_{20}		
Data Set	Actual	Estimate	Rel. Error	Actual	Estimate	Rel. Error
1S	0.01	0.0101	1.00 %	0.20	0.1787	10.65 %
2S	0.01	0.0100	0.20 %	0.20	0.1837	8.15 %
3S	0.01	0.0106	6.00 %	0.20	0.1416	29.2 %

ues except a_{20}. We note than all parameter estimates have a relative error of 7% or less (except parameter a_{20}) when using the MCR method compared to a relative error of over 10% in all estimated parameters except one when using the deterministic approach. In addition to improved estimates for the large population model, the utilization of the MCR method also resulted in drastically improved results for most parameter estimates in the initial small population model (Table 13) when compared to the deterministic approach (Table 6). Notice that for the small predator-prey population data set 2S in Table 13, the estimate for a_{12} has a relative error of about 4% compared to over 300% for the deterministic method as seen in Table 6. Similar results hold for every parameter in each data set besides a_{20} for data set 3S.

In addition to using selected individual data sets, we also construct confidence intervals for the parameter estimates to ensure the MCR method works more generally in the small initial population model. Table 14 gives the con-

Table 14: MCR Confidence intervals for the $N = 60$ Predator-Prey CTMC model using $M = 10$ in J_{MCR}.

Parameter	True Value	Confidence Interval	Max Rel. Error CI
a_{10}	0.50	(0.4921, 0.4991)	1.57 %
a_{12}	0.05	(0.0510, 0.0517)	3.35 %
a_{21}	0.01	(0.0098, 0.0100)	1.55 %
a_{20}	0.20	(0.1994, 0.2033)	1.65 %

fidence intervals for the parameters estimates with $N = 60$ using the MCR method with $M = 10$. Comparing this with Table 8, we see an astonishing improvement in estimating all the parameters, especially a_{21} and a_{20}, with the MCR method. The MCR method estimated each parameter with maximum relative error no greater than 4%, an improvement across the board compared to the deterministic approach.

4. Closer Examination of MCR Method

In each of the estimates in the previous section, $M = 10$ was used in the MCR algorithm. In this section, we begin by examining the effect of the choice of *which* M realizations are used in the cost function. We also examine how the *value* of M, i.e., how many realizations of the CTMC model are compared to the data, might effect the accuracy of the estimate. Finally, in all of the simulations to this point, we have used synthetic data which is a direct realization of the given CTMC model. In Section 4.3, we assess the effect of noise in the synthetic data on the results of the parameter estimation problem.

4.1. Effect of *which M* realizations are used

In Section 3.1, the MCR method was implemented using $M = 10$ in the cost function J_{MCR}, i.e., the data was compared, in the least squares sense, to *one* set of randomly chosen realizations of the CTMC model. Depending on the choice of *which* ten realizations are chosen, the results of the parameter estimation will be different. Recall, in the MCR method, for each m, $m = 1, ..., M$,

$$J_m(\boldsymbol{\theta}) = \sum_{j=1}^{n} [\boldsymbol{c}_{\mathsf{j}} - \boldsymbol{h}_m(t_j; \boldsymbol{\theta})]^2$$

Table 15: **Median** estimated parameter SIS CTMC model using the MCR method with 100 different randomly chosen $M = 10$ realizations.

	β			γ		
Data Set	Actual	Med. Est.	Rel. Error	Actual	Med. Est.	Rel. Error
1S	0.125	0.1160	7.20%	0.1	0.0992	0.80 %
2S	0.125	0.1220	2.40 %	0.1	0.0953	4.70 %
3S	0.125	0.1097	12.24 %	0.1	0.1045	4.50%

is computed. For each $\boldsymbol{\theta}$, $J_{MCR}(\boldsymbol{\theta})$ is the minimum of $\{J_m\}_{m=1}^{M}$ where the same M realizations are used throughout the optimization process. We repeated the estimation problem using 100 different randomly chosen sets of 10 realizations for the MCR method. The median parameter estimates are given in Table 15. These results are different than the results using just one data set indicating the choice of *which* M realizations is compared to the given data set is an important factor in the accuracy of the estimate. We note that the median error in the estimate of β for data set 3S is much better than the estimate using the randomly chosen set which gave the results in Table 10; however, the median estimate for β using data set 1S is worse. Nonetheless, the median relative error is low in all estimates; the only estimate with more than 10% median relative error is when using data set 3S to estimate β. Nonetheless, the estimate is still within an acceptable error bounds. Therefore, both the median estimates given in Table 15 as well as the estimates using only one randomly chosen set of 10 realizations given in Table 10 provided good estimates for the parameters.

Even though the results given in Table 10 are acceptable using *that* set of ten realizations, if one is unlucky in the choice of ten realizations, the relative error in the parameter estimate may be extremely considerable. Figure 10 shows the potential spread in the percent relative error when considering 1000 different sets of 10 realizations in the cost function J_{MCR}. We note that in one case, the relative error in β is as large as 95.9% while the relative error in γ is even larger with a 97.6% relative error. Recall that in Figure 9 we illustrated how a realization from the CTMC model could fit well to the data. Figure 11 illustrates why the estimate is so poor in the extreme case with over 95% relative error. The dotted blue curve is the data set and the solid red line is the realization from the CTMC model which results in the lowest value of J_m, $m = 1, .., M$ in Equation (12), i.e., the *best* fit to the data out of the ten realizations from the model. Notice the best fit realization does not follow the trend of the data at all which is why the parameter estimate is so

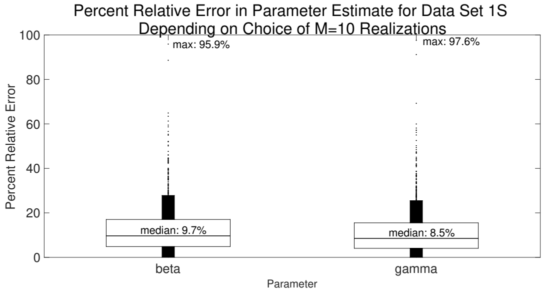

Figure 10: Boxplot for percent relative error in the parameter estimates for Data Set 1S when using 1000 different randomly chosen sets of $M = 10$ realizations in the cost function J_{MCR} for the MCR method.

poor. Therefore, in implementing the MCR method on a given model, it is **not** advisable to assume the parameter estimate is correct given only *one* set of M *randomly* chosen realizations. There is a potential for the estimate to be extremely poor. However, for the example case of estimating β and γ in the small population model using Data Set 1S, the median relative error is less than 10% when using either 100 trials (Table 15) or 1000 trials (Figure 10). Therefore, the MCR method works extremely well when using multiple trials and median parameter estimates. We are already exploring *a priori* methods for initially choosing the set of M realizations to avoid the need to repeat the procedure more than once and still have a high confidence in the estimated value with low computational time. These techniques and results will be reported in a future paper.

4.2. Effect of *value* of M

One conclusion one might draw from Section 4.1 is that one only needs to compare the data to *more* realizations of the CTMC model, i.e., use a larger value for M. The number of realizations, M, *does* have an effect on the accuracy of the estimate. Using 1000 different trials for each $M = 10, 25, 50, 100$, we calculate the median relative error in the parameter estimate for both β and γ in the SIS CTMC model using Data Set 1S. The trend for the median relative error is given in Figure 12 which illustrates a converging behavior as M increases. There is about a 1% improvement in the relative error when going from using $M = 10$ to using $M = 25$ realizations in the cost function; there is further improvement when using $M = 50$ and even more with $M = 100$. As M is increased, the error is decreased; however, the computational time required to

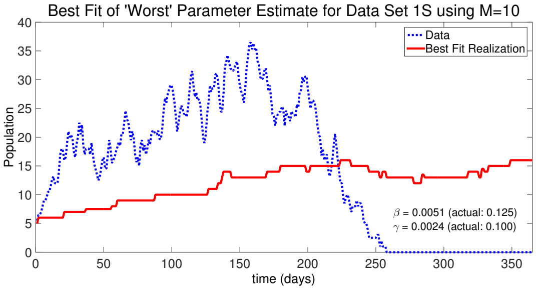

Figure 11: The blue dotted curve represents Data Set 1S. The red curve is the realization that best 'fits' the data set, in the least squares sense, for the estimate resulting in the largest percent relative error in the 1000 trials.

use $M = 100$ realizations in the cost function is much greater than when using $M = 50$. Therefore, there is a trade-off in error reduction and computational time.

In addition to analyzing the trend for the median percent relative error in the parameter estimate, we analyzed the ability to accurately estimate the parameters with only *one* set of $M = 100$ *randomly* chosen realizations. The hypothesis was that when comparing the data to more random realizations of the CTMC model, we would be more likely to avoid the *largest* outlier scenario depicted in Figure 11 in which it was possible for none of the realizations to match the trend in the data closely enough to achieve a good estimate. In Figure 13, we have a boxplot of the percent relative error when using 1000 different sets of $M = 100$ randomly chosen realizations in J_{MCR} given Data Set 1S. We notice that the increase in M *does* reduce the median relative error with a lot fewer outliers when we compare Figure 13 to Figure 10; nonetheless, there still is one large outlier with a percent relative error of 82.1% in the estimate for β and 88.8% in the estimate for γ. This, again, is an unacceptable estimate. Therefore, if one *randomly* chooses M realizations of the CTMC model, it is necessary to *repeat* the estimation multiple times to assure reliable results in the median estimate. In a future paper, we will explore methods in which one can a priori choose *one* set of M realizations of the CTMC model in a systematic manner to ensure a trend in the outlier error measurements similar to that displayed for the median error measurement in Figure 12 as M is increased. If the set of M realizations are chosen well, in other words chosen such that the behavior exhibited in the data is also displayed initially in the set of M

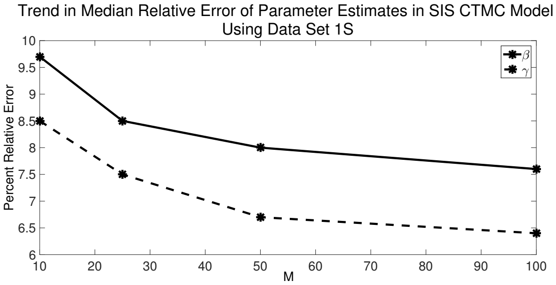

Figure 12: The trend for the median relative error in the parameter estimates for the SIS CTMC model depending on the value of M using Small Pop Data Set 1S

.

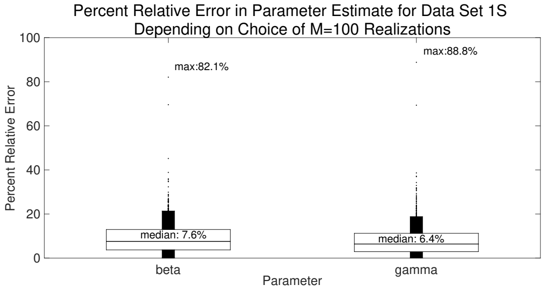

Figure 13: Boxplot for percent relative error in the parameter estimates for Data Set 1S when using 1000 different randomly chosen sets of $M = 100$ realizations in the cost function J_{MCR} for the MCR method.

realizations of the model, the likelihood of extreme errors in the measurement will be reduced.

We further examine the effect of the value of M for several synthetic data sets from the small population SIS CTMC model by creating confidence intervals using Algorithm 2. We see a similar trend in the confidence intervals for 1000 different synthetic data sets as we did when examining the trend in the median relative error calculated using Data Set 1S. There is a decrease in the maximum relative error in the confidence interval as M increases until reaching a minimum expected error. For this data set, the limiting maximum relative error for β is approximately 7% and for γ approximately 1.5-2%. When utilizing the MCR method, we are able to estimate the value of γ fairly accurately with only $M = 10$ realizations in the algorithm for which there is only slightly less

Table 16: MCR 95% Confidence intervals for the $N = 125$ SIS CTMC model with varying M.

M	Parameter	True Value	Confidence Interval	Max Rel. Error CI
10	β	0.125	(0.1143, 0.1170)	8.58 %
	γ	0.1	(0.1033, 0.1087)	8.73 %
25	β	0.125	(0.1150, 0.1173)	7.97 %
	γ	0.1	(0.1011, 0.1039)	3.91 %
50	β	0.125	(0.1156, 0.1175)	7.51 %
	γ	0.1	(0.1003, 0.1028)	2.84 %
75	β	0.125	(0.1163, 0.1183)	6.93 %
	γ	0.1	(0.1008, 0.1028)	2.80 %
100	β	0.125	(0.1163, 0.1182)	6.94 %
	γ	0.1	(0.0998, 0.1016)	1.56 %

than a 9% maximum relative error in parameter estimates in the 95%confidence interval. This maximum relative error drops to slightly more than 1.5% relative error when using $M = 100$ realizations in the algorithm. The parameter β is slightly harder to estimate; however, for all M, we have less than 8.5% relative error in the estimate. We do note that in all the confidence intervals for β, β is underestimated with the true value not contained within the confidence interval. Nonetheless, even with the consistently low estimation, the maximum relative error in the 95% confidence interval is still quite low proving this is a plausible method for parameter estimation for this CTMC model, regardless of the value of M. We did a similar analysis with the small population predator-prey CTMC model with the results given in Table 17. In this example, there is minimal improvement, if any improvement, in the parameter estimate when using $M = 10$ versus $M = 75$. Therefore, the effect of the value of M on the accuracy of the parameter estimate is also model dependent. Increasing M in the SIS CTMC model improved parameter estimates; whereas, increasing M provided little to no improvement in parameter estimates for the predator-prey CTMC model.

4.3. Effect of noisy data

The results up to this point were obtained using exact synthetic data from the CTMC model, i.e. no noise was added to the data. We explore the effect on the parameter estimate when noise is added to the synthetic data for the SIS CTMC model. We start with one example data set, Data Set 1S. Previously, $\{\boldsymbol{x}_j\}_{j=1}^{n}$

Table 17: MCR 95% Confidence intervals for the $N = 60$ Predator-Prey CTMC model with varying M.

M	Parameter	True Value	Confidence Interval	Max Rel. Error
10	a_{10}	0.50	(0.4921, 0.4991)	1.57 %
	a_{12}	0.05	(0.0510, 0.0517)	3.35 %
	a_{21}	0.01	(0.0098, 0.0100)	1.55 %
	a_{20}	0.20	(0.1994, 0.2033)	1.65 %
25	a_{10}	0.50	(0.4925, 0.5001)	1.50 %
	a_{12}	0.05	(0.0509, 0.0517)	3.42 %
	a_{21}	0.01	(0.0098, 0.0100)	1.72 %
	a_{20}	0.20	(0.1999, 0.2036)	1.81 %
50	a_{10}	0.50	(0.4919, 0.4991)	1.62 %
	a_{12}	0.05	(0.0509, 0.0516)	3.23 %
	a_{21}	0.01	(0.0098, 0.0100)	1.60 %
	a_{20}	0.20	(0.1997, 0.2034)	1.68 %
75	a_{10}	0.50	(0.4911, 0.4981)	1.78 %
	a_{12}	0.05	(0.0508, 0.0515)	2.96 %
	a_{21}	0.01	(0.0098, 0.0100)	1.77 %
	a_{20}	0.20	(0.1995, 0.2030)	1.48 %

was computed as a realization of the CTMC model given exact parameter value $\boldsymbol{\theta}$, i.e., the synthetic data without noise is given by

$$\boldsymbol{x}_j = h(t_j, \boldsymbol{\theta}_0),$$

where $h(t, \boldsymbol{\theta})$ is a realization of the model at time t with parameter value $\boldsymbol{\theta}$. We define the noisy data as

$$\boldsymbol{x}_j = \boldsymbol{h}(t_j; \boldsymbol{\theta}_0) + \boldsymbol{e}_j, \quad j = 1, \ldots, n$$

where $\boldsymbol{e}$ is normally distributed with mean 0. An example of Data Set 1S with noise added is shown in Figure 14 as the dashed line.

Table 18: Estimated parameters using the MCR method for SIS CTMC model with **Noisy** Data Set 1S.

M	Param	True	Med. Est. (Error)	Mean Est. (Error)	95% CI
10	β	0.125	0.1180 (5.6%)	0.1160 (7.2%)	(0.1149, 0.1171)
	γ	0.1	0.0997 (0.3%)	0.0971 (2.9%)	(0.0962, 0.0981)
100	β	0.125	0.1182 (5.4%)	0.1187 (5.0%)	(0.1179, 0.1195)
	γ	0.1	0.0985 (1.5%)	0.0984 (1.6%)	(0.0978, 0.0990)

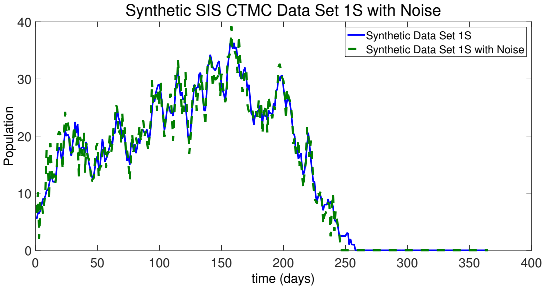

Figure 14: Synthetic Data Sets 1S with noise added.

As before, we seek to estimate β and γ using the MCR method (Algorithm 3). Since we know that a single selection of M realizations used in the cost function can result in an inaccurate estimate (Section 4.1), we go ahead and generate 1000 different trials using $M = 10$ and $M = 100$ realizations and calculate the median estimated parameter, mean estimated parameter and 95% confidence intervals. The results are given in Table 18. The results are exceptional. Finally, we generate 1000 different synthetic sets of noisy data from the same small population SIS CTMC model with the exact values for β and γ and use the MCR method to estimate the parameters. The confidence intervals are given in Table 19. Again, noisy data did not seem to effect the ability to accurately estimate parameters using the MCR method. In this example model, the MCR method seems to be robust against noisy data.

5. Conclusions and Future Work

In this paper, we first summarized and implemented a parameter estimation method for CTMC models developed by Ortiz et. al. [18] in which the CTMC model is first approximated by an appropriate deterministic model and then the parameter estimation is performed using this deterministic approximation.

Table 19: 95% Confidence intervals using MCR Method for SIS CTMC Model with Noisy Data.

n_s	Parameter	True Value	Confidence Interval	Max Rel. Error CI
10	β	0.125	(0.1155, 0.1178)	7.6%
	γ	0.1	(0.1016, 0.1049)	4.9%
100	β	0.125	(0.1173, 0.1192)	6.2%
	γ	0.1	(0.1001, 0.1038)	3.8%

Results were given for two example models of varying population sizes. In both examples, the deterministic approach worked well when the population size was 'large enough'; however, the results were unacceptable for models with small population sizes. Therefore, we modified this approach by developing the minimum cost realization or MCR method. The strengths of the MCR method lie in the fact that already established optimization algorithms can still be utilized while obtaining drastically improved parameter estimates, especially for small population models.

If one were modeling a physical system which requires a stochastic model, the data for the system will most definitely exhibit random perturbations. Therefore, in the MCR method, the data is 'compared' to realizations of the CTMC model which also contains random fluctuations. The basis for this method is that there is a 'best fit' realization of the CTMC model for the given data. Using this best fit, one can estimate the parameters which results in the lowest sum of squared differences. This method provided comparable results to the deterministic approach for both large population example models. However, the accuracy in parameter estimates was radically improved for the small population models with percent relative error below 10% in most cases. Furthermore, the method was shown to be robust when using synthetic data with noise added.

Upon further exploration of the MCR method, it was noted that the parameter estimation problem should be performed more than once if using randomly chosen realizations of the CTMC model, since the resulting estimate may be quite inaccurate in the rare extreme case in which none of the realizations are a good fit for the data. In a future paper, we plan to explore how many times the estimate needs to be performed to obtain acceptable results. We are also exploring ways in which one might *a priori* find a subset of realizations which better mimic the trend in the data and then use these in the MCR method; in this situation, the parameter estimate might only need to be performed once instead of multiple times. We also determined that increasing the number of realizations, M, used in the MCR method made a difference for one model while provided little improvement in another model when the estimates already had low relative error. Refining the method to *a priori* find suitable realizations for implementation might also change the effect of M. However, even with a small number of realizations used in the MCR method multiple times, the median parameter estimates were quite accurate for both example models. In summary, this method seems to be a plausible and an easily implementable method for parameter estimation in CTMC models.

References

[1] Linda Allen. *An introduction to stochastic processes with applications to biology.* Taylor and Francis Group, LLC, 2011.

[2] Hakan Andersson and Tom Britton. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer-Verlag, 2000.

[3] H. T. Banks, Shuhua Hu, and W Clayton Thompson. *Modeling and inverse problems in the presence of uncertainty.* CRC Press, 2014.

[4] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[5] Robert L. Borrelli and Courtney S. Coleman. *Differential equations: a modeling persepctive.* John Wiley and Sons, Inc., 2004.

[6] Richard J Boys, Darren J Wilkinson, and Thomas BL Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135, 2008.

[7] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[8] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[9] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[10] Andrew Golightly and Darren J Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005.

[11] Andrew Golightly and Darren J Wilkinson. Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13(3):838–851, 2006.

[12] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface focus*, page rsfs20110047, 2011.

[13] Thomas G Kurtz. Solutions of ordinary differential equations as limits of pure jump markov processes. *Journal of applied Probability*, 7(1):49–58, 1970.

[14] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[15] MATLAB. *version 8.1.0.604 (R2013a).* The MathWorks Inc., Natick, Massachusetts, 2013.

[16] Peter Milner, Colin S Gillespie, and Darren J Wilkinson. Moment closure based parameter inference of stochastic kinetic models. *Statistics and Computing*, 23(2):287–295, 2013.

[17] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, 2006.

[18] AR Ortiz, HT Banks, Carlos Castillo-Chavez, G Chowell, and X Wang. A deterministic methodology for estimation of parameters in dynamic markov chain models. *Journal of Biological Systems*, 19(01):71–100, 2011.

[19] Suresh Kumar Poovathingal and Rudiyanto Gunawan. Global parameter estimation methods for stochastic biochemical systems. *BMC bioinformatics*, 11(1):1, 2010.

[20] S Reinker, RM Altman, J Timmer, et al. Parameter estimation in stochastic biochemical reactions. *Systems biology*, 153(4):168, 2006.

[21] Yuanfeng Wang, Scott Christley, Eric Mjolsness, and Xiaohui Xie. Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent. *BMC Systems Biology*, 4(1), 2010.

[22] C Zimmer and S Sahle. Parameter estimation for stochastic models of biochemical reactions. *J Comput Sci Syst Biol*, 6:011–021, 2012.