Parameter Estimation for Individuals-based Models of Biochemical Reactions

LU SHAOKUI
Tianjin Polytechnic University
School of Science
No. 399 Bin Shui Xi Road, Tianjin
CHINA
Lshaokui@163.com

PEI YONGZHEN
Tianjin Polytechnic University
School of Computer Science and Software Engineering
No. 399 Bin Shui Xi Road, Tianjin
CHINA
yongzhenpei@163.com

LI CHANGGUO
Academy of Military Transportation
General course department
Dong Ju Zi Road no.1, Tianjin
CHINA
bayesmcmcli@sina.com

Abstract: Parameter estimation is crucial for us to analyse the models, and such works of individuals-based models is still in the early stage of development. For the individuals-based models, there is no efficient methods to estimate the parameters due to the observed data with noise produced by inherent randomness of model. This paper, we utilize different methods that are well developed for parameter estimation of determined model which is constituted by ordinary differential equations(ODE) are also adapted to stochastic models. In this article, We use the population changes of aphids as a case study. We want to estimate the birth rate and the mortality of the aphids. An intuitive approach is least square method to estimate the parameters, and this application is very extensive. However, the problem of parameter identification is the most common issue of least square method in estimating parameters. In this article we show the latest progress in parameter estimation for individuals-based models of our study which bases on moment closure approximation technique. The combination of MCMC and likelihood function is a less used method in the estimation of stochastic model parameters. These two methods can overcome the problem of parameter identification in the least square.

Key–Words: Parameter estimation, Individuals-based Models, Moment closure, Least squares method, Likelihood function, MCMC

1 Introduction

Research in recent years, mathematical models have become more and more important in the study of biochemical reaction systems. According to the different population level, these mathematical models can be divided into two major categories: population-level and individuals-based models. The first type of model, which we call deterministic models, is usually composed of a set of differential equations(ODE). Population-level models surreptitiously describe the dynamic behavior of the population with an infinite population size, which offers a general description of the population dynamics behavior. The most significant merit of these models is that we can analyze the dynamical behavior of these models by using some theories of ODEs. But they elide some results that may arouse by assuming the population scale is finite or by the inherently randomness of communication between individuals. Thus Matis et al. put forward a different model known as stochastic model and using it to describe the dynamics of the aphid population based on the parameters [1]. Parameter estimation is of great significance to the analysis of biological system model. This paper considers the problem of parameter estimation in the individual-based model. These models are dominated by a chemical master equation (CME) that is often difficult to solve. Consequently traditional methods of parameter estimation which depend on iterative of parameter likelihoods function are computationally intractable. To avoid this problem, we can approximate the likelihood function according to the Bayesian theorem which has many successful applications for parameters estimation [3]. This paper proposes recent advance methods in pa-

rameter estimation for individuals-based models and the parameter likelihood can be approximated by the moment closure.

The least squares technique (LSQ) is one of the most widely used method of the parameter estimation [2]. In this article, we use three methods to estimate parameters. Although the least squares method is a powerful tool in parameter estimation, the problem of parameter non-identifiability greatly limits the application of this method. Thus, we suggest a different program for parameter inference based on an expression that is obtained by approximating the likelihood function of the parameter through Bayesian theorem [4]. In order to get this expression, we need to solve the moment equations that are exhibited by ODEs which describe the time evolution of first moment (mean) and second moment (variance, covariance) even more the higher order moments of the interacting species. Then the solution of these equations can be obtained by moment closure techniques, and these results can be used to approximate calculation the likelihood function according to Bayesian theorem. Although the application of the Moment closure technique in population biology has a long history [5], and this method is rarely used in parameter estimation [6]. On the other hand, we use the MCMC method to estimate the parameter based on the likelihood of the parameters. Practice has been proved, Bayesian theorem for the individual-based models of the biochemical reaction systems is a very powerful tool. The basic method of Bayesian inference tries to integrate the prior information of model parameters with prior information, and then the posterior distribution of the parameter can be obtained according to the Bayesian theorem. Thus we can infer the model parameters with the posterior distribution [7]. The MCMC method can be described as a revolution in Bayesian statistics, which is a simple and effective Bayesian method of the calculation. The MCMC method is mainly composed of Metropolis-Hastings and Gibbs sampling algorithm [8]. The basic idea of this method is that through repeated sampling and establish a stationary distribution for the request of the posterior distribution of Markov chain, and the sample is obtained for the the posterior distribution, and then based on these samples to do all kinds of statistical inference, such as the estimated parameters of the mean, variance, and the correlation, etc

The paper is structured as follows. In the next section, we will introduce the individual-based model. Section 3 illustrates the methods of the parameter estimation and their results. A brief discussion is given in the finial section.

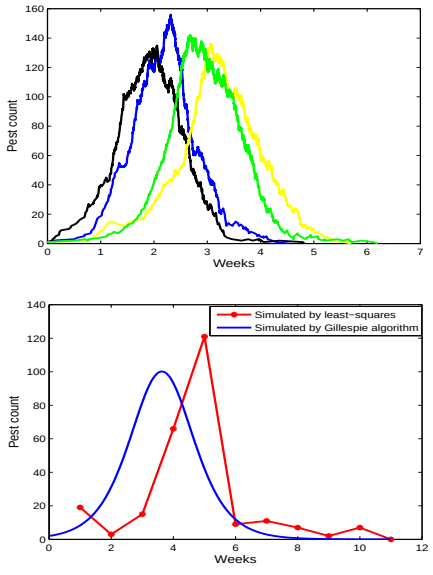
2 Model

2.1 Description of the model

For this paper we modeling including two stochastic processes $Z = \{\overline{N}, \overline{C}\}$. As discussed in [9], denote $\overline{N}(t)$ the number of the pest at current time. Assuming that $\lambda\overline{N}(t)$ represents pest population birth rate [10]. In particular, denote $\overline{C}(t)$ the environment deteriorated, up to time t, by the infestation. Let $\mu\overline{N}(t)\overline{C}(t)$ be mortality of the pest. and for simplicity, we ignore the condition of immigration and emigration. Modelling these two biochemical reactions as follow,

$$\begin{aligned} \overline{N} &\xrightarrow{\lambda} 2\overline{N} + \overline{C} \\ \overline{N} + \overline{C} &\xrightarrow{\eta} \overline{C} \end{aligned} \tag{1}$$

For (1), the first reaction means both $\overline{N}$ and $\overline{C}$ increasing one unit while the second reaction shows that $\overline{N}$ decreasing a unit whereas $\overline{C}$ is unchanged. These models are called a stochastic dynamical model in the literature [11]. Take parameter values $\alpha = 2.453$, $\eta = 0.0094$ and the initial values of the $\overline{N}$ and $\overline{C}$ are $\overline{N}(0) = 1, \overline{C}(0) = 1$. Simulations for the dynamic of the pest by Gillespie algorithm [12] are illustrated in Fig. 1(a)



(a)

(b)

Figure 1: (a) Time evolution of the pest population simulated by Gillespie algorithm [13]. (b) shows the Curve fitting by the least squares estimation.

By using the probabilistic laws, we can express the model considered in $(t, t + dt]$, as follow

$$Prob\{\overline{N}(t + dt) = n(t) + 1, \overline{C}(t + dt) = c(t) + 1|$$

$$n(t), c(t)\} = \lambda n(t)dt + o(dt), . \tag{2}$$

$$Prob\{\overline{N}(t+dt)=n(t)-1,\overline{C}(t+dt)=c(t)|$$
$$n(t),c(t)\}=\lambda n(t)dt+o(dt). \tag{3}$$

Where dt is enough small. Let $p_{n,c}(t)$ be the probability that there are n pests in the population at time t and a cumulative quality of c. In particular, the number of $\overline{C}$ is greater than $\overline{N}$. Then, the $p_{n,c}(t)$ can be obtained by solving the forward Komogorov equations. This is achieved by writing $p_{n,c}(t+\Delta t)$ as the sum of the probabilities of arriving in state (n,c) for small interval $(t,t+dt]$:

$$\begin{aligned} p_{n,c}(t+\Delta t)= &\ \lambda(n-1)p_{n-1,c-1}(t)\Delta t+\eta(n+1) \\ &\ cp_{n+1,c}(t)-n(\lambda+\eta c)p_{n,c}(t). \end{aligned} \tag{4}$$

Clearly, for (4) the probability of a birth event is the first term, and the second gives a probability of the death event. Removing from the state (n,c) at time t, and such events are occur in $(t,t+dt]$. Hence, from the (4) we obtain the forward Kolmogorov equation which is called the chemical master equation(CME):

$$\begin{aligned} \frac{dp_{n,c}(t)}{dt}= &\ \lambda(n-1)p_{n-1,c-1}(t)+\eta(n+1)cp_{n+1,c} \\ &\ (t)-n(\lambda+\eta c)p_{n,c}(t). \end{aligned} \tag{5}$$

The analytic solution of (5) is impossible to obtain. These two stochastic processes can be formulated as the moment equations of the system. Defining the bivariate moment generating function as

$$M(\theta_1,\theta_2,t)=\sum_{n,c=0}^{\infty}e^{n\theta_1+c\theta_2}p_{n,c}(t) \tag{6}$$

there is a relationship between the moment and cumulant generating function

$$K(\theta_1,\theta_2,t)=\log[M(\theta_1,\theta_2,t)]=\sum_{n,c=0}^{\infty}\kappa_{n,c}\theta_1\theta_2/i!j! \tag{7}$$

On multiplying equation (5) by $e^{n\theta_1}e^{c\theta_2}$, then we obtain

$$\begin{aligned} e^{n\theta_1}e^{c\theta_2}\frac{dp_{n,c}(t)}{dt}= &\lambda(n-1)e^{n\theta_1}e^{\theta_2}e^{(n-1)} \\ &e^{(c-1)\theta_2}p_{n-1,c-1}(t)+(\eta c(n+1) \\ &e^{-\theta_1}e^{(n+1)\theta_1}e^{c\theta_2}p_{n+1,c}(t)- \\ &n(\lambda+\eta c)e^{n\theta_1}e^{c\theta_2}p_{n,c}(t). \end{aligned} \tag{8}$$

We sum both sides of the (8) for n, c and then obtain

$$\frac{\partial M}{\partial t}=\lambda(e^{\theta_1+\theta_2}-1)\frac{\partial M}{\partial t}+\mu(e^{-\theta_1}-1)\frac{\partial^2 M}{\partial\theta_1\partial\theta_2} \tag{9}$$

From the (7) we can get the cumulant generating function as

$$\begin{aligned} \frac{\partial K}{\partial t} &= \lambda(e^{\theta_1+\theta_2}-1)\frac{\partial K}{\partial t}+\mu(e^{-\theta_1}-1)(\frac{\partial^2 K}{\partial\theta_1\partial\theta_2}+ \\ &\ \frac{\partial K}{\partial\theta_1}\frac{\partial K}{\partial\theta_2}) \end{aligned} \tag{10}$$

The ODEs for the cumulants κ_{10}, κ_{01}, κ_{02}, κ_{20} and κ_{11} can be obtained from (10). It seems worth noting here that κ_{10}, κ_{01} are the means of the $N(t)$ and $C(t)$ respectively, and κ_{20}, κ_{02} are the corresponding variance, and κ_{11} is covariance.

We now equate coefficients of θ_1, θ_2, θ_1^2, $\theta_1\theta_2$ and θ_2^2 on both sides of (10) to give the differential equations

$$\begin{cases} \dot{\kappa_{10}}=\lambda\kappa_{10}-\eta(\kappa_{10}\kappa_{01}+\kappa_{11}) \\ \dot{\kappa_{01}}=\lambda\kappa_{10} \\ \dot{\kappa_{20}}=\lambda(\kappa_{10}+2\kappa_{20})+\eta(\kappa_{11}-2\kappa_{10}\kappa_{11}-2\kappa_{21} \\ +\kappa_{01}(\kappa_{10}-2\kappa_{20})) \\ \dot{\kappa_{02}}=\lambda(\kappa_{10}+2\kappa_{11}) \\ \dot{\kappa_{11}}=\lambda(\kappa_{10}+\kappa_{20}+\kappa_{11})-\mu(\kappa_{10}\kappa_{02}+\kappa_{01}\kappa_{11} \\ +\kappa_{12}) \end{cases} \tag{11}$$

For simplicity, we set $\boldsymbol{\psi}=(\lambda,\eta)$. In order to more accurately describe the dynamics of pest populations, we need to estimate these two parameters.

3 Method

3.1 The least squares methods

The individuals-based model describes a continuous time Markov process. Parameter estimation of the individuals-based model can help us make the best prediction. As described in the second section, we use different methods to estimate parameters of the model (11).

For the parameter estimation, the LSQ is the most commonly used method. From the definition of absolute least-squares, we want to find the optimal parameters to minimize the objective function that is the sum of squares of the differences between the observed data y_i and predicted values y_i^* of the model. Therefore the absolute least-squares can be defined as

$$J=\sum_{i=1}^{n}[y_i-y_i^*]^2 \tag{12}$$

In our model, we can obtain the $\overline{N}(t)$ by the Gillespie algorithm which treated as the measured variables and the model-predicted data can be obtained by the (11). We use the absolute least-squares to estimate the parameters λ and η as follow:

Algorithm 1.

(i) Given the initial values that are $\overline{N}(0) = 1$, $\overline{C}(0) = 1$, $\lambda = 2.453$ *and* $\eta = 0.0094$

(ii) To simulate the evolution of $\overline{N}(t)$ *by solving the system (11), then we obtain the measured variables, denoted* y_i*. The first moment of* $\overline{N}(t)$ *can be obtained by solving equation (11), which we treat it as predicted values* y_i^* *of the model.*

(iii) Minimizing the (12) then we get the estimate values of λ *and* η*.*

We get the estimate values of the λ and η are $\lambda = 1.4648$, $\eta = 0.0072$. Although the least squares are a powerful tool for parameter estimation, The problem of parameter identifiability is the biggest bottleneck of least squares estimation (see Fig. 1(b)).

3.2 The maximum likelihood and MCMC methods

In order to circumvent this obstacle, we utilize maximum likelihood method to estimate the parameter of the system (11). As the previous introduced, for the model (1), taking the initial values $Z(0) = z(0)$ of species and the initial values $\lambda = 2.453$, $\eta = 0.0094$ of the parameters. The count of the Z at different time point can be obtained. Substituting these results into equation (5). We get the probability of the Z at different observation points of the time by solving the (5). We can calculate the likelihood function of the parameters as defined later by these probabilities. Specifically, giving the likelihood is

$$p(\boldsymbol{\psi}|Z) = \prod_{t=1}^{T} \prod_{n=1}^{N} \prod_{c=n}^{C} p(n, c, t|\boldsymbol{\psi}) \tag{13}$$

Where $p(\cdot, t|\boldsymbol{\psi})$ is the probability of Z at time t given that $\boldsymbol{\psi}$ is parameter of the model. N and C are the maximum values of $\overline{N}$ and $\overline{C}$. Maximizing the likelihood function can obtain the estimated value of the model parameters.

MCMC method is a another tool for parameter estimate of stochastic model [14]. Using the MCMC method to estimate parameter, we need to determined the prior distribution $p(\boldsymbol{\psi})$ of the parameters by the prior information. The MCMC method takes parameter as random variable. So the sample distribution family should be understood as conditional distribution, namely, whether it is a continuous or discrete random variable, it can be expressed as a conditional distribution that depends on the parameter. In this paper, it can be formulated as $p(Z|\boldsymbol{\psi})$. From the Bayesian point, the sample is produced in two steps. The first step is that we construct a prior distribution to generate a parameter. And the second step is that a sample is produced from the distribution of $p(Z|\boldsymbol{\psi}^{'})$ depend on the parameters obtained from the first step. For this step, it can be understood as to solve the model for the given parameter $\boldsymbol{\psi}^{'}$ and the solution of the model for the given time is the sample. The probability of sample occurrence is

$$p(Z|\psi^{'}) = \prod_{i=1} p(z_i|\psi^{'}) \tag{14}$$

According to the marginal distribution of the sample Z, from the Bayesian inference we have

$$p(\boldsymbol{\psi}|Z) = \frac{p(Z|\psi)p(\psi)}{p(Z)} \tag{15}$$

Where

$$P(Z) = \int_{\boldsymbol{\psi}} p(Z|\psi)p(\psi)d\psi \tag{16}$$

After given the prior distribution and the sample distribution, the posterior distribution of the parameter can be calculated according to the (15). Since all possible parameter values are independent of the marginal density function of the sample. So the equation (15) be transformed into

$$p(\psi|Z) \propto p(Z|\psi)p(\psi) \tag{17}$$

When the distribution of the sample is a likelihood function the equation can be reformulated as

$$p(\psi|Z) \propto L(Z|\psi)p(\psi) \tag{18}$$

Equation (18) is equivalent to formula (13). So the MCMC method for the parameter estimate as follows:

Algorithm 2.

(i) Given the initial values vector $\boldsymbol{\psi}^{(0)} = (\lambda^{(0)}, \eta^{(0)})$.

(ii) Given the initial time $t = 0$.

(iii) The cycle: From the Proposal distribution $q(\psi^{'}|\psi^{(t)})$ *draw the* $\psi^{'}$*; From the Uniform distribution* $U(0, 1)$ *draw the* u*; If* $u \leq \alpha(\psi^{'}, \psi^{(t)})$*, then* $\psi_i^{(t+1)} = \psi_i^{'}$ *accept this value. Else* $\psi_i^{(t+1)} = \psi_i^{(t)}$ *reject this value.*

(iv) $t = t + 1$;*. After a burn-in storage* $\psi^{(t+1)}$ *for the per cycle;*

(v) When t is sufficiently large, the loop ends;

This method is known in the literature as Random Walk Metropolis-Hastings algorithm. In terms of ease of implementation, the Metropolis-Hastings methods are the most favourable.

3.3 Determining acceptance probability

In the process of distribution selection, the basic idea is that the acceptance probability can be maintained at a certain level. For simplicity, we use symmetric propose distribution $q(\psi^{'} \mid \psi) = q(\psi \mid \psi^{'})$ [15]. So the acceptance probability can be determined by

$$\alpha(\psi^{'}, \psi) = min\left\{1, \frac{p(\psi^{'}|Z)}{p(\psi|Z)}\right\} \qquad (19)$$

From the calculation form of the (19), we conclude that the $p(\psi|Z)$ only appears in the form of quotient in the whole algorithm. Therefore, the complete form of $p(\psi|Z)$ in the entire calculation is not necessary. In particular, constant terms can be omitted. So in the calculation, we remove any non-parameter-dependent constant factors in the likelihood function, which simplify the calculation.

3.4 Determining burn-in

The 'burn-in' problem is the question of how much of a run should be thrown away on grounds that the chain may not yet have reached equilibrium. The length of burn-in m depends on ψ^0, on the rate of convergence of $q(\psi^{'}|\psi^{(t)})$ to the stationary distribution and on how similar between the proposal and stationary distributions are required to be. Theoretically, having specified a criterion of 'similar enough', m can be determined analytically. However, this calculation is far from computationally feasible in most situations. Starting the chain close to the mode of stationary distribution doesnot remove the need for a burn-in, as the chain should still be run long enough for it to 'forget' its starting position.

Calculating the the length of burn-in is unnecessary, as it is likely to be less than 1% of the total length of a run sufficiently long to obtain adequate precision in the estimator. It does not seem necessary to throw away many more iterations than the time it takes for the autocovariances to decay to a negligible level [16].

3.5 Performance the maximum likelihood and MCMC methods

We chose the same initial value as the least squares method that is $\lambda = 2.453$, $\eta = 0.0094$. The observed data can be obtained by the Gillespie algorithm. As followed the describe in the section 3.1, The parameters are estimated by these methods and the results are listed in Table 1.

Table 1: Estimated the model (1) parameters by lSQ, MLE, MLE-MCMC methods

	initial values	LSQ	MLE	MLE-MCMC
λ	2.453	1.4648	0.2514	0.2956
η	0.0094	0.0072	0.0075	0.0096

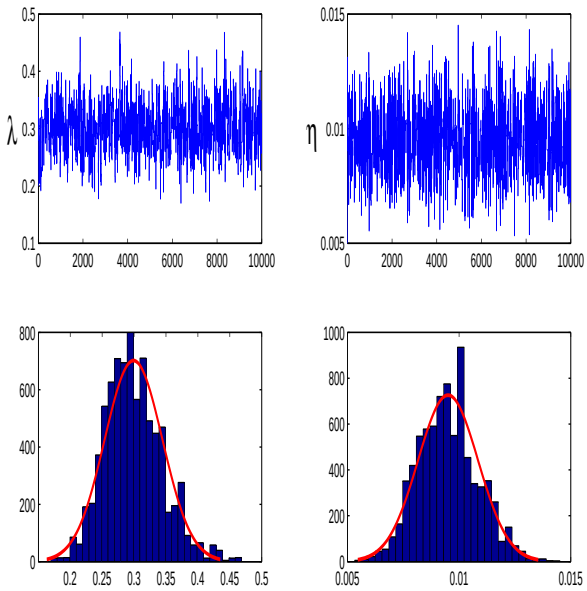


Figure 2: The panels delineate the distribution of the λ and η by MH algorithm.

In order to implement this algorithm, we set the burn-in period $m = 5000$ and sample size $M = 10000$. To illustrate the effectiveness of the MH algorithm, we estimate the parameters by using simulated data. The advantage of this is that we know the true value of the parameters and we can determine the validity of the algorithm by comparing the difference between the estimated value and the real value. In practice, we only know actual observation or experimental data, not the model. To realize the prediction function of the model, it is necessary to use the actual data to determine the parameters of the alternative model, that is, to determine the model. It turns out that the MH algorithm is also effective for model selection [17].

4 Discussion

The parameter estimation is the premise of the model analysis. The paper suggests three methods for parameter estimation in stochastic individual-based model for biochemical reaction systems, based on the moment equations. These methods are available for the parameters estimation, even though the process of individuals-based models is different from the deterministic model [18]. For the least square method, we need the observed data for the objective function (12). In this article, the data is synthesized by the Gillespie algorithm. The least-squares method relies heavily on system parameters.

In order to overcome the problem of parameter identifiability, we introduce maximum likelihood and the MLE-MCMC estimation. But we must acknowledge that it is very difficult to get the solution of a high dimensional CME when the system contains exceed three species. The solution of likelihood function is still the biggest bottleneck in the application of these two methods.

In this paper, the solution of likelihood function can be obtained. The accuracy of maximum likelihood estimation is higher than the least squares estimation [19]. From a mathematical aspect, how to improve the efficiency for the MLE-MCMC method is very interesting, and it would be left as a future work.

To conclude the paper, suggests different methods for parameter estimation of stochastic models. The last two methods need to calculate CME integration for a small time interval. It shows that these approaches are very practical for parameter estimation. As described in the introduction, it demonstrate these methods are also useful for parameter estimation of stochastic individual models of in biochemical reaction system. As a mathematical aspect, how to improve the efficiency of he algorithm as well as implement optimal control for the pest management is very promising. In the future work, we will consider this two problems.

Acknowledgements: The authors thank the referees for their careful reading of the original manuscript and many valuable comments and suggestions, which greatly improved the presentation of this paper. This work was supported by National Natural Science Foundation of China (11471243, 11501409, 11501410) .

References:

[1] J. H. Matis, T. R. Kiffe, T. I. Matis, D. E. Stevenson, Stochastic modeling of aphid population growth with nonlinear, power-law dynamics,*Mathematical Biosciences.* 208, 2007, pp. 469–494.

[2] Johnson, M. L. and Faunt, L. M, Parameter estimation by least-squares methods *Methods in Enzymology,* 210, 1992.

[3] Mckinley, Trevelyan and Cook, Alex R and Deardon, Robert, *International Journal of Biostatistics,* 5, 2009, pp. 24–24.

[4] Daigle, Bernie J and Min, K Roh and Petzold, Linda R and Niemi, Jarad, Accelerated maximum likelihood parameter estimation for stochastic biochemical systems, *Bmc Bioinformatics.* 13, 2012, pp. 1–68.

[5] Hespanha, Jo and Xe, Moment closure for biochemical networks *International Symposium on Communications, Control and Signal Processing* 2010, pp. 142–147.

[6] Milner, Peter and Gillespie, Colin S and Wilkinson, Darren J, Moment closure based parameter inference of stochastic kinetic models, *FStatistics and Computing.* 23, 2013, pp. 287–295.

[7] Diamond, G. A, Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem, *Medical Decision Making An International Journal of the Society for Medical Decision Making.* 12, 1992, pp. 1–22.

[8] Arminger, Gerhard and Muthn, Bengt O, A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the metropolis-hastings algorithm, *Psychometrika.* 63, 1998, pp. 271–300.

[9] Prajneshu, A nonlinear statistical model for aphid population growth, *Journal of the Indian Society of Agricultural Statistic.* 51, 1998.

[10] Matis, J. H. and Kiffe, T. R. and Matis, T. I. and Stevenson, D. E, Application of population growth models based on cumulative size to pecan aphids, *Journal of Agricultural Biological and Environmental Statistics.* 11, 2006, pp. 425–449.

[11] Koblents, Eugenia and Młguez, Joaqułn, A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models, *Statistics and Computing.* 25, 2015, pp. 407–425.

[12] Golightly, A and Gillespie, C. S, Simulation of stochastic kinetic models, *Methods Mol Biol.* 1021, 2013, pp. 169–187.

[13] Daniel T. Gillespie, Exact Stochastic Simulation of Coupled Chemical Reactions, *The Journal of chemical physics.* 126, 2007.

[14] Mauro Gasparini, Markov Chain Monte Carlo in Practice, *Technometrics.* 2, 1999, pp. 9236–9240.

[15] Liu, Jun S, Metropolized Independent Sampling with Comparisons to Rejection Sampling and Importance Sampling, *Statistics and Computing*. 6, 1996, pp. 113–119.

[16] Geyer, Charles J, Practical Markov Chain Monte Carlo *Statistical Science.* 7, 1992, pp. 473–483.

[17] Green, P, Reversible jump MCMC computation and Bayesian model determination *Biometrika.* 82, 1995, pp. 104–111.

[18] Kummer, U and Krajnc, B and Pahle, J and Green, A. K. and Dixon, C. J. and Marhl, M, Transition from stochastic to deterministic behavior in calcium oscillations *Wiley.* 2013.

[19] Wang, Sichun and Jackson, Brad R. and Inkol, Robert, Hybrid RSS/AOA emitter location estimation based on least squares and maximum likelihood criteria *Communications.* 2013, pp. 24–29.