Прилози, Одд. маш. шех. науки, МАНУ, **XXXI**, 1–2 (2010), с. 73–98
Contributions, Sec. Math. Tech. Sci., MANU, **XXXI**, 1–2 (2010), p. 73–98
ISSN 0351–3246
UDC: 517.9 : 57.01

APPROXIMATE BAYESIAN PARAMETER INFERENCE FOR DYNAMICAL SYSTEMS IN SYSTEMS BIOLOGY

Jovan Tanevski, Sašo Džeroski, Ljupčo Kocarev

Abstract: This paper proposes to use approximate instead of exact stochastic simulation algorithms for approximate Bayesian parameter inference of dynamical systems in systems biology. It first presents the mathematical framework for the description of systems biology models, especially from the aspect of a stochastic formulation as opposed to deterministic model formulations based on the law of mass action. In contrast to maximum likelihood methods for parameter inference, approximate inference method-sare presented which are based on sampling parameters from a known prior probability distribution, which gradually evolves tward a posterior distribution, through the comparison of simulated data from the model to a given data set of measurements. The paper then discusses the simulation process, where an overview is given of the different exact and approximate methods for stochastic simulation and their improvements that we propose. The exact and approximate simulators are implemented and used within approximate Bayesian parameter inference methods. Our evaluation of these methods on two tasks of parameter estimation in two different models shows that equally good results are obtained much faster when using approximate simulation as compared to using exact simulation.

Key words: systems biology, reaction kinetics, stochastic models, exact stochastic simulation, approximate stochastic simulation, approximate parameter inference

1. INTRODUCTION

Building mathematical models is needed for the analysis and better understanding of the behavior of dynamical systems. This includes observation and measurement of the behavior of the dynamical system under different conditions, choosing a set of variables that describe the system, and creating a mathematical description of the model. After an adequate model has been chosen for a certain dynamical system, an appropriate set of parameters has to be inferred. After a set of appropriate parameters has been chosen, a simulation of the proposed model is performed for comparison with existing experimental data and establishing the correctness of the model.

In systems biology, a model is described as a complex network of chemical reactions driven by known kinetic laws. If the initial conditions of the system are known, a simulation can be made by evolving the system through time.

The temporal evolution of the biological systems is traditionally considered as a deterministic process with known dynamical behavior described by the law of mass action, which can be described by a system of ordinary differential equations. The deterministic approach to modeling dynamical systems is inadequate for the description of cellular interactions [1] as it can be used only for systems with a large number of molecules, where the noise at the molecular level has no macroscopic effect. In the more general case, this approach does not allow for a complete and physically correct representation of the basic stochastic processes that take place in a living cell. On the other hand, the discrete and stochastic evolution takes into account the discrete number of entities in the system and the random nature of the events taking place, drawing nearer to the theories of thermodynamics and stochastic processes [2].

In this paper we considerapproximate Bayesian methods for parameter inference in dynamical models from the area of systems biology. These methods, based on a sequential Monte Carlo technique, are being used for determining a posterior probability distribution over a space of parameter values. The approximate Bayesian approach can be used for a range of modeling methods without any significant modifications. Given a prior parameter distribution I can determine an appropriate posterior parameter distribution from incomplete or partially observable data. By using approximate methods, the determination of a likelihood function based on experimental data is replaced by a simulation based procedure.

The remainder of the paper is organized as follows. Section 2 presents the stochastic formulation of the problem of modeling reaction dynamics in the domain of systems biology. In Section 3, we give an overview of exact and approximate methods for stochastic simulation of reaction dynamics. We then discuss (in Section 4) the approximate Bayesian framework for parameter estimation, which determines a proper/posterior probability distribution over the parameter value space, starting from a prior distribution and taking into account observed data. The approximate Bayesian framework based on a sequential Monte Carlo approach performs many stochastic simulations: We propose to use approximate stochastic simulation instead of exact simulation. Section 5 compares the use of approximate and exact simulation in the context of parameter estimation in two models of different dynamical systems. Section 6 concludes and indicates a major direction for further work.

2. MODELING REACTION DYNAMICS

The deterministic approach to modeling reaction dynamics considers the temporal evolution as a continuous process with known behavior, described by a system of ordinary differential equations (one for each entity present in the system), called reaction rate equations. The deterministic approach presumes that, for a sufficiently large number of molecules the stoichiometric changes that take place in the system as a result of a single reaction are negligible and the change of the concentrations of the entities in the system is continuous. In this way, the small changes in the behavior of the system are approximated by the average behavior. Namely, the small changes in the molecular population are a result of an occasional firing of a certain reaction that introduces negligible changes in the macroscopic trend of the concentrations and the average behavior can be a good overall approximation of the evolution of the system [3]. In addition, the system is in a thermodynamic equilibrium and there are no changes in the temperature or the space in which reactions take place.

In contrast to the deterministic mathematical formulation for which only one possible evolution through time exists the stochastic formulation introduces uncertainty in the evolution. This uncertainty is described by a probability distribution. As a result, even if the initial conditions are known, some trajectories are more probable than other. By exploring a sufficiently large number of possible evolutions, given a set of final states, the system will reach some states more frequently than other.

In the stochastic formulation of modeling reaction dynamics, a system consists of a system of N molecular entities $\{S_1, .., S_N\}$ that participate in M chemical reactions $\{R_1, .., R_M\}$. The entities are well stirred in a reaction space with volume V and are in a state of thermal, but not chemical, equilibrium. A large number of collisions that take place in this setting are elastic (nonreactive). This behavior results in a uniform spatial distribution of the molecules in the reaction volume V and a distribution of the molecular speeds approximate to the Maxwell-Boltzmann distribution [4]. Given a described system, the non-reactive collisions are ignored and only the events that result in change of the concentrations of the molecules are considered.

The state of the system can be described as a vector of number of molecules $X_i(t)$ for each entity at a given time t, $X(t) \equiv (X_1(t), \ldots, X_N(t))$. The evolution trough time is a result of reactions fired. Every reaction is considered a separate, instantaneous, elementary and random physical event. Every reaction R_j is characterized by a propensity function $\mathrm{a_j}$ and a vector of state changes $v_j \equiv (v_{1j}, \ldots, v_{Nj})$. Let $X(t) = x$. $a_j(x)dt$ then represents the probability that the reaction R_j will occur in the next differential time period $[t, t + dt)$, and v_{ij} represent the change in the molecular population of S_i as a result of the reaction R_j being fired.

The probability of a reaction happening, for the purposes of the stochastic formulation, must include a constant c_j which depends on the physical characteristics of the molecules and the current temperature. Multiplying the probability c_jdt with the total number of different combinations of reactants of reaction R_j in V at time t, gives the probability $P_j(dt)$ of the reaction R_j happening in the interval $[t, t + dt)$, if the system is in state X at time t. That is

$$P_j(dt) = a_j dt = c_j h_j dt\,.$$

Gillespie [5] presents the physical principle of propensity in reactions. If R_j is an unimolecular reaction in the form $S_1 \xrightarrow{c_j} \text{Products},$ there exists a constant c_j, so that c_jdt is the probability that a molecule from S_1 will participate in a reaction at the next differential time dt. The laws of probability state that if there currently are x_1 molecules of the entity S_1 present in the system, the probability that any one of them will participate in a reaction at the next dt is x_1c_jdt. The propensity then can be calculated as $a_j(x) = c_jx_1$. If R_j is a bimolecular reaction in the form $S_1 + S_2 \xrightarrow{c_j} \text{Products},$ there exists a constant c_j so that c_jdt is the probability that a random pair of molecules from S_1 and S_2 respectively will participate in a reaction at the next differential time dt. The probability that any of the x_1x_2 pairs of molecules will come into reaction R_j at the next differential time dt is $x_1x_2c_jdt$. In this case, $a_j(x) = x_1x_2c_jdt$ is the propensity function. If the bimolecular reaction is in the form, the number of different pairs is $x_1(x_1 - 1)/2$ and the propensity is $a_j(x) = (x_1(x_1 - 1)/2)c_jdt$ and so on. Given a termolecular reaction the propensity will depend on the number of triplets that can be formed. Termolecular reactions are very rare because of the low probability of simultaneous collision of three molecules, and are modeled as a sequence of bimolecular reactions.

3. STOCHASTIC SIMULATION METHODS

3.1. Exact methods

Gillespie proposes a Monte Carlo technique, known as the stochastic simulation algorithm, for generating solutions for $X(t)$ [6]. The numerical simulation of the evolution through time is based on answering two crucial questions at every step: (1) When will the next reaction occur? (2) Which reaction will be executed? The algorithm proposed by Gillespie represents an algorithm for exact stochastic simulation of a system based on a probability density function of the reactions.

One simulation for times t_1 through t_{stop} results in only one possible realization. Several independent realizations are needed for a good estimate of the average concentrations of the entity S_i at time t.

Gillespie originally proposes two exact methods for stochastic simulation. These arethe direct method and the first reaction method. Although these methods are most widely used, another frequently used method is the next reaction method of Gibson and Bruck [7] for the simulation of models with many species and many channels.

In the direct method, to generate a pair of a time and a reaction index, two independent random values r_1 and r_2 are sampled from theuniformly distributed interval $U(0, 1)$. Using these random values the time is calculated as

$$\tau = \frac{1}{a}\ln\left(\frac{1}{r_1}\right),$$

and the index of the most probable reaction happening in the interval $(t+\tau,$ $t+\tau+d\tau)$ is the smallest integer such that

$$\sum_{v=1}^{\mu-1} a_v < r_2 a \leq \sum_{v=1}^{\mu} a_v .$$

For the first reaction method, tentative reaction times are generated for each reaction. The method then takes the smallest τ_v. The index v for which τ_v is the smallest is taken as a parameter for execution of every simulation step.

In theory the asymptotic complexity of both methods coincides. In practice the direct method performs better, this is due to the smaller number of random numbers being drawn in each step.

The method of next reaction is an improvement of these methods and outperforms them when applied to systems with a large number of reactions. This method is a popular and efficient implementation of the method of first reaction. The method uses an indexed binary tree as a priority queue *P* to find the next reaction and the time of its occurrence. It also implements a directed graph *G*, calculating only propensities and reaction times that are being influenced by the selected reaction.

As opposed to the exact methods, the approximate stochastic simulation methods trade a certain amount of precision to speed up the simulation process. The τ-leaping method and its improvements are the most widely used approximate methods.

3.2. Approximate methods

The first τ-leaping method was proposed by Gillespie et al. [8], where the authors consider a leaping condition in the propensity functions. If there exists a time period t in which the propensities a_j are almost constant, then the number of occurrences of a reaction R_j in the time interval $[t, t+\tau)$ can be approximated with a Poisson random variable. Instead of calculating reaction times at every step, this method selects the largest time τ for which the leaping condition is being met and generates, for every reaction R_j, a random sample with Poisson distribution $k_j = \mathcal{P}_j(a_j, \tau)$ and updates the system according to the formula.

$$x(t+\tau)=x(t)+\sum_{j=1}^{M}k_j v_j$$

A problem with the τ-leaping method appears while generating a random Poisson variable. The Poisson approximation of k_j may result in firing reaction R_j so many times that the number of molecules of a reactant becomesinsufficient and its population becomes negative.

Tian et Burrage [9] and in another independent research Chatterjee et al. [10] propose a way of addressing the negative population problem by approximating k_j with a binomial random variable and defining an upper bound of the distribution that will not allow the selection of large values for k_j.

Cao et al. [11] on the other hand, propose an approach that uses the original Poisson random variables and avoids the negative population problem. A negative population may appear only in cases where the number of molecules of a certain reactant is sufficiently small, so that the reactions are split in two groups: a group of critical reactions, which may result in negative populations, and a different group of noncritical reactions, with a smaller probability of negative populations appearing upon firing. This separation of the reactions in two groups allows for simulating the critical group with the standard exact methods and simulating the reactions in the noncritical group by τ-leaping.

Taking into consideration a real system, the change of the number of molecules for each of the participants in the reaction can be quick at first, especially in systems with a small number of molecules, and then continue to evolve slowly. The original τ-leaping method will speed up the process of simulation for systems with a large number of molecules which reach a stable state very quickly, but reaching the same effect with the original methodon a more general chemical (biological) system is a matter of discussion.

An implicit version of the τ-leaping method is proposed for the solution of stiff models [12]. The choice of explicit or implicit τ is being made based on previous knowledge of the behavior of the system. Having considered the original explicit and the implicit version of τ-leaping, an adaptive method is proposed for the simulation of general biochemical systems [13]. The adaptive method automatically chooses one of the two proposed methods depending on the stiffness of the system during simulation.

In the trivial case where none of the propensity functions depend on the current state of the system, the leaping condition will be met for every τ. Every reaction introduces a small change in the number of molecules present in the system, so in a system in which large numbers of molecules appear as reactants, large numbers of reactions need to be fired to the effect of the propensity functions being significantly changed. So, for systems with a large number of participating molecules, where the exact methods will be slower in their execution, it can be said that the leaping condition will be met for a time τ that will allow a large number of reactions to be fired in the interval $[t, t+\tau)$. If a relatively small value for τ is selected ($\tau = 1/a$ or smaller), the resulting time for the leap will match the time selected using exact methods, which in turn is an inefficient use of the τleaping method. On the other hand, selecting the largest possible value for the leap time speeds up the simulation process, so the largest τthat meets the leaping condition has to be selected in computationally efficient way. The formulas for τ-selection, both explicit and implicit, are given by the following formulas

$$\tau^{(ex)} = \min_{i\in I_{rs}} \left\{ \frac{\varepsilon a}{|\mu_i(x)|}, \frac{(\varepsilon a)^2}{[\sigma_i(x)]^2} \right\},$$

$$\tau^{(im)} = \min_{i\in I_{rs}} \left\{ \frac{\max\left\{\frac{\varepsilon x_i}{g_i}, 1\right\}}{|\mu_i(x)|}, \frac{\max\left\{\frac{\varepsilon x_i}{g_i}, 1\right\}^2}{[\sigma_i(x)]^2} \right\}.$$

Here, I_{rs} is the set of all reactions, the expected change of the propensities is bounded by ε ($0 < \varepsilon < 1$), and g_i is given by a formula which guarantees that bounding the relative change of states is sufficient for bounding the relative change of propensity functions [13]. In both cases μ_i and σ_i are calculated using the following formulae:

$$\mu_i(x) \triangleq \sum_{j\in J_{necr}} v_{ij} a_i(x), \qquad \forall i \in I_{rs},$$

$$[\sigma_i(x)]^2 \triangleq \sum_{j\in J_{necr}} v_{ij}^2 a_i(x), \qquad \forall i \in I_{rs}.$$

Here, J_{necr} is the subset of reactions that are neither critical nor in a partial equilibrium. The implicit $\tau^{(\mathrm{im})}$ is chosen when its value is larger than $N_{\mathrm{stiff}}\,\tau^{(\mathrm{ex})}$ and the system is ruled to be stiff. The N_{stiff} parameter is usually chosen to be 100.

4. APPROXIMATE BAYESIAN PARAMETER ESTIMATION

The methods used for parameter inference on complex multidimensional systems are based on a group of Monte Carlo algorithms based of the knowledge of a likelihood function $P(D|\theta)$. While the methods based on a maximum likelihood approach tend to infer the parameter vector by maximizing the likelihood function $\hat{\theta} = \arg\max_{\theta} P(D,\theta)$ [14] [15], the Bayesian approachesgiven an prior distribution $\pi(\theta)$ give as a result a posterior distribution of the parameters $\pi(\theta \mid D)$ such that $\pi(\theta \mid D) = P(\mathrm{D}|\theta)\pi(\theta)$ [16], [17]. If the likelihood isn't known, the use of these approaches is impossible and an alternative approach is needed for the process of parameter inference which isn't based on the likelihood function. These methods are known as approximate Bayesian computational methods. These methods are being used more frequently in genetics [18], [19], epidemiology and ecology where they have proven to be useful [20]. One of the main advantages of these methods is the ability to be used with different kind of models, deterministic of stochastic. In this paper, an analysis of the application of these methods is made, using the exact and approximate stochastic simulation methods described.

Let D be a set of discrete data, which is generated from a model M using a vector of parameters θ with a prior distribution π (θ). The posterior distribution of the parameters after observing the given data D can be obtained by using the formula

$$\pi(\theta \mid D) = \frac{P(D \mid \theta)\pi(\theta)}{P(D)},$$

in which $P(D) = \int P(D \mid \theta)\pi(\theta)d\theta$ is evidence in favor of the model M or marginal likelihood. The analytical solution of $\pi(\theta \mid D)$ may not be given for most of the models if the marginal likelihood cannot be computed by integration. As an alternative, it is possible to apply a numerical solution to the problem using Monte Carlo techniques developed for this kind of problems. If we are unable to calculate the marginal likelihood, the likelihood term $P(D \mid \theta)$ cannot be calculated either, but simulated data from the model can still be obtained.

The basic approximate Bayesian algorithm for parameter inference is the algorithm proposed by Pritchard [18] which in turn is based on the rejection method. The rejection method is based on simulating data from a model and does not depend on the likelihood function. Relaxing the condition in which simulated data *D'* are accepted only when it is identical to the given set of real data *D* and by introducing a measure of closeness between the real and simulated data (ε) we get the first approximate Bayesian method based on the rejection method for parameter estimation. The methodis executed by going through the following steps:

1. Sample parameters θ from the given prior distributions $\pi(\cdot)$;
2. Simulate new data set *D'* using the sampled parameters;
3. Accept the parameters θ if $d(D, D') \leq \Sigma$,

where the function d is a distance metric on the state space.

This basic method has been extended in various ways. The Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling a probability distribution by constructing a Markov Chain whose stable distribution conforms to the desired posterior distribution. The most widely known MCMC algorithms are the Metropolis-Hastings [21] [22] sampler and the Gibbs sample [23], the latter being a special case of the first one.

The MCMC methods are not Bayesian in nature, but are frequently used in applications that use the Bayesian approach. These methods maximize the statistical efficiency of every the Monte Carlo estimation. Although the marginal likelihood does not have to be given when using these methods, the likelihood function is needed.

As mentioned before, the likelihood functions of some models cannot be computed, so these algorithms are not a good solution for the problem. Marjoram et. al. [24] proposed an algorithm that can be a solution of this problem. The main difference between this and the previous approaches is the use of simulated data instead of calculation of the likelihood function.

An approximate version of this algorithm uses a tolerance ε of the distance between the simulated data and the real data analogous to the basic method. The result of this algorithm is a Markov Chain, which contains the posterior distribution of the parameter space and will always converge towards a solution. The main disadvantage of this method is the relatively small acceptance rate which may result in long chains. Also, given the nature of the process, the algorithm may get stuck in regions of low probability for a long time. Apartial solution of this problem is the use of sequential Monte Carlo methods that use the obtained approximated distributions as marginal probabilities of a larger family of distributions through which the sampler can easily traverse.

The sequential Monte Carlo simulation based method with partial rejection control can be used as an alternative solution [25]. A population of parameters $\theta^{(1)}\ldots\ \theta^{(n)}$ is propagated from an initially defined parameter distribution, through intermediate distributions until the population becomes a sample from the required posterior distribution. Because the approach includes a whole population of parameter distributions, this method allows analysis of systems with parameters that may have complex (multimodal) distributions. The distributions in the population that does not correspond to the required posterior distribution are being easily rejected in contrast to the suitable ones by allocating importance weights.

The rejection process can also be realized by introducing a gradient of tolerances for the populations. In every step, every particle foregoes a perturbation by a Markov kernel with the goal of improving the particle dispersion. This kernel can be a standard Gaussian kernel or a Metropolis-Hastings acceptance step. The kernel is also used for calculating the importance weights. Because there is a possibility of degeneration of a certain population, a resampling step is introduced to remove from the population unwanted particles that have small importance weight and normalize others.

While running the algorithm, there is a possibility of biased posterior samples being produced by approximating the weights with two unbiased Monte Carlo scores. The solution is implementing a backwards kernel for obtaining an unbiased posterior probability. This approach is being used in the Monte Carlo method [26] that implements a special case ofthe sequential importance sampling algorithm by Del Moral [27].

The algorithm for the approximate Bayesian method, based on sequential Monte Carlo, proceeds as follows:

```
1. Initialize $\varepsilon_1, \ldots, \varepsilon_T$. Set the population indicator $t = 0$.
2. Set the particle indicator $i = 1$.
2.1. If $t = 1$independently sample the prior distribution $x^{**} \sim \pi(x)$.
     If $t > 1$ sample $x^*$ from the previous population $\{x_{t-1}^{(i)}\}$ with
     weights $\{w_{t-1}^{(i)}\}$ and pertur $b$ the particle to obtain $x^{**} \sim K_t(x|x^*)$,
     where $K_t$ is a perturbation kernel.
     If $\pi(x^{**}) = 0$ return to 2.1
     Simulate a candidate dataset $D_{(b)}(x^{**}) \sim f(D|x^{**})$,
     $B_t$times ($b = 1, \ldots, \mathrm{B}_t$) and calculate $b_t(x^{**})$.
     If $b_t(x^{**}) = 0$, return to 2.1
     IF $d(D_{(b)}(x^{**}), D_0) \geq \varepsilon_t$return to 2.1
```

2.2. Set $x_t^{(i)} = x^{**}$ and calculate the weight for particle $x_t^{(i)}$

$$w_t^{(i)} = \begin{cases} b_t(x_\tau^{(i)}), & \text{if } t = 0 \\ \dfrac{\pi(x_\tau^{(i)}) b_t(x_\tau^{(i)})}{\sum_{j=1}^{N} w_{\tau-1}^{(j)} K_\tau\left(x_{\tau-1}^{(j)}, x_\tau^{(i)}\right)} & \text{if } t > 0 \end{cases}$$

If $i < N$, set $i = i + 1$, go to 2.1

3. Normalize the weights. If $t < T$, set $t = t + 1$, go to 2.

It is important to mention that this algorithm will follow the defined steps when estimating the parameters of a stochastic model. If the model is deterministic, we have $B_t = 1$ and the algorithm is much simpler.

Within Step 2.1 of the algorithm, exact stochastic simulation is typically used. In this paper we propose the use of approximate stochastic simulation. In the next section, we compare the two alternatives.

5. EVALUATION

The evaluation of the precision of the stochastic simulator can be made only probabilistically by simulating a model a large number of times and examining the distribution of the results of the simulation. A group of models from the discrete stochastic models test suite [28] are used for the evaluation. Every model used for testing contains a description in SBML [29] format, as well as the expected mean values and acceptable standard deviation of the obtained concentrations of the entities in the model (given a regular time frame). Our results on the models tested obtained by using an average values from 10000 independent simulations for each test modelshow values within the acceptable intervals of deviation given by the authors.

A modified version of the ABC-SysBio package is then used for approximate Bayesian parameter inference [30]. The exact and approximate stochastic methods for simulation previously discussed namely the direct method and the adaptive explicit-implicit tau leaping method, are implemented and tightly integrated within the package for the purpose of evaluating the precision of the inferred posterior probabilities of the parameters.

The approximate Bayesian framework based on the sequential Monte Carlo method is being used with the possibility of defining a number of simulations being made for each particle of the population. The role of the parameter B_t is investigated given the inferred posterior distributions of the parameters of the system. Another issue we consider is the influence of the number of sampled particles within a population on the inference of a required posterior distribution. The results obtained using exact and approximate simulations are given in parallel for two models in each of the following subsections.

5.1. Dimerization model

Wilkinson's dimerization model [31] is a basic model of a biochemical process having an important regulatory role in a large number of biological proceses. A description of the model in a discrete and stochastic formulation is made for compatibility with the simulators. The model consists of two species P and $P2$ which take part in two reactions. The first is a reaction of dimerization $2P \rightarrow P2$, with a rate $c1$, in which two moleculs from the species P form a dimer $P2$. The second is a reaction of dissociation $P2 \rightarrow 2P$, with a rate $c2$, in which a molecule from the dimer $P2$ is converted into two molecules of species P. From this model with parameters $(c1, c2) = (0.00166, 0.2)$, a synthetic dataset is generated. To each point of the dataset, Gaussian noise from the distribution $N(0, (0.5)^2)$ is added. The dataset, created under thegiven conditions represents a set of measurements from a process of dimerization.

For the estimation of the parameters of the model of dimerization and the evaluation of the number of simulations of each particle in the population, the following prior distributions of the parameters are used $\pi(c1) \sim U(0, 0.05)$, $\pi(c2) \sim U(0, 0.5)$. The initial conditions are $(P, P2) = (301, 0)$. The process of parameter inference goes through the following tolerance gradient $\varepsilon = (80, 70, 60, 50, 45)$. The distance metric, being used is the square root of the sum of squared distances. The Perturbation kernels are uniformly distributed according to $K_t = \sigma U(-1, 1)$, where $\sigma_{c1} = 0.002$, $\sigma_{c2} = 0.2$. For each population, 100 accepted particles are considered..

Figure 1 and 2 shows the results of performing only one simulation per particle using the exact and approximate simulations.

The posterior distributions obtained by using the exact and the approximate simulation are similar. Considering the amplitudes of the posterior distribution, we can conclude that a good approximation of the parameters is being made.

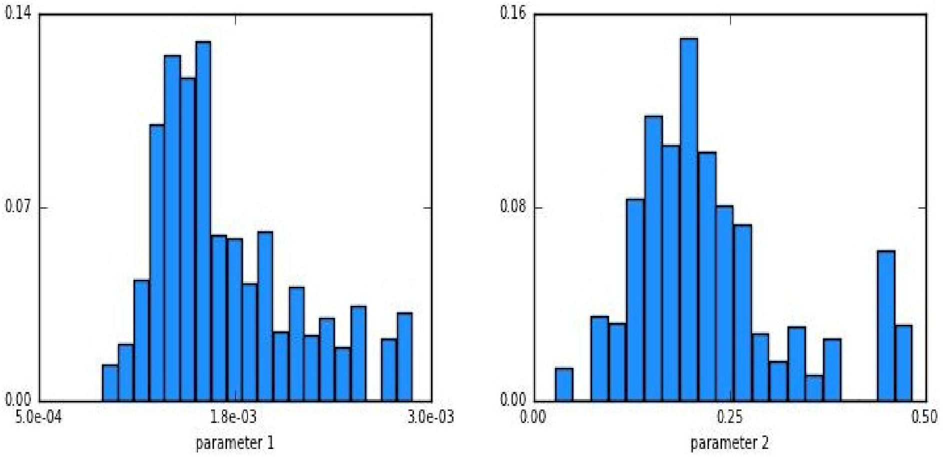


Fig. 1. The posterior distributions of the parameters of the dimerization model using exact simulation ($B_t = 1$)

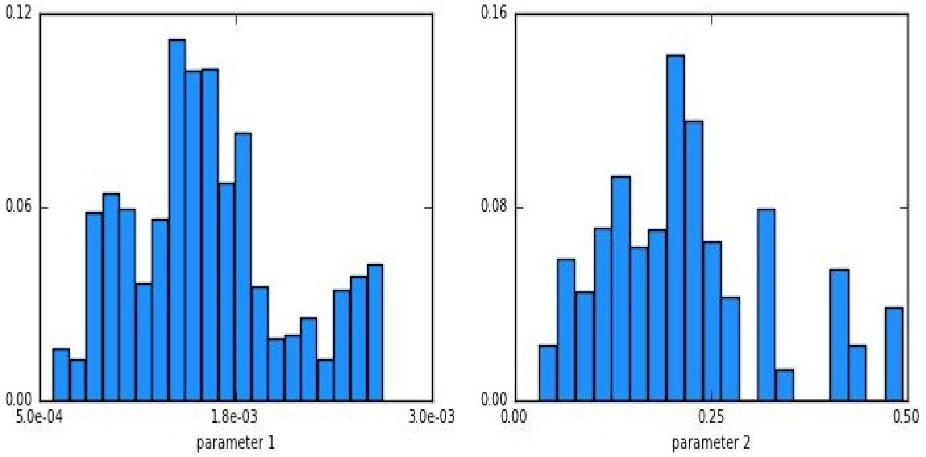


Fig. 2. The posterior distributions of the parameters of the dimerization model using approximate simulation ($B_\mathrm{t} = 1$)

In the next step, we increase the number of simulations per particle to $B_t = 10$. The posterior distributions of the parameters are shown in Figures 3 and 4.

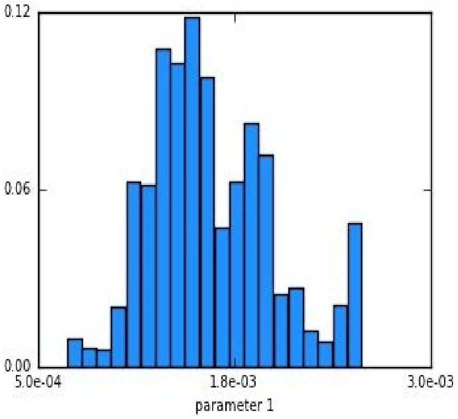


Fig. 3. The posterior distributions of the parameters of the dimerization model using exact simulation ($B_t = 10$)

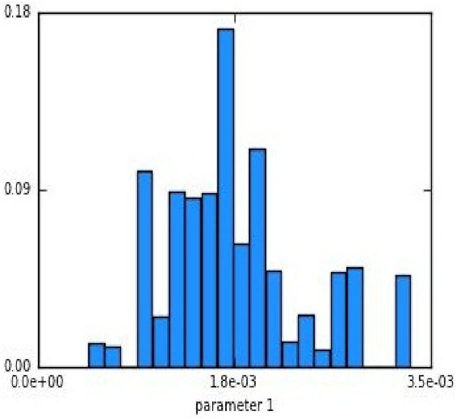
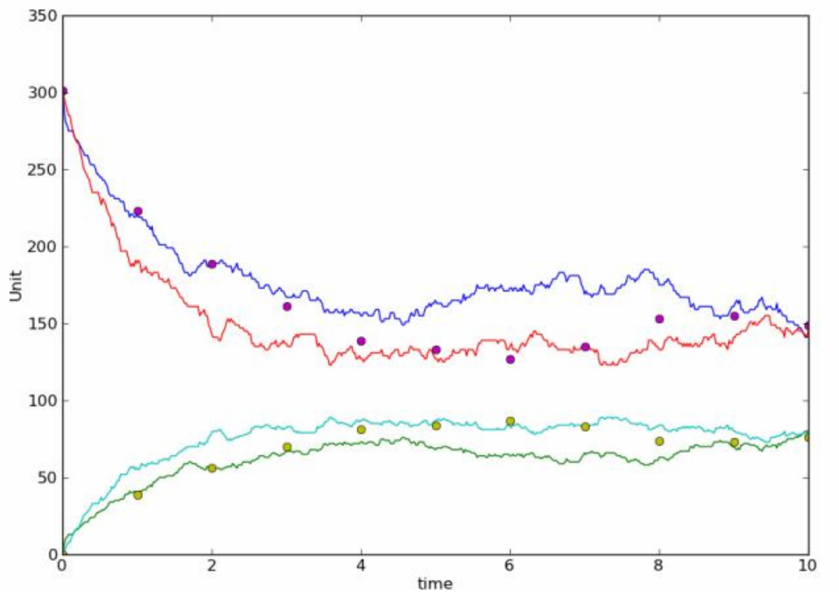


Fig. 4. The posterior distributions of the parameters of the dimerization model using approximate simulation ($B_t = 10$)

Looking at the posterior distributions, clearly differentiated regions are noticeable where the exact required parameter values should lie. Same as before, the posterior distributions obtained using exact and approximate simulation are similar and a good approximation of the parameters is being made. In addition we can conclude that the precision of the inference depends on the number of simulations per particle. This can also be seen in Figure 5 where the evolution of the concentrations of the species is shown with respect to the inferred parameters.



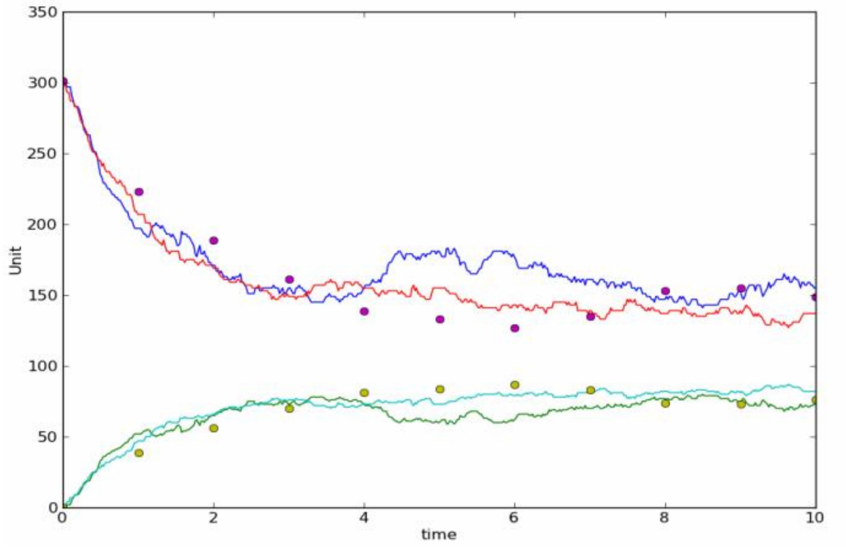


Fig. 5. Evolution of the concentrations of the species with inferred parameters. (B_t = 1 (upper graph) and B_t = 10 (lower graph)). The dots represent the values from the synthetic dataset. In the region between the solid lines are the data generated by simulationsusing the inferred parameters

We next consider a larger number of accepted particles per population in the parameter inference process for the dimerization model.

Parameter inference using B_t = 10 and P = 1000 is performed. The results can be seen in Figures 6 and 7. An improvement in the distribution of the parameter values can be seen and the distributions now resemble a Gaussian distribution with mean value close to the original parameters.

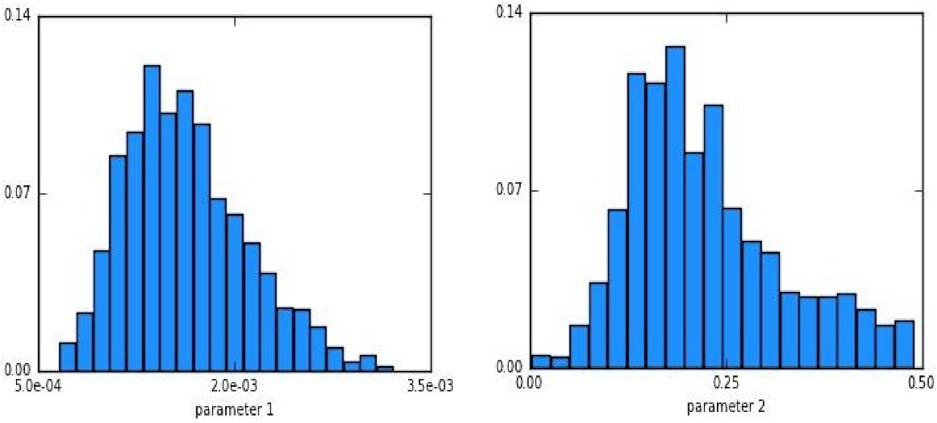


Fig. 6. The posterior distributions of the parameters of the dimerization model using exact simulation (B_t = 10 and population size P = 1000)

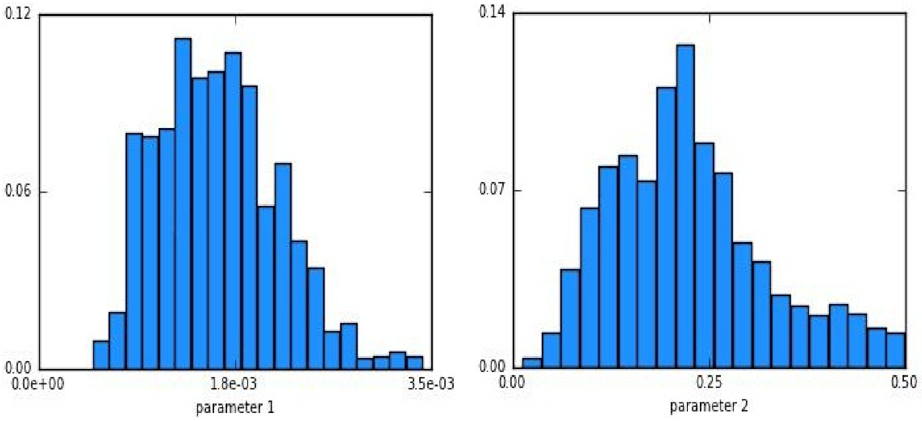


Fig. 7. The posterior distributions of the parameters of the dimerization model using approximate simulation (B_t = 10 and population size P = 1000)

5.2. Michaelis-Menten model

The second model being considered in detail is the Michaelis-Menten model of a biomolecular reaction between an enzyme E, which reversibly reacts with a substrate S forming an enzyme substrate complex C, which (as a result of an unimolecular reaction) is converted to product P and enzyme E. The stochastic model consists of three reactions: the reaction of forming an enzyme substrate complex $S + E \rightarrow \mathrm{C}$ (with rate $c1$), the reverse process of converting the enzyme substrate complex to a molecule of an enzyme and a molecule of a substrate $C \rightarrow S + E$ (with rate $c2$) and a reaction of forming a product $C \rightarrow E + P$ with rate $c3$. This model is used in a various biochemical processes. As in the previous case a synthetic dataset is generated from the model with parameters $(c1, c2, c3) = (0.00177, 0.0001, 0.1)$, to which a Gaussian noise from the distribution $N(0, (0.5)^2)$ is added, resulting in dataset of measurements.

Taking into account the influence of the number of simulations per particle on the precision of the approximation, we vary the number of accepted particlesper population for the parameter inference in the Michaelis-Menten model. The parameters that are being inferred have the following prior distributions $\pi(c1) = \pi(c2) \sim U(0, 0.05)$, $\pi(c3) \sim U(0, 0.5)$. The initial conditions for the species are as follows $(E, S, C, P) = (301, 120, 0, 0)$.

The process of parameter inference goes through the tolerance gradient $\varepsilon = (90, 70, 50, 40, 35)$, the distance metric being the square root of the sum of squared distances. The perturbation kernels are uniformly distributed according to $K_t = \sigma U(-1, 1)$, where $\sigma_{c1} = 0.001$, $\sigma_{c2} = 0.00001$, $\sigma_{c3} = 0.01$.

The results from the parameter inference performing 10 simulations per particle on a population of 100 particles are given in Figures 8 and 9. Figure 8 gives the results of using exact simulation, while Figure 9 gives the results of using approximate simulation.

It can be noticed that, the posterior distributions obtained by using approximate simulation the mean value have a small difference as compared to the original parameter values. In both cases (exact and approximate simulation), the distributions of the parameter values have a large spread. This is the result of choosing a small number of accepted particles per population. Our next step was to increase the number of accepted particles per population. The results can be seen in Figures 10 and 11.

Looking at the posterior distribution in the case of 1000 accepted particles per population, we can say that the parameters have good distributions that closely resemble a Gaussian distribution with a mean value close to the value of the original parameters. In Figure 12 the evolution of the concentrations through time using the inferred particle value distribution using exact and approximate simulation can be seen.

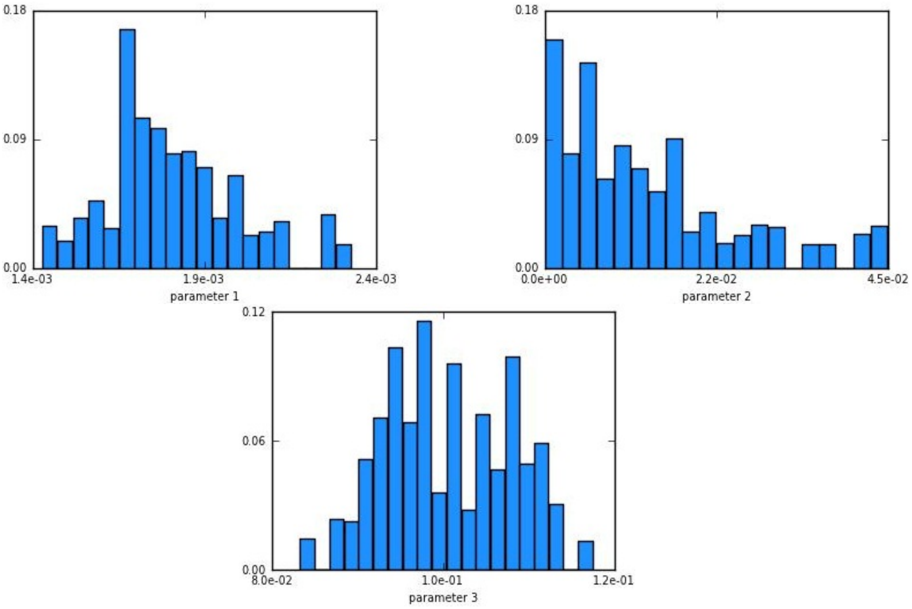


Fig. 8. The posterior distributions of the parameters of the Michaelis-Menten model estimated by using exact simulation (B_t = 10 and population size P = 100)

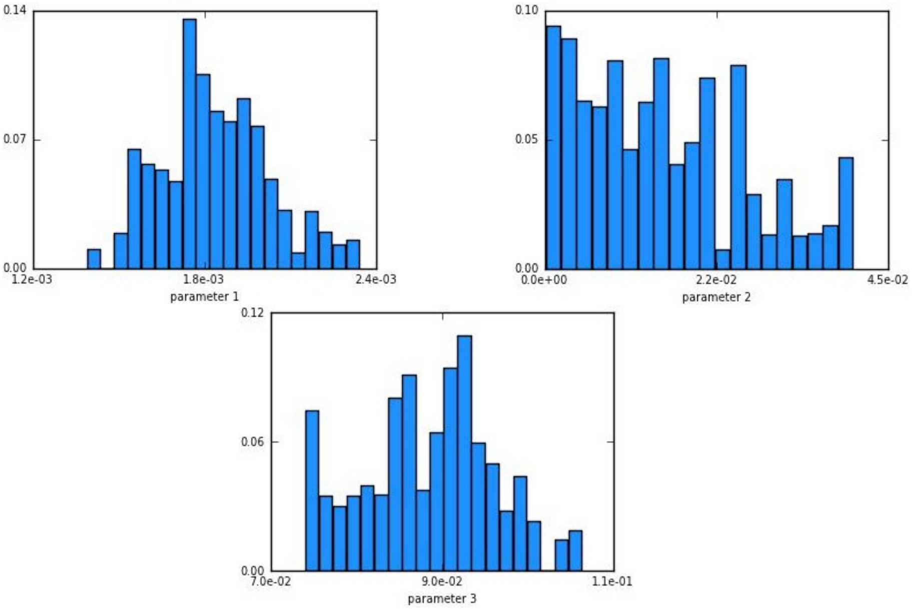


Fig. 9. The posterior distributions of the parameters of the Michaelis-Menten model estimated by using approximate simulation (B_t = 10 and population size P = 100)

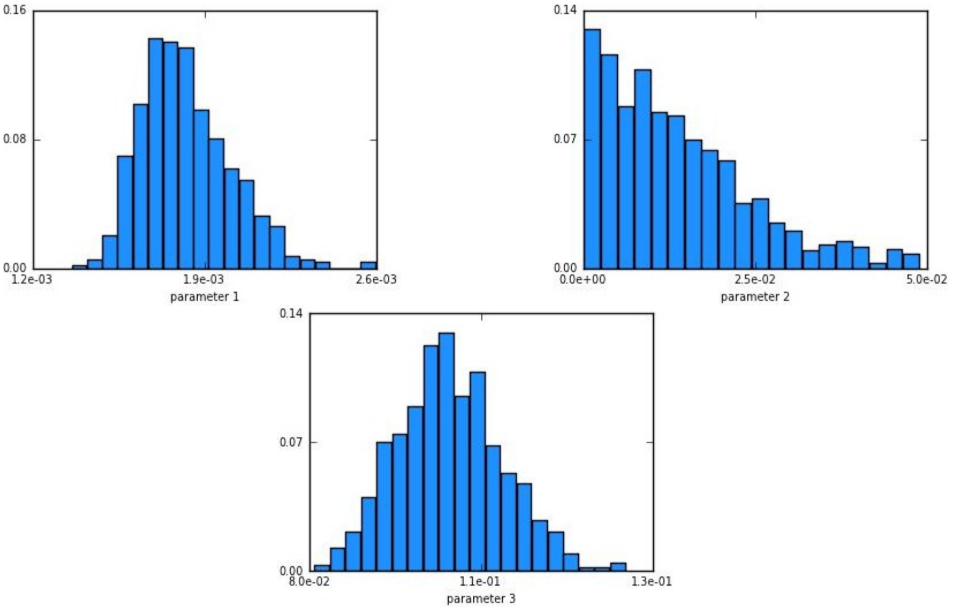


Fig. 10. The posterior distributions of the parameters of the Michaelis-Menten model estimated by using exact simulation ($B_t = 10$ and population size $P = 1000$)

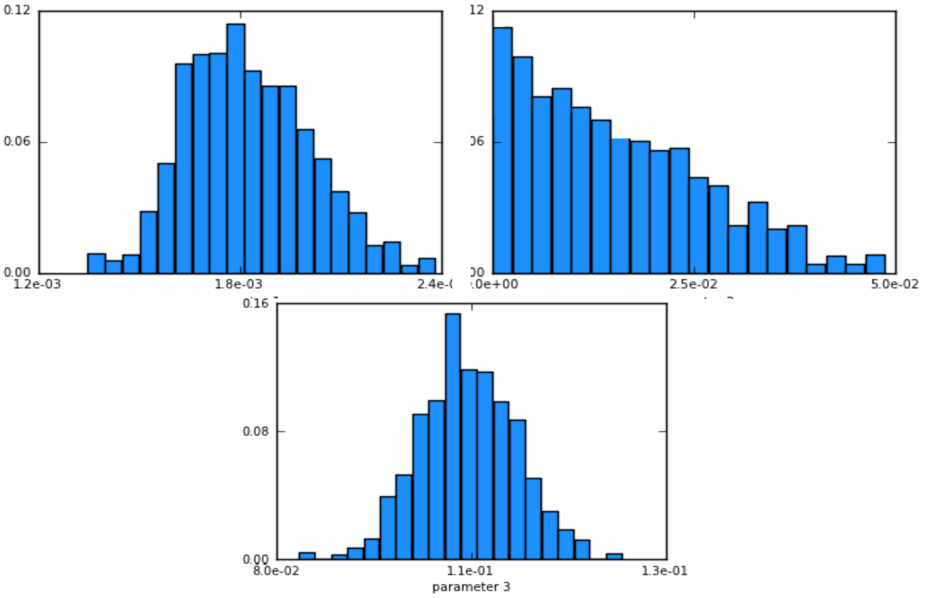


Fig. 11. The posterior distributions of the parameters of the Michaelis-Menten model estimated by using approximate simulation ($B_t = 10$ and population size $P = 1000$)

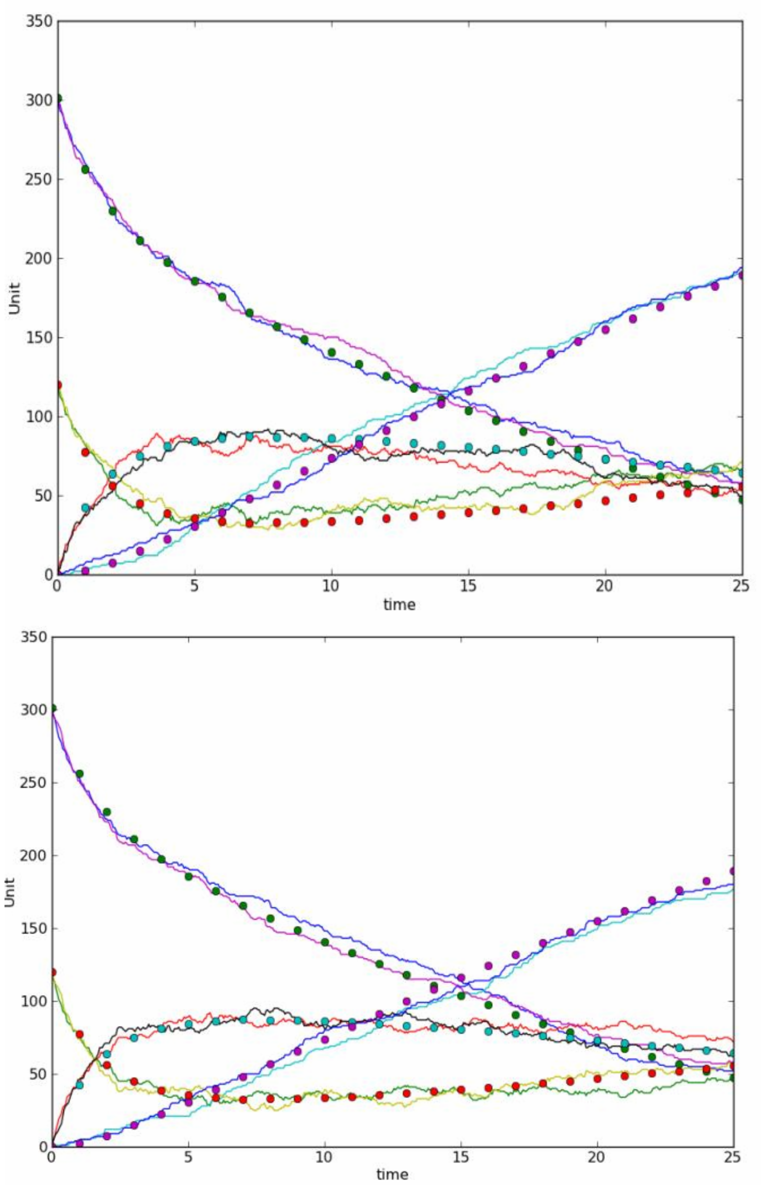


Fig. 12. Evolution of the concentrations of the species with inferred parameters. $B_t = 10$ and $P = 1000$ using exact (upper graph) and approximate (lower graph) simulation. Thedots represent the values from the synthetic dataset. In the region between the solid lines are the data generated by simulation using the inferred parameters

6. CONCLUSION

In this paper, we present the stochastic view of reaction kineticsas used for modeling dynamical systems in systems biology.

We consider the tasks of deductive and inductive inference with such stochastic models, i.e., simulation of and parameter inference in such models. We first give a review of exact and approximate simulation methods. We then discuss approximate Bayesian methods for parameter inference and the use of different simulation methods in this context.

The widely used methods for parameter inference, based on the determination of a likelihood function (e.g., maximum-likelihood parameter estimation), give point estimates of the parameter values in a dynamical system. Bayesian methods, on the other hand, infer a probability distribution of the values of the parameters in the system. Taking observed data into account, a prior distribution of the parameter values is evolved into a posterior distribution. Approximate Bayesian methods perform a large number of simulations of the model with parameter values sampled from the prior, comparing the simulation outcomes to the observed data.

The main contribution of this paper is the proposal to use approximatesimulation methods in the context of approximate Bayesian parameter inference, where many simulations are performed. Approximate simulation methods are much faster than exact methods: We expected this to result in much faster parameter estimation, hopefully without reducing the quality of the parameter estimates.

We have implemented the approximate simulation methods in an approximate Bayesian framework based on the sequential Monte Carlo method. We have compared the results of parameter inference with approximate vs. exact simulation on two parameter estimation tasks for two models of different dynamical systems. Parameter inference using approximate simulation methods gives satisfactory results, precise and very similar to the results obtained by using exact simulation.

The number of simulations per particle was evaluated for its effect on the acceptance rate and the variance of the posterior distribution of the parameter values. It was shown that the number of accepted particles per population has an effect on the variance of the posterior distribution. A larger number of accepted particles results in a smaller variance and more precise estimates of the parameter values.

Previous knowledge of the system is needed for adequate parameter estimation. The choice of the interval of the prior distribution has to be made accordingly. The increase of the interval results in a decrease of the acceptance rate, due to the random nature of the selection of the particles.

The results of our research set the stage for using approximate Bayesian parameter inference in the context of computational scientific discovery [32]. Methods for automated modeling of dynamic systems in systems biology [33] consider a large number of models structures, for which parameter inference needs to be performed. The efficiency and effectiveness of parameter inference is critical issue in this context.

REFERENCES

[1] H. Kuthan, Self-organisation and orderly processes by individual protein complexes in the bacterial cell, *Progress in Biophysics and Molecular Biology*, vol. **75**, pp. 1–17, 2001.

[2] S. Reineker, R. Altman, and J. Timmer, Parameter estimation in stochastic biochemical reactions, in *IEEE Proceedings in Systems Biology*, vol. **153**, 2006, pp. 168–178.

[3] J. H. Espenson. *Chemical kinetics and reaction mechanisms*. McGraw-Hill, New York, 2nd ed edition, 1995.

[4] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Reviewof Physical Chemistry*, **58** (1), pp. 35–55, 2007.

[5] D. T. Gillespie. A general method for numerically simulating the stochastictime evolution of coupled chemical reactions. *The Journal of Computational Physics*, **22** (4), pp. 403–434, 1976.

[6] D. Gillespie, Exact stochastic simulation of coupled chemical reactions, *Journal of Physical Chemistry*, vol. **81**, pp. 2340–2361, 1977.

[7] M. A. Gibson and J. Bruck, Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels, *Journal of Physical Chemistry*, vol. **104**, pp. 1876–1889, 2000.

[8] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, **188** (1–3), pp. 404–425, 1992.

[9] T. Tian, K. Burrage. Binomial leap methods for simulating stochasticchemical kinetics. *The Journal of Chemical Physics*, **121** (21), pp. 10356–10364, 2004.

[10] A. Chatterjee, D. G. Vlachos, M. A. Katsoulakis. Binomial distributionbased tau-leap accelerated stochastic simulation. *The Journal of Computational Physics*, **122** (2), 024112, 2005.

[11] Y. Cao, D. T. Gillespie, and L. R. Petzold. Avoiding negative populations inexplicit poisson tau-leaping. *The Journal of chemical physics*, **123** (5): 054104, 2005.

[12] M. Rathinam, L. Petzold, Y. Cao, D. Gillespie. Stiffness in stochastic chemically-reactingsystems: the implicit tau-leaping method. *J. Chem. Phys.*, **119,** pp. 12784–94, 2003.

[13] Y. Cao, D. T. Gillespie, L. R. Petzold. Adaptive explicit-implicit tau-leaping method with automatic tau selection. *The Journal of Chemical Physics*, **126** (22): 224101, 2007.

[14] T. E.Turner, S. Schnell, and K. Burrage, Stochastic approaches for modelling, *Computational Biology and Chemistry*, vol. **28**, no. 3, pp. 165–178, 2004.

[15] J. Timmer and T. Muller, Modeling the nonlinear dynamics of cellular signal transduction., *International Journal of Bifurcation and Chaos*, vol. **14**, pp. 2069–2079, 2004.

[16] A. Golightly and D. Wilkinson, Bayesian Inference for Stochastic Kinetic Models Using a Diffusion Approximation, *Biometrics*, vol. **61**, pp. 781–788, 2005.

[17] A. Golightly and D. Wilkinson, Bayesian sequential inference for stochastic kinetic biochemical network models, *Journal of Computational Biology*, vol. **13**, pp. 838–851, 2006.

[18] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, M. W. Feldman. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16** (12), pp. 1791–1798, 1999.

[19] S. Tavare, D. J. Balding, R. C. Griffiths, P. Donnelly. Inferring coalescencetimes for molecular sequence data, *Genetics*, **145**, pp. 505–518, 1997.

[20] S. A. Sisson. Genetics and stochastic simulation do mix!, *The American Statistician*, **61**, pp. 112–119, 2007.

[21] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, **21** (6), pp. 1087–1092, 1953.

[22] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57** (1), pp. 97–109, 1970.

[23] S. Geman, D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6** (6), pp. 721–741, 1984.

[24] P. Marjoram, J. Molitor, V. Plagnol, S. Tavare. Markov chain Monte Carlowithout likelihoods. *Proc. Natl. Acad. Sci. USA*, **100** (26), pp. 15324–15328, 2003.

[25] S. A. Sisson, Y. Fan, M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA,* **104**, 1760–1765, 2007.

[26] T.Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems, *Journal of The Royal Society Interface*, vol. **6**, pp. 187–202, 2009.

[27] P. Del Moral, A. Doucet, A. Jasra. Sequential Monte Carlo samplers. *J. R. Stat. Soc*. **B 68**, 411–432, 2006.

[28] T. W. Evans, C. S. Gillespie, and D. J. Wilkinson, The SBML discrete stochastic models test suite, *Bioinformatics*, vol. **24**, no. 2, pp. 285–286, 2008.

[29] *The Systems Biology Markup Language*. [Online]. http://www.sbml.org.

[30] J. Liepe, C. Barnes, E. Cule, K. Erguler, P. Kirk, T. Toni, M. P. H. Stumpf ABC-SysBio – Approximate Bayesian Computation in Python with GPU support *Bioinformatics*. **15**, 26 (14), pp. 1797–9, 2010.

[31] D. J. Wilkinson, *Stochastic modelling for systems biology*. London, UK: Chapman & Hall/CRC. 2006.

[32] S. Džeroski, L. Todorovski (eds.) *Computational Discovery of Scientific Knowledge*., Springer, 2007.

[33] S. Džeroski, L. Todorovski. Equation discovery for systems biology: finding the structure and dynamics of biological networks from time course data. *Current Opinion in Biotechnology*, **19** (4), pp. 360–368, 2008.

Резиме

АПРОКСИМАТИВНО БАЕСОВО ОДРЕДУВАЊЕ НА ПАРАМЕТРИ НА ДИНАМИЧКИ СИСТЕМИ ВО СИСТЕМСКАТА БИОЛОГИЈА

Овој труд предлага употреба на апроксимативни наспроти егзактни алгоритми за стохастичка симулација за апроксимативно Баесово одредување на параметри на динамички системи во системската биологија. Прво ја претставува математичката рамка за опис на модели од системската биологија, особено од аспект на стохастичката формулација наспроти детерминистичката формулација на моделите која е базирана на законот за дејство на масата. Наспроти методите за одредување на параметри базирани на максимална веродостојност, претставени се апроксимативни методи кои прават избор на параметри од позната почетна веројатносна распределба, која постепено еволуира кон постериорна распределба, преку споредба на симулирани податоци добиени од моделот со познато множество од мерења. Овој труд понатаму се осврнува на процесот на симулација и дава преглед на различните егзактни и апроксимативни методи на стохастичка симулација и нивните подобрувања кои ги предлагаме за употреба. Егзактните и апроксимативните симулатори се имплементирани и употребени во методите за апроксимативно Баесово одредување на параметри. Нашата евалуација на овие методи на две задачи за одредување на параметри на два различни модела покажува дека еднакво добри резултати се добиваат побрзо при употребата на апроксимативна симулација во споредба со користењето на егзактната симулација.

Клучни зборови: системска биологија, кинетика на реакции, стохастички модели, егзактна стохастичка симулација, апроксимативна стохастичка симулација, апроксимативно одредување на параметри

Adress:

Jovan Tanevski
Ss. Cyril and Methodius University in Skopje
Faculty of Electrical Engineering and Information Technologies
Rugjer Boshkovik bb
PO Box 574, MK-1001 Skopje
Republic of Macedonia
tanevski@gmail.com

Sašo Džeroski
Jožef Stefan Institute,
Department of Knowledge Technologies,
Jamova 39, SI-1000 Ljubljana
Slovenia
saso.dzeroski@ijs.si

Ljupčo Kocarev
Macedonian Academy of Sciences and Arts
Bul. Krste Misirkov, 2, P.O. Box 428, 1000 Skopje
Republic of Macedonia
lkocarev@ucsd.edu