Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2020, 2020:11
https://doi.org/10.28919/cmbn/4489
ISSN: 2052-2541

APPROXIMATE BAYESIAN COMPUTATION BASED ON PSEUDO-PRIOR ADJUSTMENT AND ITS ADHIBITION IN BIOSCIENCE

GAN LIU[1], YONGZHEN PEI[2,*], CHANGGUO LI[3]

[1]Department of Computer Science and Technology, Tiangong University, Tianjin, 300387, China

[2]Department of Mathematical Sciences, Tiangong University, Tianjin, 300387, China

[3]Department of Basic Science, Military Transportation University, Tianjin, 300161, China

Abstract. Approximate Bayesian Computation (ABC) is a powerful tool to solve problem in likelihood-free methods. Markov Chain Monte Carlo and Sequential Monte Carlo based on ABC are effective techniques for obtaining the posterior sample points. However, without consideration of convergence criterion and choice of proposal kernels, these methods will lead to inefficient sampling or large deviations in statistical inference. By contrast, for ABC rejection sampling, despite being computationally inefficient sampling, independent identically distributed samples are obtained from approximate posterior. In order to combine the advantages of the methods mentioned, an alternative method is proposed for the acceleration of likelihood-free Bayesian inference that uses the pseudo-prior to replace the prior in ABC rejection algorithm and weights each sample point obtained, where the prior is obtained based on historical information and experience and the pseudo-prior is a distribution different from the prior. And the weighted sample are considered to be from the target distribution. In our method, choosing a suitable pseudo-prior not only greatly improves the efficiency of the algorithm but also retains the accuracy advantages of rejection sampling. The approach is illustrated by parameter estimation in bioscience.

Keywords: likelihood-free methods; approximate Bayesian Computation; weight; pseudo-prior; RNA interference.

2010 AMS Subject Classification: 92B15.

*Corresponding author

E-mail address: yzhpei@tiangong.edu.cn

Received January 30, 2020

1. INTRODUCTION

Approximate Bayesian Computation (ABC) is considered to be the first member of a class of likelihood-free methods, which is indispensable in promoting study of Monte Carlo methods in the past decade. ABC was originally applied to solving challenging inferential problems in population genetics, where the complexity of a model means that the likelihood function is computationally difficult to handle. The essence of ABC method is that a numerical assessment of the likelihood function is replaced with an evaluation of the probability that the model generates the observed data, based on comparing the observed data and simulated data generated by the model. For this reason, it has been gaining popularity as an analysis tool in systems biology [1], population genetics [2][3], dynamic ecological models [4], cosmology and astrophysics [5][6].

The earliest ABC method [7][8] is basic rejection sampling algorithm (ABC-REJ), which originates from population genetics. In ABC-REJ, first of all, the candidate parameter $\theta^{'}$ is sampled from the prior distribution $\pi(\theta)$, and then substituting it into the model to obtain simulated data $x^{'}$. Subsequently, comparing the simulated data $x^{'}$ and observed data x_{obs}, if the distance from $x^{'}$ to x_{obs} is within the tolerance ε, then the observed data generated by $\theta^{'}$ is credible for this model, and so $\theta^{'}$ is accepted. Conversely, if $x^{'}$ and x_{obs} are dissimilar, then the reliability of generating observed data by $\theta^{'}$ is low for this model, and so $\theta^{'}$ is rejected. This method results in independent identically distributed samples from the approximated posterior $\pi_{\varepsilon}(\theta|x_{obs})$. However, ABC-REJ algorithm admits the inefficient sampling, that is, as most candidate parameter vectors are rejected, caused by the large difference between the prior and the posterior distribution. Various ABC methods were proposed to improve the acceptance rate of candidate parameters.

Regression approaches based on ABC (ABC-REG) was introduced to sovle the complex problems in population genetics [2]. After ABC-REJ algorithm is implemented by retaining parameters that produce simulated data close enough to observed data, parameters are adjusted to explain the discrepancy between simulated and observed data. ABC-REG method won't materially affect the accuracy of the estimated parameters if the tolerance is increased properly [2]. Therefore, to some extent this method can avoid the inefficient sampling of ABC-REJ.

However, the support of prior is not taken into account in this regression adjustment, that is, it may cause the support of posterior not to be in the support of prior[9].

ABC method based on Markov chain Monte Carlo (ABC-MCMC) was originally developed by Marjoram et al. What we obtain is a Markov Chain with the stationary distribution $\pi(\theta|x_{obs})$ by the ABC-MCMC [10]. If the current chain state is at $\theta^{(i)}$, and then sampling a candidate parameter $\theta^{'}$ from a transition kernel $q(\theta|\theta^{(i)})$. Subsequently, $\theta^{'}$ is substituted into the model to obtain simulated datasets $x^{'}$. Next, comparing the simulated and observed data, if $x^{'}$ and x_{obs} are similar, then accepting $\theta^{'}$ with probability $\alpha = \min\{1, \frac{\pi(\theta^{'})q(\theta^{(i)}|\theta^{'})}{\pi(\theta^{(i)})q(\theta^{'}|\theta^{(i)})}\}$, and otherwise stay at $\theta^{(i)}$. ABC-MCMC method solves the problem of inefficient sampling by constantly adjusting the transition kernel, so that the candidate parameters sampled from the transition kernel are easier to simulate the observed data. However, an inappropriate transition kernel will have significant impact on the efficiency of ABC-MCMC [11]. The problem of choosing a transition kernel is non-trivial. In general, the transition kernel is determined heuristically. However, the optimal proposal can be obtained by adaptive schemes in some cases [12][13]. Another challenge is to determine when the Markov Chain has converged [14].

ABC method based on sequential Monte Carlo (ABC-SMC) sampling was introduced by Sisson et al [15], which is driven from a sequential importance sampling (SIS) algorithm [16][17]. In ABC-SMC, A set of samples $\theta_1, ..., \theta_N$ (called particles) is evolved from the prior distribution $\pi(\theta)$. And it is propagated through a sequence of intermediate distributions, $\pi(\theta|\rho(x, x_{obs}) \leq \varepsilon_i), i = 1, ..., T-1$, up till the particles denotes a sample from the target distribution $\pi(\theta|\rho(x, x_{obs}) \leq \varepsilon_T)$. The tolerances are selected such that $\varepsilon_1 > ... > \varepsilon_i > ... > \varepsilon_T \geq 0, i = 1, ..., T$. Therefore, the intermediate distributions gradually approach the target distribution. The advantage of ABC-SMC method is that it can avoid the above-mentioned disadvantages of ABC-REJ and ABC-MCMC methods at least to some extent. The disadvantage of ABC-SMC is to choose a suitable sequence of acceptance thresholds. In order to resolve this problem, some schemes were proposed to generate these sequences adaptively [18][19].

The above algorithms respectively use different ideas to solve the problem that the ABC rejection algorithm is inefficient when the prior and posterior are dissimilar. In this paper, we propose a new idea based on a pseudo-prior adjustment to solve this problem. First, we find

a suitable pseudo-prior to replace the given prior in ABC rejection algorithm. The resulting samples are derived from the posterior corresponding to the pseudo-prior. Then, each sample point is given a weight to ensure that it comes from the posterior corresponding to the given prior. In this method, choosing a suitable pseudo-prior not only greatly improves the efficiency of the algorithm but also retains the accuracy advantages of rejection sampling.

2. METHODS

2.1. ABC Rejection Sampling. The parameter to be estimated in the model is θ, and the prior is $\pi(\theta)$. Given the observed data, x_{obs}, and the likelihood of the model, $L(x|\theta)$, the ABC-REJ algorithm is as follows.

Algorithm 1 ABC Rejection sampler

1: **for** $i = 1; i < N; i++$ **do**
2: **repeat**
3: $\theta^* \sim \pi(\theta)$;
4: $x^* \sim L(x|\theta^*)$
5: **until** $\rho\{S(x^*), S(x_{obs})\} \leq \varepsilon$
6: set $\theta^{(i)} = \theta^*$
7: **end for**

where the marks of the algorithm are

$-S$, a function on x defining a statistic which most often is not sufficient.

$-\rho > 0$, a distance on $S(x)$.

$-\varepsilon > 0$, a tolerance level.

2.2. Pseudo-prior ABC. We propose two innovations based on ABC-REJ sampling: using the pseudo-prior $\phi(\theta)$ instead of the prior $\pi(\theta)$ and weighting each sample point. The pseudo-prior is different from the prior but the same support as the prior. We choose a pseudo-prior based on the following two points. The first is to improve the efficiency of ABC-REJ sampling. The second is that the mode of the pseudo-prior chooses the parameter corresponding to the point that makes the likelihood function as large as possible. As for weighting each sample point, we realize that the weighted sample are from the target distribution. First, we replace the

prior $\pi(\theta)$ with the pseudo-prior $\phi(\theta)$ at step 3 of the Algorithm 1, the resulting sample are derived from posterior $\phi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$ (called it pseudo-posterior) corresponding to the pseudo-prior $\phi(\theta)$. To ensure that the sample are from the posterior $\pi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$, each sample point is given a weight $\omega_i = \frac{\pi(\theta_i)}{\phi(\theta_i)}$, $i = 1,...,N$. The sampler of ABC algorithm based on pseudo-prior adjustment (ABC-PPA) is as follows.

Algorithm 2 ABC-PPA sampler.

Input: Observation, x_{obs};

Number of iterations N;

Output: The sample $\{\theta_1^{'},\theta_2^{'},...,\theta_N^{'}\}$;

1: Determine the pseudo-prior $\phi(\theta)$;

2: Simulate $\theta_1,...,\theta_N$ from the posterior $\phi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$ corresponding to the pseudo-prior $\phi(\theta)$ by ABC Rejection Sampling;

3: Weight θ_i by $\omega_i = \frac{\pi(\theta_i)}{\phi(\theta_i)}, i = 1,2,...,N$

4: Compute the sum of the weights, $\omega_0 = \sum_{i=1}^{N} \omega_i$

5: Compute the normalised weights $\omega_i^{'} = \omega_i/\omega_0, i = 1,2,...,N$

6: Sample N times, with replacement from the set $\{\theta_1,\theta_2,...,\theta_N\}$ using the probabilities $\{\omega_1^{'},\omega_2^{'},...,\omega_N^{'}\}$ (for example, using the lookup method) to generate a new sample $\{\theta_1^{'},\theta_2^{'},...,\theta_N^{'}\}$

7: **return** $\{\theta_1^{'},\theta_2^{'},...,\theta_N^{'}\}$

2.2.1. *Pseudo-prior.* when the value of the likelihood function is larger, the probability of simulating the observed data is greater for this model. If a parameter value that makes the likelihood function as large as possible is chosen as the mode of the pseudo-prior, it will greatly improve the efficiency of our algorithm. Now, finding a parameter value that makes the likelihood function as large as possible is a problem that needs to be solved. But ABC is a likelihood-free method, it is not feasible to directly obtain the maximum value of the likelihood function. So an alternative method is proposed for the approximation of the likelihood function that uses the distance between the simulated data and the observed data [20].

$$L(x_{obs}|\theta) \propto exp(-E^2) \tag{1}$$

where $E^2 = \rho\{S(x^*), S(x_{obs})\}$. Equation (1) shows that the closer the distance between the simulated data and the observed data is, the larger the likelihood will be. Therefore, the mode of the pseudo-prior is the parameter values that make the simulated data and the observed data as close as possible. The algorithm for finding the mode of the pseudo-prior is as follows.

Algorithm 3 The search for a mode of the pseudo-prior.

Input: Observation, x_{obs};

Simulation times, p;

threshold, δ;

Output: the mode of the pseudo-prior, $\theta^{(*)}$;

Generate a candidate value $\theta^{(0)}$ from prior distribution $\pi(\theta)$ and $\theta^{(*)} = \theta^{(0)}$;
Simulate $\{x_1^{'}, x_2^{'}, ..., x_p^{'}\}$ from the model $L(x|\theta^{(0)})$;
Make $x^{'}$ be the mean of $\{x_i^{'}\}_{i=1,...,p}$;
Calculate the distance $dist\_min = ||x^{'} - x_{obs}||$;
while $dist\_min \leq \delta$ **do**
    Generate a proposed value $\theta^{(i)}$ from proposal distribution $q(\theta|\theta^{(*)})$;
    Simulate $\{x_1^{'}, x_2^{'}, ..., x_p^{'}\}$ from the model $L(x|\theta^{(i)})$;
    Make $x^{'}$ be the mean of $\{x_i^{'}\}_{i=1,...,p}$;
    Calculate the distance $dist = ||x^{'} - x_{obs}||$;
    if $dist < dist\_min$ **then**
        $dist\_min = dist$;
        $\theta^{(*)} = \theta^{(i)}$;
    end if
end while
return $\theta^{(*)}$;

The mode $\theta^{(*)}$ of the pseudo-prior is obtained by Algorithm 3. The specific form of the pseudo-prior can be determined according to the actual situation. In general, for continuous random variables, we choose a Gaussian distribution where the mean is equal to the mode and the standard deviation is given, for discrete random variables, we choose Poisson distribution.

2.2.2. *Weighting.* The prior $\pi(\theta)$ is replaced by the pseudo-prior $\phi(\theta)$ at step 3 of the Algorithm 1, and the resulting sample are derived from the pseudo-posterior $\phi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$. To ensure that the sample are from the posterior $\pi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$, each sample point is given a weight $\omega_i = \frac{\pi(\theta_i)}{\phi(\theta_i)}, i = 1,2,...,N$. The following is the process of explaining the form of weights.

Based on Importance Resampling (IR) framework, we now define ABC-PPA as a special case of IR algorithm, where the target distribution is the posterior, and the proposal distributions is chosen as the pseudo-posterior, it is as follow:

Algorithm 4 Importance Resampling.

Input: Proposal distributions $\phi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$;

Target distribution $\pi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$;

Number of iterations N;

Output: The sample $\{\theta_1^{'},\theta_2^{'},...,\theta_N^{'}\}$;

1: Simulate $\theta_1,...,\theta_N$ from the pseudo-posterior $\phi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$;

2: Weight θ_i by $\omega_i = \frac{\pi(\theta|\rho(S(x_{obs}),S(x))\leq\varepsilon)}{\phi(\theta|\rho(S(x_{obs}),S(x))\leq\varepsilon)}$, i=1, 2, . . . , N

3: Compute the sum of the weights, $\omega_0 = \sum_{i=1}^{N} \omega_i$

4: Compute the normalised weights $\omega_i^{'} = \omega_i/\omega_0, i = 1,2,...,N$

5: Sample N times, with replacement from the set $\{\theta_1,\theta_2,...,\theta_N\}$ using the probabilities $\{\omega_1^{'},\omega_2^{'},...,\omega_N^{'}\}$ (for example, using the lookup method) to generate a new sample $\{\theta_1^{'},\theta_2^{'},...,\theta_N^{'}\}$

6: **return** $\{\theta_1^{'},\theta_2^{'},...,\theta_N^{'}\}$

In Algorithm 4, the sample $\{\theta_1^{'},\theta_2^{'},...,\theta_N^{'}\}$ is from the posterior $\pi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$, but the target distribution $\pi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)$ is unknown so that the weight of Algorithm 4 cannot be directly obtained. So we use the Bayesian formula to simplify the weights of Algorithm 4. The process is as follows.

For prior $\pi(\theta)$, using Bayes theorem, the resulting posterior distribution:

$$\pi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon) = \frac{\pi(\theta)L(\rho(S(x_{obs}),S(x)) \leq \varepsilon|\theta)}{\int \pi(\theta)L(\rho(S(x_{obs}),S(x)) \leq \varepsilon|\theta)d\theta} \tag{2}$$

For pseudo-prior $\phi(\theta)$, using Bayes theorem, the resulting pseudo-posterior distribution:

$$\phi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon) = \frac{\phi(\theta)L(\rho(S(x_{obs}),S(x)) \leq \varepsilon|\theta)}{\int \phi(\theta)L(\rho(S(x_{obs}),S(x)) \leq \varepsilon|\theta)d\theta} \tag{3}$$

In equation (2) and equation (3), the likelihood function $L(\rho(S(x_{obs}),S(x)) \leq \varepsilon|\theta)$ are the same for the same tolerance ε. So the following relationship can be obtained equivalently by equations (2) and (3).

$$\frac{\pi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)}{\phi(\theta|\rho(S(x_{obs}),S(x)) \leq \varepsilon)} = \frac{\pi(\theta)}{\phi(\theta)} \times C \tag{4}$$

where $C = \frac{\int \phi(\theta)L(\rho(S(x_{obs}),S(x)) \leq \varepsilon|\theta)d\theta}{\int \pi(\theta)L(\rho(S(x_{obs}),S(x)) \leq \varepsilon|\theta)d\theta}$, and C is a constant.

The weights of Algorithm 4 can be written as follows according to equation (4).

$$\omega_i = \frac{\pi(\theta_i)}{\phi(\theta_i)} \times C \tag{5}$$

Since the weights need to be normalised, the weights of Algorithm 4 can be further written as follows.

$$\omega_i = \frac{\pi(\theta_i)}{\phi(\theta_i)} \tag{6}$$

3. RESULTS

In this section, we first verify the effectiveness of our algorithm with a toy model, then illustrate sampling efficiency and accuracy of the estimation results with stochastic process model of pest (Continuous random variable) and RNA interference model (Discrete random variable).

3.1. Toy model. We first examine how ABC-PPA performs in a toy example, where the posterior is known. The model that we consider is binomial distribution $B(100,\theta)$, where θ is the parameter to be estimated. For the prior we specify the beta distribution $Beta(2,5)$, and for the observed data we take $x_{obs} = 80$.

From the above known conditions, ABC-PPA algorithm is applied to estimate parameter θ. We assume that the pseudo-prior is Gaussian distribution under the support of prior. The mode of the the pseudo-prior is obtained by Algorithm 3, and in the Gaussian distribution, the mode is equal to the mean and the standard deviation is given 0.2. The posterior sample are obtained by the Algorithm 2, where $N = 10000$ and the tolerance $\varepsilon = 0$.

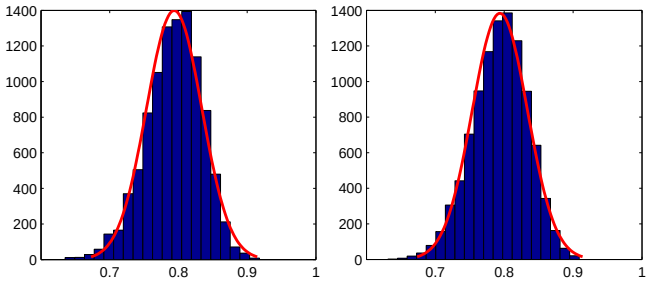


FIGURE 1. The bar chart on the left shows that it is obtained with ABC-PPA algorithm, and the bar chart on the right shows the sample taken directly from the true posterior distribution.

The true posterior can be directly obtained by Bayesian formula as the beta distribution $Beta(81,21)$. Figure 1 shows that the sample obtained directly from the true posterior distribution is very similar to the sample obtained by the ABC-PPA algorithm. The parameter estimated by the true posterior distribution is 0.7941, and the parameter estimated by ABC-PPA algorithm is 0.7938. The two estimates are very similar, which also illustrates the feasibility of ABC-PPA.

3.2. Stochastic process model of pest. In this section, we deliberate a cotton aphids model proposed in [21][22], where $\overline{N}(t)$ denotes the number of the aphid at current time, $\lambda\overline{N}(t)$ represents aphid population birth rate [23], $\overline{C}(t)$ denotes the environment deteriorated, and $\eta\overline{N}(t)\overline{C}(t)$ be mortality of the aphid. and for simplicity, we ignore the condition of immigration and emigration. Modelling these two biochemical reactions as follow,

$$\begin{aligned} \overline{N} &\xrightarrow{\lambda} 2\overline{N}+\overline{C} \\ \overline{N}+\overline{C} &\xrightarrow{\eta} \overline{C} \end{aligned} \tag{7}$$

For 7, 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 [24]. The parameter values $\lambda = 2.453$, $\eta = 0.0094$ and the initial values $\overline{N}(0) = 1$, $\overline{C}(0) = 1$ are given. Simulations for the dynamic of the aphid by Gillespie algorithm [25][26] are illustrated in figure 2.

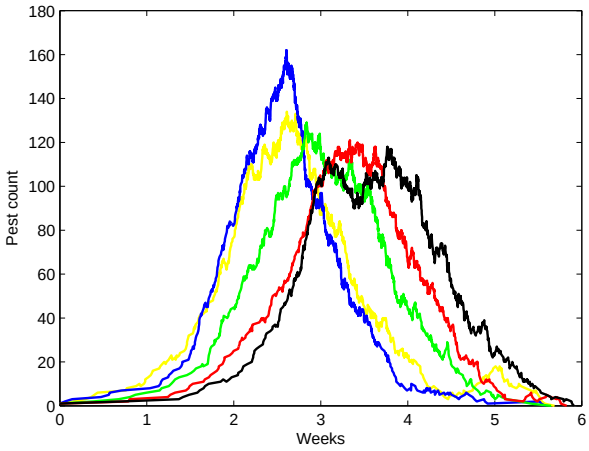


FIGURE 2. Time evolution of the pest population simulated by Gillespie algorithm

The observed data consist of 8 data points for the aphid with rates $(\lambda,\eta) = (2.453, 0.0094)$ and initial conditions $(\overline{N}(0),\overline{C}(0)) = (1,1)$. The prior distributions for λ and η are taken to be uniform, $\lambda \sim U(0,5)$, $\eta \sim U(0,0.1)$.

First, we apply the ABC-REJ sampler approach with ε =50. The inferred posterior distributions are shown in figure 3(a).

We apply the ABC-REG approach, where the regression model is linear and the weighting function is the Epanechnikov kernel. The 3% of simulated $x^{'}$ that are closest to x_{obs} are assigned a nonzero weight, and simulation sizes $N = 20000$. The inferred posterior distributions are shown in figure 3(b).

Applying the ABC-MCMC approach. We ran a Markov chain for 20000 iterations, and discarded the first 4000 iterations from this chain as burn-in. The inferred posterior distributions are shown in figure 3(c).

Next, we apply the ABC-SMC approach. The perturbation kernels for both parameters are Gaussian distribution, where the standard deviation of the random variable λ is taken 0.1, and the standard deviation of the random variable η is 0.0001. The number of particles in each population is $N = 1000$. To ensure the gradual transition between populations, we take $T = 6$ populations with ε =(300, 200, 100, 80, 60, 50). The inferred posterior distributions are shown in figure 3(d).

Finally, we apply the ABC-PPA approach, where the pseudo-prior takes the form of Gaussian distribution. And the mean of the pseudo-prior can be obtained by Algorithm 3. The standard deviation of the pseudo-prior of the random variable λ is taken 0.5. The standard deviation of the pseudo-prior of the random variable η is 0.002. The inferred posterior distributions are shown in figure 3(e).

When different ABC methods are applied, the corresponding approximate posterior samples will be obtained (figure 3), and the number of iterations required to obtain these samples (Table 2). The obtained samples will be averaged to approximate the estimated parameters (Table 1).

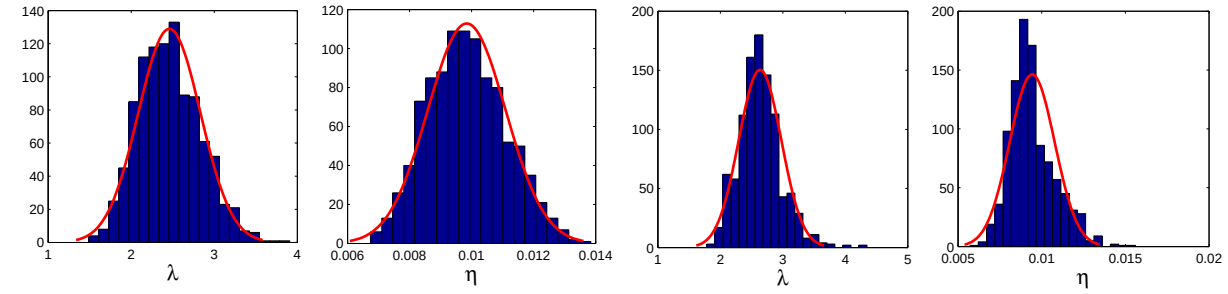
TABLE 1. Parameter estimation results.

Method	True value	ABC-REJ	ABC-REG	ABC-MCMC	ABC-SMC	ABC-PPA
λ	2.453	2.459762	2.621292	2.470784	2.422911	2.430638
η	0.0094	0.009841	0.009402	0.009972	0.009720	0.009767

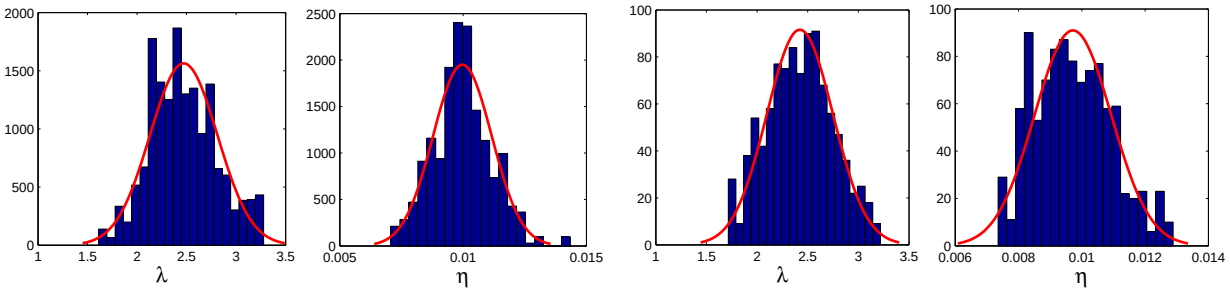
TABLE 2. The number of iterations required to obtain 1000 samples.

Method	ABC-REJ	ABC-SMC	ABC-PPA
times	561135	(1046, 2266, 27093, 4525, 6534, 7729)	10674

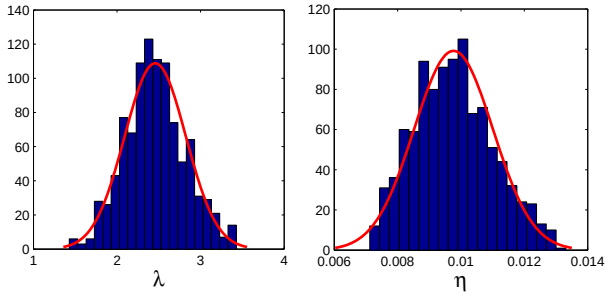
Table 1 shows that ABC-PPA has similar accuracy to ABC-REJ in terms of estimation results. And in the condition of obtaining the same number of samples, Table 2 shows that the number of iterations required by ABC-PPA is far less than that required by ABC-REJ.



(a) ABC-REJ (b) ABC-REG



(c) ABC-MCMC (d) ABC-SMC



(e) ABC-PPA

FIGURE 3. (a):ABC-REJ method for estimating the posterior distribution of parameters λ and η.(b):ABC-REG method for estimating the posterior distribution of parameters λ and η.(c):ABC-MCMC method for estimating the posterior distribution of parameters λ and η.(d):ABC-SMC method for estimating the posterior distribution of parameters λ and η.(e):ABC-PPA method for estimating the posterior distribution of parameters λ and η.

3.3. RNA interference model. RNA interference (RNAi) is a gene silencing phenomenon, which usually caused by the specific transcription of the double-stranded RNA (dsRNA) molecules [27][28]. First, long dsRNA is cleaved into siRNA by Dicer enzyme with the participation of ATP. Then the siRNA is derotated by the RNA helicase into a sense strand and an antisense strand with the participation of ATP, where the antisense strand directs the formation of an activated RNA-induced silencing complex (RISC). The activated RISC recognizes the target mRNA under the guidance of single-stranded siRNA, and cleaves the target mRNA from the target gene corresponding to the siRNA-guided strand center under the action of the endonuclease in RISC, thereby interfering with gene expression.

But, the main obstacle to effective siRNA uptake is the membrane. Although siRNA molecules are small in size, they still cannot enter the membrane directly because of their negative charge and hydrophilicity, and they enter the membrane by endocytosis and exocytosis. However, most of the siRNA is degraded during this process, and only a small fraction can escape to participate in the effects of RNA interference. And the study found that the escaped siRNA will have an amplification process. So estimating the amount of escape plays an important role in our study of RNA interference processes.

Locust is an important agricultural pest in the world, which can harm more than 20 crops. In recent years, locust outbreaks have become more frequent and serious in China [29]. We know that the growth of locusts depends strictly on the biosynthesis and degradation of chitin, which does not exist in plants and vertebrates. And excess or absence of chitinase can cause locust death. Therefore, chitin metabolism is an attractive target for the development of safe and effective pesticides.

The number of siRNAs injected into locusts is controlled by random processes including amplification, degradation, immigration and emigration, which are controlled by parameter set $\Phi = \{\alpha, \beta, \gamma, \eta, \}$, where α is the amplification rate, β is the degradation rate, γ is the immigration rate and η is the migration rate. And $\bar{S}(t)$ denotes the current number of siRNAs, and when $t = 0$, it is the amount of escape $s_1 = \bar{S}(0)$. Four kinds of random processes are simulated

by four biochemical reactions as follows.

$$
\begin{aligned}
&\overline{S} \xrightarrow{\alpha} 2\overline{S} \\
&\overline{S} \xrightarrow{\beta} \varnothing \\
&\varnothing \xrightarrow{\gamma} \overline{S} \\
&\overline{S} \xrightarrow{\eta} \varnothing
\end{aligned} \tag{8}
$$

It is assumed that the initial injection amount of siRNAs $s_0 = 1000$. The parameter values $\alpha = 0.6$, $\beta = 0.3$, $\gamma = 0.6$, $\eta = 0.23$ and the target amount of escape $s_1 = \bar{S}(0) = 700$ are given [30]. And the number of siRNAs simulated by Gillespie algorithm varies with time as shown in figure 8. When $t = 12$ hours, the number of amplified siRNA is used as observed data s_2.

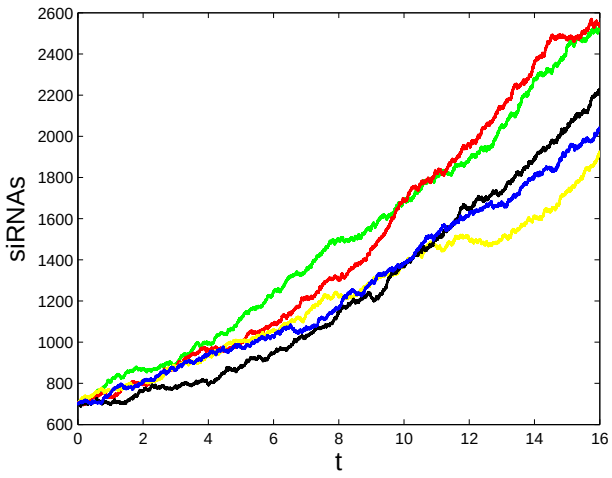


FIGURE 4. Time evolution of the siRNA simulated by Gillespie algorithm.

We apply the ABC-REJ sampler approach with $\varepsilon = 40$. The prior distributions for s_1 is taken to be discrete uniform distribution, $s_1 \sim U(1,1000)$. The inferred posterior distributions are shown in figure 4.

Applying the ABC-PPA approach with $\varepsilon = 40$. The prior distributions for s_1 is taken to be discrete uniform distribution, $s_1 \sim U(1,1000)$, the form of pseudo-prior can be constructed by the distance to the mode that obtained by Algorithm 3. The inferred posterior distributions are shown in figure 4.

The result of the parameter estimation: the amount of escape s_1 estimated by the ABC-REJ method is 701, and it estimated by ABC-PPA is 699. The estimation of both methods are very accurate.

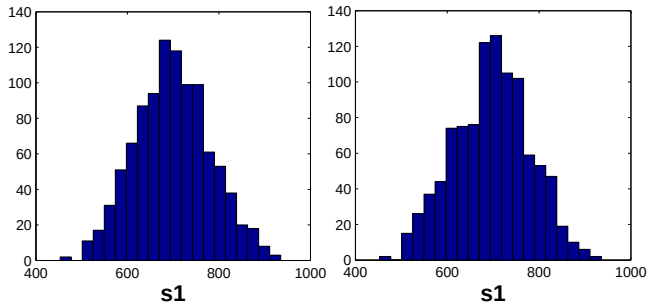


FIGURE 5. The posterior distribution of the escape amount: the bar graph on the left is obtained by the ABC-REJ method, and the bar graph on the right is obtained by the ABC-PPA method.

4. DISCUSSION

Bayesian statistical inference without likelihood is growing in common, especially in systems biology and dynamic ecological models, so it is vital to provide efficient and feasible methods to the practitioner. The deficiencies of existing likelihood-free method have been mentioned in the introduction. So we introduce ABC-PPA algorithm. The results indicate that, our method not only greatly improves the efficiency of the algorithm but also retains the accuracy advantages of ABC-REJ sampling. However, the problem of how to choose the variance of pseudo-prior is non-trivial. On the one hand, if the variance of pseudo-prior is given to be small, it will cause a big deviation in the results. On the other hand, if the variance of pseudo-prior is given to be large, it will affect the efficiency of ABC-PPA.

In stochastic evolution of pest populations, various Bayesian methods are used to estimate parameters. Through comparison, it is found that ABC-PPA method not only retains the accuracy advantage of ABC-REJ, but also greatly improves the sampling efficiency.

In the RNA interference mechanism, the part of siRNA that escapes from the endosome is one of the important factors that determine the efficiency of target mRNA silencing. Estimation

of the escape amount plays an important role in studying the mechanism of RNA interference, and also provides a direction for quantitative research to enhance endosomal escape. Here we use ABC-REJ and ABC-PPA methods to estimate the amount of escape, respectively. The estimation results of both methods are very accurate. And ABC-PPA method is faster than ABC-REJ method in sampling. Our study provides two new statistical methods to infer the amount of siRNAs in the actual RNAi reaction.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] Wilkinson D J. Stochastic modelling for quantitative description of heterogeneous biological systems. Nat. Rev. Genet. 10 (2) (2009), 122-133.

[2] M.A. Beaumont, W. Zhang, D. J. Balding, Approximate bayesian computation in population genetics, Genetics, 162 (4) (2002), 2025–2035.

[3] T. Thorne, M. P. Stumpf, Graph spectral analysis of protein interaction network evolution, J. R. Soc. Interface, 9 (75) (2012), 2653–2666.

[4] M. Fasiolo, N. Pya, S. Wood, Statistical inference for highly non-linear dynamical models in ecology and epidemiology, arXiv:1411.4564v1, 2014.

[5] T. Kacprzak, J. Herbel, A. Amara, A. Réfrégier, Accelerating approximate bayesian computation with quantile regression: application to cosmological redshift distributions, J. Cosmol. Astro Part. Phys. 2018 (2018), 042.

[6] C.H. Hahn, M. Vakili, K. Walsh, A. P. Hearin, D. W. Hogg, D. Cambpell, Approximate bayesian computation in large scale structure: constraining the galaxy-halo connection, Mon. Not. R. Astron. Soc. 469 (3) (2018), 2791–2805.

[7] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, M. W. Feldman, Population growth of human y chromosomes: a study of y chromosome microsatellites, Mol. Biol. Evol. 16 (12) (1999), 1791–1798.

[8] S. J. Tavare, D. J. Balding, R. C. Griffiths, P. Donnelly, Inferring coalescence times from dna sequence data, Genetics, 145 (2) (1997), 505–518.

[9] C.L.D.W.L. Excoffier, Bayesian computation and model selection in population genetics. arXiv:0901.2231v1, 2009.

[10] P. Marjoram, J. Molitor, V. Plagnol, S. Tavare, Markov chain monte carlo without likelihoods, Proc. Natl. Acad. Sci. USA, 100 (26) (2003), 15324–15328.

[11] P.J. Green, K. Latuszyński, M. Pereyra, et al. Bayesian computation: a summary of the current state, and samples backwards and forwards. Stat. Comput. 25(4)(2015), 835-862.

[12] S. Cabras, M.E.C. Nueda, E. Ruli, Approximate Bayesian computation by modelling summary statistics in a quasi-likelihood framework. Bayesian Anal. 10 (2) (2015), 411-439.

[13] G.O. Roberts, J.S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. Ann. Appl. Probab. 16 (4) 2006, 2123-2139.

[14] G. O. Roberts, J. S. Rosenthal, et al., General state space markov chains and mcmc algorithms, Probab. Surv. 1 (2004), 20–71.

[15] S. A. Sisson, Y. Fan, M. M. Tanaka, Sequential monte carlo without likelihoods, Proc. Natl. Acad. Sci. USA, 104 (6) (2007), 1760–1765.

[16] P. Del Moral, A. Doucet, A. Jasra, Sequential monte carlo samplers, J. R. Stat. Soc., Ser. B, Stat. Methodol. 68 (3) (2006), 411–436.

[17] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, M. P. Stumpf, Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems, J. R. Soc. Interface 6 (31) (2008), 187–202.

[18] C. C. Drovandi, A. N. Pettitt, Estimation of parameters for macroparasite population evolution using approximate bayesian computation, Biometrics, 67 (1) (2011), 225–233.

[19] D. Silk, S. Filippi, M. P. H. Stumpf, Optimizing threshold-schedules for sequential approximate bayesian computation: applications to molecular systems, Stat. Appl. Genet. Mol. Biol. 12 (5) (2013), 603–618.

[20] A. Pandey, A. Mubayi, J. Medlock, Comparing vector-host and sir models for dengue transmission, Math. Biosci. 246 (2) (2013), 252–259.

[21] C.S. Prajneshu, A nonlinear statistical model for aphid population growth. J. Indian Soc. Agric. Stat. 51 (1998), 73-80.

[22] Y. Pei, S. Lu, C. Li, B. Liu, Y. Liu, Optimal pest regulation tactics for a stochastic process model with impulsive controls using regression analysis-taking cotton aphids as an example, J. Biol. Syst. 27 (01) (2019), 107–129.

[23] J.H. Matis, T.R. Kiffe, T. I. Matis, D. E. Stevenson, Application of population growth models based on cumulative size to pecan aphids, J. Agric. Biol. Environ. Stat. 11 (4) (2006), 425–449.

[24] E. Koblents, J. Miguez, A population monte carlo scheme with transformed weights and its application to stochastic kinetic models, Stat. Comput. 25 (2) (2015), 407–425.

[25] A. Golightly, C.S. Gillespie, Simulation of Stochastic Kinetic Models. In: Schneider M. (eds) In Silico Systems Biology. Methods in Molecular Biology (Methods and Protocols), vol 1021. Humana Press, Totowa, NJ, 2013, pp 169-187.

[26] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Chem. Phys. 126 (12) (2007), 124108.

[27] A. Fire, S.Q. Xu, Montgomery M K, et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature, 391 (6669) (1998), 806-811.

[28] T. Ma, Y. Pei, C. Li, M. Zhu, Periodicity and dosage optimization of an RNAi model in eukaryotes cells, BMC bioinform. 20 (2019), 340.

[29] Xia J Y, Huang H. Analysis on the outbreak of Locusta migratoria manilensis and its control strategies. Plant Protect. Technol. Ext. 22 (2002), 7-10.

[30] T. Liu, Y. Pei, C. Li, et al. Amount of escape estimation based on Bayesian and MCMC approaches for RNA interference. Mol. Ther. Nucleic Acids, , 18 (2019), 893-902.