Approximate Bayesian Computation using Auxiliary Model Based Estimates

Anthony N Pettitt[1], Christopher C Drovandi[1] and Malcolm J Faddy[1]

[1] Mathematical Sciences, Queensland University of Technology. 2 George St, Brisbane, QLD, Australia, 4001. Email: a.pettitt@qut.edu.au.

Abstract: We present a novel approach for developing summary statistics for use in approximate Bayesian computation (ABC) algorithms using indirect inference. We embed this approach within a sequential Monte Carlo algorithm that is completely adaptive. This methodological development was motivated by an application involving data on macroparasite population evolution modelled with a trivariate Markov process. The main objective of the analysis is to compare inferences on the Markov process when considering two different indirect models. The two indirect models are based on a Beta-Binomial model and a three component mixture of Binomials, with the former providing a better fit to the observed data.

Keywords: Approximate Bayesian computation, Beta-Binomial model, Binomial mixture model, Indirect inference, Markov process, Sequential Monte Carlo.

1 Introduction

In approximate Bayesian computation (ABC), we seek to make inferences about the parameters of the posterior distribution when the likelihood function is computationally intractable. While the likelihood function itself cannot be computed easily, it is assumed that simulation from the model is relatively straightforward. The likelihood is replaced by a comparison of p summary statistics, $S(\cdot) = [S_1(\cdot), \ldots, S_p(\cdot)]^T$, of the observed and simulated data using a distance function, $\rho(y, x)$

$$\rho(y,x) \quad = \quad \|S(y) - S(x)\|.$$

ABC is particularly effective when the statistics are sufficient. However, in many applications sufficient statistics are not available and the practitioner must resort to a selection of carefully chosen data summaries.

In this paper we investigate an alternative approach to obtaining summary statistics based on indirect inference (Heggland and Frigessi 2004). In indirect inference an auxiliary model is proposed whose likelihood function is tractable and provides a good description of the data. The objective is to

search for parameter values of the model of interest that produce simulated data that lead to auxiliary parameters close to those based on maximum likelihood of the original data. Therefore a comparison of summary statistics involves computing a distance between such auxiliary parameters.
We consider a stochastic process model developed by Riley et al (2003) for a macroparasite population within a host. A Beta-Binomial model or a Binomial mixture is employed as an auxiliary model to provide a description of the data, while the stochastic model encapsulates the biological system which drives the observed data. We investigate the sensitivity of the inferences on the Markov process model to the indirect model. In particular we analyse any inefficiencies by introducing a three component Binomial mixture, which does not fit the data as well as the Beta-Binomial model.

2 Data and Modelling

Here the data is described as well as the stochastic process model of Riley et al (2003) used to explain the data. We also outline the auxiliary models.

2.1 Data

The data consist of mature parasite counts at particular autopsy times for 212 hosts (Denham et al 1972). Each host was injected with roughly 100 or 200 larvae and necropsy time ranged between 24 and 1193 days after the initial infection. The data are in the form of proportions (the mature count divided by the initial infection). From Figure 1 there is clear evidence of overdispersion, which a Binomial distribution alone cannot describe.

2.2 Markov Process Model

The following stochastic model was developed by Riley et al (2003) to help explain the population dynamics of *Brugia pahangi*. At time t any host is described by three random variables $\{M(t),\, L(t),\, I(t)\}$, where $M(t)$ is the number of mature parasites, $L(t)$ is the number of larvae and $I(t)$ is a discrete immunity variable. Initially cats are infected with L_I larvae and after a certain time the hosts are autopsied and the number of mature parasites are recorded. It is assumed that larvae can mature at a rate of γ per larva per day. Larvae die at a rate $\mu_L + \beta I(t)$ per larva where μ_L represents natural death of larvae and β describes the death of larvae due to the immune response of the host. The acquisition of immunity occurs at rate $\nu L(t)$, and a host loses immunity at a rate μ_I per unit of immunity. Mature parasites die at a rate of μ_M adults per day. In its deterministic form, the above model can be re-written as a set of differential equations

$$\frac{dL}{dt} = -\mu_L L - \beta I L - \gamma L,\ \frac{dM}{dt} = \gamma L - \mu_M M,\ \frac{dI}{dt} = \nu L - \mu_I I.$$

We consider the stochastic version of this model via a continuous time discrete trivariate Markov process as developed by Riley et al (2003). Data can be simulated from the model using the algorithm of Gillespie (1977). We consider $\mu_M = 0.0015$ and $\gamma = 0.04$ fixed as per Riley et al (2003).

2.3 Auxiliary Models

For the auxiliary model we propose a Beta-Binomial model, which contains an extra parameter to capture the dispersion. More specifically, the i^{th} observation has a Beta-Binomial distribution with Beta parameters, α_i and β_i. It is convenient to use a reparameterisation in terms of the proportion, $p_i = \alpha_i/(\alpha_i + \beta_i)$, and overdispersion, $\theta_i = 1/(\alpha_i + \beta_i)$, so that the mean and variance are given by $l_i p_i$ and $l_i p_i(1-p_i)(1+(\theta_i/(1+\theta_i))(l_i - 1))$. We relate these parameters to the necropsy time, t_i, and initial larvae burden, l_i, through the following functions chosen to optimise the fit to the data

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + \beta_1 \log(t_i) + \beta_2 \log(t_i)^2 \\ \log(\theta_i) &= \begin{cases} \eta_{100}, & \text{if } l_i \approx 100 \\ \eta_{200}, & \text{if } l_i \approx 200 \end{cases} . \end{aligned}$$

We also consider an alternative auxiliary model based on a three component Binomial mixture. This model was chosen purposefully as it still provides a reasonable description of the data but does not fit the data as well as the Beta Binomial model. The ith observation has the density

$$f(m_i|\Theta) \quad = \binom{l_i}{m_i} \sum_{k=1}^{3} w_k (\theta_i^k)^{m_i} (1-\theta_i^k)^{l_i - m_i},$$

where $w_3 = 1 - w_1 - w_2$. We reparameterise the θ_i^k, $\text{logit}(\theta_i^k) = \gamma_0^k + \gamma_1 \log(t_i)$, so that each component has the same slope but a different intercept. Therefore this model has six parameters, $\Theta = (w_1, w_2, \gamma_0^1, \gamma_0^2, \gamma_0^3, \gamma_1)$. The Beta-Binomial model provides a more optimal fit, with an improvement of about 170 points in the loglikelihood using one less parameter. From Figure 1 it is clear that the Beta-Binomial is explaining more variability. Furthermore, the Beta-Binomial simulations are spread across the range of observed matures while the mixture simulations are 'clumpy'.

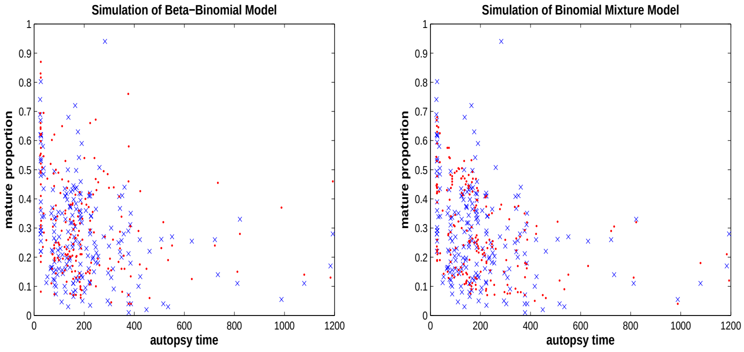
Our main investigation is to determine whether using the mixture auxiliary model leads to inefficient parameter estimates of the stochastic process model compared to when using the Beta-Binomial auxiliary model. Or, are inferences sensitive to the choice of the auxiliary model and how much effort should be spent on finding a well-fitting indirect model?

3 ABC using Indirect Inference

We consider a sequential Monte Carlo ABC (Sisson et al 2007) algorithm to sample from the sequence of targets

$$\pi(\theta, x|\rho(y,x) \le \epsilon_t) \quad \propto \quad f(x|\theta)\pi(\theta)1_{\rho(y,x)\le\epsilon_t} \text{ for } t = 1, \ldots, T.$$

FIGURE 1. A typical simulation from the Beta-Binomial (left) and the Binomial mixture (right) models. A cross denotes observed and a dot denotes simulated.



Our ABC algorithm is based upon the SMC ABC replenishment algorithm of Drovandi and Pettitt (2010). Here N particles are traversed through the sequence of target distributions. The algorithm determines the sequence of tolerances adaptively by dropping a proportion, α, of the particles with the highest discrepancy value. The population is replenished by resampling from the remaining particles. Diversity is ensured by moving the particles according to an MCMC kernel invariant for the current target. The proposal distribution of this MCMC step is also updated dynamically. We iterate the MCMC kernel sufficiently to ensure that each particle gets moved with a theoretical probability of $1 - c$ (with c set small). Our algorithm differs from Sisson et al (2009) and Beaumont et al (2009) since they use a forward kernel and also require pre-specification of the sequence of tolerances.

ABC with indirect inference requires an extra step. After data are simulated from the model, an auxiliary model is fitted to the data. The parameter estimates of this auxiliary model become the simulated summary statistics, θ_a^x, which are then compared to the observed summary statistics, $\hat{\theta}_a$.

4 Results

We inferred the parameters of the stochastic model successfully using the indirect inference approach with both auxiliary models. In the ABC algorithm we used $N = 1000$ particles, dropped half the particles with the worst discrepancy, $\alpha = 0.5$, and iterated the MCMC kernel so that theoretically 99% of the particles are moved, $c = 0.01$. For both cases the process was stopped when the MCMC kernel had about a 3% acceptance rate.

Parameter summaries of the Markov process model when applying each of the auxiliary models is presented in Table 1. Unfortunately the parameters μ_I and β are imprecisely estimated. This occurred since only mature counts are available and the immunity variable in simulations mostly takes a value

TABLE 1. Posterior summaries. Shown are the posterior mode, mean, standard deviation and the (2.5%,50%,97.5%) quantiles. BB = Beta-Binomial indirect model and BM = Binomial mixture indirect model. † estimates for these parameters have been multiplied by 100.

model	param	mode	mean	std dev	(2.5%,50%,97.5%)
BB	ν†	0.13	0.13	0.03	(0.07,0.13,0.20)
BB	μ_I	1.08	1.03	0.47	(0.15,1.02,1.88)
BB	μ_L†	0.55	0.85	0.60	(0.04,0.73,2.35)
BB	β	1.34	1.20	0.44	(0.34,1.22,1.96)
BM	ν†	0.08	0.11	0.04	(0.05,0.11,0.22)
BM	μ_I	1.05	1.03	0.46	(0.20,1.03,1.89)
BM	μ_L†	2.44	2.07	0.68	(0.37,2.22,3.05)
BM	β	1.03	1.18	0.43	(0.40,1.17,1.95)

no higher than 1 and is short lived. This meant that the parameters were sensitive to the prior but the ratio μ_I/β was relatively less sensitive.

The parameters ν and μ_L were precisely estimated. It can be seen from the Table that the posterior summaries for ν were similar regardless of which auxiliary model was applied, with a smaller variance when the Beta-Binomial model was used. The posterior summaries for μ_L were more dependent on the indirect model. The posterior for μ_L when using the Beta-Binomial distribution is shifted, tighter and is skew to the right compared with the posterior when the Binomial mixture is used.

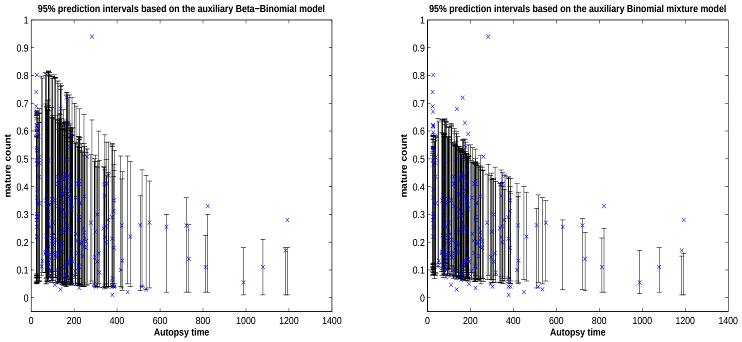
To compare the results from the two auxiliary models we produced predictions of the Markov process model based on the posterior modes in Table 1. We approximated 95% prediction intervals based on each auxiliary model, shown in Figure 2. It is clear that predictions from the stochastic model using Beta-Binomial auxiliary estimates account for more variability in the data. However, it appears that the 'clumpiness' of the Binomial mixture fit does not cause any problems as the stochastic model cannot predict such an effect. However, the most important summary would seem to be the range of the data at each time point, which the Beta-Binomial model explains better than the Binomial mixture model.

Acknowledgments: The authors would like to thank Edwin Michael and David Denham for access to the data. The authors also wish to thank Chris Glasbey and Steven Riley.

References

Beaumont, M. A., Cornuet, J-M, Marin, J-M and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, **96**, 983-990.

FIGURE 2. 95% predictions intervals based on the posterior modes of the stochastic model when applying the Beta-Binomial (left) and the Binomial mixture (right) as auxiliary models.



Denham, D. A., Ponnudurai, T., Nelson, G. S., Guy, F. and Rogers, R. (1972). Studies with *Brugia pahangi.* I. Parasitological observations on primary infections of cats (Felis catus). macroparasite models. *International Journal for Parasitology*, **2**, 239-247.

Drovandi, C. C. and Pettitt, A. N. (2010). Estimation of Parameters for Macroparasite Population Evolution using Approximate Bayesian Computation. *To appear in Biometrics.*

Gillespie, D. T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, **81**, 2340-2361.

Heggland, K. and Frigessi, A. (2004). Estimating functions in indirect inference. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **66**, 447-462.

Riley, S., Donnelly, C.A. and Ferguson, N.M. (2003). Robust parameter estimation techniques for stochastic within-host macroparasite models. *Journal of theoretical biology*, **225**, 419-430.

Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 1760-1765.

Sisson, S. A., Fan, Y. and Tanaka, M. M. (2009). Correction for Sisson et al., sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* , doi:10.1073/pnas.0908847106.