Mathematical Biosciences 000 (2016) 1–12

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation

Theodore Kypraios[a,*], Peter Neal[b], Dennis Prangle[c]

[a] School of Mathematical Sciences, University of Nottingham, UK
[b] Department of Mathematics and Statistics, Lancaster University, UK
[c] School of Mathematics and Statistics, Newcastle University, UK

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Bayesian inference
Epidemics
Stochastic epidemic models
Approximate Bayesian Computation
Population Monte Carlo

ABSTRACT

Likelihood-based inference for disease outbreak data can be very challenging due to the inherent dependence of the data and the fact that they are usually incomplete. In this paper we review recent Approximate Bayesian Computation (ABC) methods for the analysis of such data by fitting to them stochastic epidemic models without having to calculate the likelihood of the observed data. We consider both non-temporal and temporal-data and illustrate the methods with a number of examples featuring different models and datasets. In addition, we present extensions to existing algorithms which are easy to implement and provide an improvement to the existing methodology. Finally, R code to implement the algorithms presented in the paper is available on https://github.com/kypraios/epiABC.

1. Introduction

The past two decades have seen a significant growth in the field of mathematical modelling of communicable diseases and this has led to a substantial increase in our understanding of infectious-disease epidemiology and control. Understanding the spread of communicable infectious diseases is of great importance in order to prevent major future outbreaks and therefore it remains high on the global scientific agenda, including contingency planning for the threat of a possible influenza pandemic. The main purpose of this paper is to give an introduction and overview of some of the recent work concerned with Approximate Bayesian Computation methods for performing (approximate) Bayesian inference for stochastic epidemic models given data on outbreaks of infectious diseases. In addition, we present novel modifications to the existing algorithms and show that such modifications can be more efficient than the existing state-of-the-art algorithms. In the present section we discuss generic ideas with the bulk of the remainder of the paper containing various algorithms and illustrative examples.

1.1. Models and inference for epidemic models

It has been widely recognised that mathematical and statistical modelling has become a valuable tool in the analysis of infectious disease dynamics by supporting the development of control strategies, informing policy-making at the highest levels, and in general playing a fundamental role in the fight against disease spread [30].

The transmissible nature of infectious diseases makes them fundamentally different from non-infectious diseases, and therefore the analysis of disease outbreak data cannot be tackled using standard statistical methods. This is mainly due to the fact that the data are i) highly dependent and ii) incomplete, in many different ways since the actual transmission process is not directly observed. However, it is often possible to construct simple stochastic models which describe the key features of how an infectious disease spreads in a population. The complexity of the models typically varies depending on the application in question as well as the data available. For example, models may incorporate a latent period during which individuals are infected but not yet infectious, reduced infectivity after control measures are imposed, etc. Similarly, aspects of the population heterogeneity can also be included such as age structure and that individuals live in households and go to workplaces, etc.

Models can then be fitted to data either within a classical or Bayesian framework to draw inference on the parameters of interest that govern transmission. In turn these parameters can be used to provide useful information about quantities of clinical or epidemiological interest. One needs always to strike a balance between model complexity and data availability. In other words, it is not wise to construct a fairly complicated model when not much data are available and vice versa.

* Corresponding author.
E-mail address: theodore.kypraios@nottingham.ac.uk (T. Kypraios).

http://dx.doi.org/10.1016/j.mbs.2016.07.001
0025-5564/

1.2. Bayesian inference

In frequentist inference, model parameters are regarded as fixed quantities. On the other hand, a Bayesian approach treats all the unknown model parameters as random variables, enabling us to quantify the uncertainty of our estimates in a coherent, probabilistic manner. The Bayesian paradigm to inference operates by first assigning to the parameters a *prior distribution* which represents our belief about the unknown parameters (θ) before seeing any data. Subsequently this prior information is updated in the light of experimental data (D) using Bayes theorem by multiplying it with the likelihood $\pi(D|\theta)$ and renormalising, thus leading to the posterior distribution $\pi(\theta|D)$ via:

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\int_\theta \pi(D|\theta)\pi(\theta)\mathrm{d}\theta} \propto \pi(D|\theta)\pi(\theta). \quad (1)$$

All Bayesian inference arises from the posterior distribution in the sense that $\pi(\theta|D)$ contains all the information regarding our knowledge about the parameters θ given the experimental data D and any prior knowledge which might be available. Point and interval summaries of the posterior distribution (such as mean, median and credible intervals) can easily be obtained. The advantage of a Bayesian approach as opposed to a frequentist inference is that the former enables the complete quantification of our knowledge about the unknown parameters in terms of a probability distribution. We highlight such advantages in subsequent Sections.

1.3. Approximate Bayesian Computation

The main task in Bayesian statistics is to derive the posterior distribution of the parameters given the data $\pi(\theta|D)$. For many models the likelihood of observed data $\pi(D|\theta)$ is costly to compute and in other cases the observed data are insufficient to write down a tractable likelihood. However, provided that it is possible to simulate from the model, then "implicit" methods such as Approximate Bayesian Computation (ABC) allows us to perform inference without having to compute the likelihood.

We have already mentioned above that one of the difficulties when fitting models to disease outbreak data is that the infection process is unobserved. The likelihood of the observed data can become very difficult to evaluate and so is the posterior distribution. This is particularly the case when analysing temporal data, since calculating the likelihood involves integration over all possible infection times, which is rarely analytically possible. On the other hand, simulating realisations from a stochastic epidemic model is relatively straightforward. Therefore, ABC algorithms are very well suited to make inference for the parameters of epidemic models based on partially observed data and this has been illustrated when both temporal [39] and non-temporal data [40] are available.

1.4. Other approaches to inference

One way to overcome this issue is to employ data imputation methods where unknown quantities (such as the infection times) are treated as additional model parameters and inferred along with the other parameters. One of the most widely used methods for doing so is Markov Chain Monte Carlo (MCMC) which have revolutionised not only Bayesian statistics, but have also been developed for fitting stochastic epidemic models to partially observed outbreak data [28,43]. Despite being successfully applied to a wide variety of applications such as Foot-and-Mouth [19,33,51], SARS outbreaks [38], healthcare-associated infections (such as MRSA and C. difficile) [26,34] and Avian, H1N1 and H3N2 influenza [17,18,31] as the population size increases and/or the model becomes more sophisticated, the likelihood can become prohibitively costly to compute. In addition, non-standard and problem-specific MCMC algorithms need to be designed to improve on the efficiency of the standard algorithms.

The remainder of the paper is structured as follows. In Section 2, we introduce the ABC algorithm including extensions to ABC-MCMC and sequential based ABC-PMC. In Section 3, we apply the ABC algorithm to non-temporal (final outcome) data, firstly to a homogeneously mixing SIR epidemic model and secondly to a household SIR epidemic model. For the latter we introduce a new partially coupled ABC algorithm which offers significant gains in efficiency. In Section 4, we turn to the analysis of temporally observed epidemic data, in particular, the effective implementation of adaptive ABC-PMC algorithms.

2. ABC algorithms

Intuitively, ABC methods involve simulating data from the model using various parameter values and making inference based on which parameter values produced realisations that are "close" to the observed data. Algorithm 1 generates *exact* samples from the Bayesian posterior density $\pi(\theta|D)$ as defined in (1).

Algorithm 1 Exact Bayesian Computation (EBC).

Input: observed data D, parameters governing $\pi(\theta)$
Output: samples from $\pi(\theta|D)$

1: Sample θ^* from $\pi(\theta)$.
2: Simulate dataset D^* from the model using parameters θ^*.
3: Accept θ^* if $D^* = D$, otherwise reject.
4: Repeat until the required number of posterior samples is obtained.

This algorithm is only of practical use if D is discrete, else the acceptance probability in Step 3 is zero. For continuous distributions, or discrete ones in which the acceptance probability in step 3 is unacceptably low, [46] suggested the following algorithm:

Algorithm 2 Approximate Bayesian Computation (vanilla ABC).

Input: observed data D, tolerance ϵ, distance function $d(\cdot,\cdot)$, summary statistics $s(\cdot)$, parameters governing $\pi(\theta)$
Output: samples from $\tilde{\pi}(\theta|D) = \pi(\theta|D, d(s(D), s(D^*)) \leq \varepsilon)$

1: Sample θ^* from $\pi(\theta)$.
2: Simulate dataset D^* from the model using parameters θ^*.
3: Accept θ^* if $d\big(s(D), s(D^*)\big) \leq \varepsilon$, otherwisereject.
4: Repeat until the required number of posterior samples is obtained.

Here, $d(\cdot,\cdot)$ is a distance function, usually taken to be the L^2-norm of the difference between its arguments; $s(\cdot)$ is a function of the data; and ε is a tolerance. Note that $s(\cdot)$ can be the identity function but in practice, to give tolerable acceptance rates, it is often the case that it is a lower-dimensional vector comprising summary statistics that characterise key aspects of the data. In addition, if the prior and the posterior distribution are rather different, for example, in the case where the data are very informative about the model parameters then the rejection sampling approach of this ABC algorithm will be very inefficient. A wide range of extensions to the original ABC (which is often termed *vanilla* ABC) algorithm have been developed over the past decade and it still remains a topic of significant research interest.

2.1. Summary statistics

As discussed above, using $s(\cdot)$ as the identity function gives an inefficient ABC algorithm if the data is high dimensional. The underlying reason is a *curse of dimensionality* issue. Roughly speaking,

Please cite this article as: T. Kypraios et al., A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation, Mathematical Biosciences (2016), http://dx.doi.org/10.1016/j.mbs.2016.07.001