Computational Statistics and Data Analysis 71 (2014) 434–447

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Simulation-based Bayesian inference for epidemic models

Trevelyan J. McKinley [a,*], Joshua V. Ross [b], Rob Deardon [c], Alex R. Cook [d]

[a] Disease Dynamics Unit, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

[b] School of Mathematical Sciences, The University of Adelaide, Adelaide, Australia

[c] Department of Mathematics and Statistics, University of Guelph, Guelph, Canada

[d] Saw Swee Hock School of Public Health, Department of Statistics and Applied Probability, and Duke-NUS Graduate Medical School Singapore, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 27 June 2012
Received in revised form 23 November 2012
Accepted 19 December 2012
Available online 9 January 2013

Keywords:
Bayesian inference
Epidemic models
Markov chain Monte Carlo
Pseudo-marginal methods
Smallpox

ABSTRACT

A powerful and flexible method for fitting dynamic models to missing and censored data is to use the Bayesian paradigm via data-augmented Markov chain Monte Carlo (DA-MCMC). This samples from the joint posterior for the parameters and missing data, but requires high memory overheads for large-scale systems. In addition, designing efficient proposal distributions for the missing data is typically challenging. Pseudo-marginal methods instead integrate across the missing data using a Monte Carlo estimate for the likelihood, generated from multiple independent simulations from the model. These techniques can avoid the high memory requirements of DA-MCMC, and under certain conditions produce the exact marginal posterior distribution for parameters. A novel method is presented for implementing importance sampling for dynamic epidemic models, by conditioning the simulations on sets of validity criteria (based on the model structure) as well as the observed data. The flexibility of these techniques is illustrated using both removal time and final size data from an outbreak of smallpox. It is shown that these approaches can circumvent the need for reversible-jump MCMC, and can allow inference in situations where DA-MCMC is impossible due to computationally infeasible likelihoods.

1. Introduction

Mathematical models of infectious disease dynamics are useful tools to help explore the biological mechanisms of disease spread and to provide predictive information to guide the implementation of control policies and interventions (see e.g. Bailey, 1975; Keeling and Rohani, 2008). A common way to model epidemic systems is to consider that individuals progress through different epidemiological states over time. A simple example for a single epidemic of a disease such as influenza is an $\mathcal{SIR}$ model, in which individuals are classified as either susceptible to infection ($\mathcal{S}$), infected and infectious ($\mathcal{I}$), or removed ($\mathcal{R}$; corresponding to recovered and immune, or dead). A functional form is then chosen to describe the movements of individuals between states, governed by a set of epidemiological parameters. Due to the inherently stochastic nature of infectious disease outbreaks, we eschew deterministic approximations in favor of fully stochastic models, in which state transitions are governed by sets of probability equations. Hence, multiple realizations of the system will result in a distribution of outcomes, even for a fixed set of parameter values (i.e. with no parameter uncertainty). Therefore the observed data are one realization of a stochastic process, the dynamics of which we are attempting to explore using the chosen model.

* Corresponding author. Tel.: +44 1223 337685.
E-mail address: tjm44@cam.ac.uk (T.J. McKinley).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.12.012

To ensure that the outputs from the model can be interpreted robustly, it is vital to account for *parameter* uncertainty, as well as stochasticity arising from the model dynamics. Various techniques exist in order to fit dynamic models to data (see e.g. Bailey, 1975; Weirman and Marchette, 2004; Ionides et al., 2006; Cook et al., 2007; Höhle and Feldmann, 2007; Yang et al., 2007; Keeling and Ross, 2008; Jewell et al., 2009; Chis Ster et al., 2009; Deardon et al., 2010; Wong et al., 2013), many of which use a likelihood function to quantify the propensity of a given model and set of parameters to explain the observed data. However, the likelihood function can be difficult to calculate in practice, particularly when data are missing or incomplete. Although techniques exist to generate maximum likelihood estimates of dynamic temporal epidemic systems when data are missing/censored (e.g. Ionides et al., 2006), since it is often useful to supplement case time-series data with other forms of information—on the incubation period, say—here we use the Bayesian paradigm.

Readers unfamiliar with the Bayesian framework are referred to many excellent texts available, such as those by Gilks et al. (1996) and Gelman et al. (2004). This framework treats all parameters and variables as random, and the aim is to estimate the *posterior distribution* for the unknown parameters, $\boldsymbol{\theta}$, given the observed data, $\boldsymbol{D}$, written as $f(\boldsymbol{\theta}|\boldsymbol{D}) \propto f(\boldsymbol{D}|\boldsymbol{\theta}) f(\boldsymbol{\theta})$, up to some normalizing constant, where $f(\boldsymbol{\theta})$ represents our *prior* knowledge about the parameters, and $f(\boldsymbol{D}|\boldsymbol{\theta})$ is the likelihood. The normalizing constant is often difficult to evaluate, and so we resort to numerical estimation methods such as Markov chain Monte Carlo (MCMC; e.g. Gilks et al., 1996) or Sequential Monte Carlo (SMC; e.g. Doucet et al., 2001). The techniques discussed in this paper relate directly to the former, and in particular are linked to the Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970).

The Bayesian framework offers a natural environment to parameterize epidemic systems, since missing/censored data can simply be included as extra parameters in the model. One implementation of this approach is through data-augmented MCMC (DA-MCMC; Gibson and Renshaw, 1998; O'Neill and Roberts, 1999), which, particularly when coupled with reversible-jump (RJ) methodology (Green, 1995), is perhaps the most flexible computational technique currently available for fitting dynamic epidemic models to data. However, implementation of DA-MCMC can be challenging, particularly in defining efficient proposal distributions for the missing data. For large amounts of missing data, it may be necessary to update each missing value, or subsets of the missing values, in turn. Furthermore, it may also be necessary to track the full history of each augmented variable. This can lead to large memory requirements for high-dimensional problems and highly autocorrelated chains. A recent paper by Andrieu et al. (2010) uses SMC methods to build efficient high-dimensional proposals for use in MCMC. Known as particle MCMC, this method has the potential to be widely applicable for inference in many epidemiological problems. However, in this paper we focus on an alternative method, based on using information from multiple repeated simulations instead of direct evaluation of the likelihood function. This idea goes back at least to Diggle and Gratton (1984), who approximate the log-likelihood through simulation, and use this to develop a numerical approximation routine for performing maximum likelihood calculations.

A general technique – based on these ideas – that is growing in popularity in various scientific fields is Approximate Bayesian Computation (ABC). For a given parameter value, multiple simulations from the model are produced and the proportion that 'match' the observed data are used to provide an estimate of the likelihood. This basic idea can be incorporated into rejection sampling (e.g. Tavaré et al., 1997; Beaumont et al., 2002), MCMC (e.g. Marjoram et al., 2003; Wilkinson, submitted for publication) or SMC routines (e.g. Sisson et al., 2007; Toni et al., 2009; Beaumont et al., 2009; Erhardt and Smith, 2012). In practice the requirement to match the observed and simulated data exactly is relaxed, and instead some metric, $\rho(\cdot)$, is defined that characterizes the distance between the observed and simulated data sets. Simulations then 'match' if $\rho(\cdot)$ is less than some tolerance ϵ. This introduces three areas of approximation: the choice of metric, tolerance and the number of simulations used to produce the approximate Monte Carlo estimate. In McKinley et al. (2009), ABC techniques were employed to produce approximate posterior estimates for the parameters of a temporal epidemic model, both with and without missing data. The authors showed that it was possible to produce simple metrics that provided accurate estimates of the true posterior (relative to the gold-standard of DA-MCMC) in the case where there is negligible missing data. However, they showed that the accuracy of the approximation begins to break down when the amount of missing data increases. Although these techniques are potentially useful to provide estimates of parameter uncertainty in complex models for which it is difficult to calculate a likelihood, it is not always clear how to define a metric sensibly, or decide on a suitable value for the tolerance. Questions also remain as to the impact of these choices on what the approximate posterior distribution actually represents (see e.g. Wilkinson, submitted for publication), although a recent paper by Fearnhead and Prangle (2012) made some exciting developments in terms of re-casting ABC as an inferential framework in its own right, as opposed to simply approximating the true posterior. The reader is also encouraged to see Tanaka et al. (2006), Blum and Tran (2010) and Neal (2010) for other applications of ABC in epidemic modeling. Nonetheless, some of these complexities motivate the interest here to explore alternative simulation methods.

Pseudo-marginal approaches (see e.g. O'Neill et al., 2000; Beaumont, 2003; Andrieu and Roberts, 2009) are based on importance sampling. O'Neill et al. (2000) employ a so-called Monte Carlo within Metropolis (MCWM) algorithm to analyze epidemiological models based on household outbreak data. Beaumont (2003) introduces a similar algorithm called grouped-independence Metropolis–Hastings (GIMH) to analyze genealogical data. The convergence properties of both MCWM and GIMH are explored more theoretically in Andrieu and Roberts (2009), where the general moniker of 'pseudo-marginal approaches' is applied to cover both cases. Although they are based on a similar central concept, GIMH can be shown to produce an *exact* marginal posterior for the parameters, despite the use of a Monte Carlo (MC) estimate for the likelihood (Beaumont, 2003; Andrieu and Roberts, 2009). MCWM produces an approximation, though we show in Section 4 that this approximation is good for the sorts of applications discussed here. Similar techniques have been implemented with some