Spatial and Spatio-temporal Epidemiology 24 (2018) 27–37

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Original Research

Approximate Bayesian computation for spatial SEIR(S) epidemic models

Grant D. Brown[a,*], Aaron T. Porter[b], Jacob J. Oleson[a], Jessica A. Hinman[c]

[a] Department of Biostatistics, University of Iowa, Iowa City, Iowa 52242 USA
[b] Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, Colorado 80401 USA
[c] Department of Epidemiology, University of Iowa, Iowa City, Iowa 52242 USA

ARTICLE INFO

Article history:
Received 18 January 2017
Revised 6 October 2017
Accepted 14 November 2017
Available online 22 November 2017

Keywords:
Approximate Bayesian computation
Compartmental Models
Chikungunya
Epidemics

ABSTRACT

Approximate Bayesia n Computation (ABC) provides an attractive approach to estimation in complex Bayesian inferential problems for which evaluation of the kernel of the posterior distribution is impossible or computationally expensive. These highly parallelizable techniques have been successfully applied to many fields, particularly in cases where more traditional approaches such as Markov chain Monte Carlo (MCMC) are impractical. In this work, we demonstrate the application of approximate Bayesian inference to spatially heterogeneous Susceptible-Exposed-Infectious-Removed (SEIR) stochastic epidemic models. These models have a tractable posterior distribution, however MCMC techniques nevertheless become computationally infeasible for moderately sized problems. We discuss the practical implementation of these techniques via the open source ABSEIR package for R. The performance of ABC relative to traditional MCMC methods in a small problem is explored under simulation, as well as in the spatially heterogeneous context of the 2014 epidemic of Chikungunya in the Americas.

1. Introduction

The study of epidemics is complicated by the fact that real human populations exhibit complex structure and interact in subtle ways over both space and time. Nevertheless, in an increasingly globalized world, the ability to model pathogen outbreaks, predict ongoing spread, and evaluate interventions represents crucial abilities of public health practitioners. In this work, we present a class of algorithms and statistical framework ideally suited to meet this need, in addition to a discussion of our open source software, ABSEIR, which implements them.

1.1. Approximate Bayesian computation

Approximate Bayesian Computing is generally attributed to the work of Rubin (1980), which concerns interpretation and implementation of practical modeling techniques for applied Bayesian statisticians. Among other contributions, this work introduced one of the most commonly used algorithmic approaches to ABC: the rejection algorithm. This procedure provides an intuitive introduction to approximate Bayesian computing techniques. We therefore begin our approach to the subject by introducing the requisite notation, and describing the basic ABC rejection algorithm.

Define a $p \times 1$ parameter vector $\boldsymbol{\theta}$ with p dimensional parameter space $\boldsymbol{\Theta}$ and prior distribution $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. Further define an $N \times 1$ vector of observed data, $\mathbf{y}$, with a likelihood or data generating distribution denoted by $f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})$. Finally, define a distance function (such as the Euclidean

* Corresponding author.
E-mail address: grant-brown@uiowa.edu (G.D. Brown).

https://doi.org/10.1016/j.sste.2017.11.001
1877-5845/

distance) between appropriately sized vectors $\mathbf{x}$ and $\mathbf{y}$: $\rho(\mathbf{y}, \mathbf{x})$. As a Bayesian sampling technique, the goal of ABC is to make inference about the posterior distribution, $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathbf{Y}) \propto f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$.

The general pattern of rejection sampling ABC is quite simple. We first generate repeated samples $\boldsymbol{\theta}_i$ from the prior distribution for $\boldsymbol{\theta}$. Each of these samples, indexed by i, is in turn used to generate a replicate data set $\mathbf{x}_i$ from the likelihood. Parameters that generate replicate data sets that are sufficiently 'close' to the observed data $\mathbf{y}$, according to the distance function ρ and a tolerance ϵ, are retained, while the rest are discarded. Details of this procedure are given in Algorithm 1.

Algorithm 1 ABC rejection algorithm.

Require: Define a tolerance $\epsilon > 0$, and let '$\leftarrow$' denote assignment

1: **for** $i \leftarrow 1$ to n **do**
2: $d \leftarrow \infty$
3: **while** $d > \epsilon$ **do**
4: draw $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\Theta})$
5: draw $\mathbf{x}_i \sim f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})_{\mathbf{i}}$
6: $d \leftarrow \rho(\mathbf{y}, \mathbf{x}_i)$

Note that this approach does not require the user to evaluate the potentially expensive or unavailable likelihood function, but does require the ability to draw samples from it (Beaumont, 2010; Rubin, 1980). In its original formulation, the tolerance, ϵ, was taken to be zero (Rubin, 1980). The key insight of the rejection approach is clear in this context: accepting only parameters that produce replicate data identical to the observed response is equivalent to conditioning on that observed data. The distribution of parameter values conditional on the observed data is the posterior distribution: our inferential target. The most commonly applied version of the algorithm, however, generally includes the aforementioned nonzero tolerance, and employs a distance measure which depends only on a set of summary statistics of $\mathbf{x}$ and $\mathbf{y}$, thus rendering the inference 'approximate'.

1.2. Sequential algorithms

Numerous improvements and extensions have been proposed to this basic algorithm, generally focusing on obtaining increased sampling efficiency. In particular, many authors note that sampling performance can be extremely poor in situations where prior distributions on the parameter vector $\boldsymbol{\theta}$ are diffuse with respect to the posterior distribution, especially for high dimensional problems (Beaumont, 2010; Beaumont et al., 2009; Blum and François, 2010; Del Moral et al., 2012; Neal and Huang, 2015; Sisson et al., 2007). Sun et al. (2015) apply several such improvements in the context of non-spatial deterministic and stochastic compartmental epidemic models. Here, we emphasize a single algorithm, though the software described in later sections is the focus of ongoing research in this area. We implement a slightly modified version of the sequential Monte Carlo algorithm proposed by Beaumont et al. (2009), which we find both intuitive and effective. As with the rejection algorithm, Beaumont et al. (2009) begin by drawing proposed parameters from their prior distribution. Instead of repeating this step, however, subsequent sets of parameters are re-sampled and then perturbed from previously accepted values according to a set of weights. Data is then simulated as before, and parameters are accepted according to a decreasing sequence of ϵ values. Weights are updated using an importance sampling step to preserve the target posterior distribution. This approach can provide dramatic efficiency gains over the rejection algorithm.

Our adaptation of this algorithm introduces four primary modifications. First, we employ a batch size, $N \geq n$, over which simulations and distance evaluations may be conducted in parallel with no need for communication between nodes. This is important, because even with the sequential parameter updates, acceptances can become quite rare as ϵ decreases. Second, we permit the first iteration to employ a larger batch size than subsequent sequential step. This ensures that the algorithm starts at a practical ϵ, rather than spending too much time at unnecessarily permissive tolerances. Third, we implement a specific ϵ schedule: $\epsilon_{t+1} = c\epsilon_t$, where $0 < c \leq 1$. This obviates the need for investigators to manually specify a sequence of ϵ values, a process that depends on the scale of observed values as well as the chosen set of prior distributions. Finally, we generalize the perturbation kernel to permit a multivariate Gaussian distribution. For problems that exhibit correlation among the parameters, we have found the multivariate approach can be more efficient.

These modifications imply two potential modes of convergence, beyond specifying a required terminating ϵ value. First, the investigator may specify a desired number of sampling epochs. Second, users may choose to specify a maximum number of batches of size N which will execute for a particular value of ϵ before the sampler will simply return the current sample of n parameter values. This latter mode enables the algorithm to adapt the termination of the sequence of ϵ values to the difficulty of sampling by specifying a termination acceptance rate. This acceptance rate is generally chosen based on the computational resources available. Background on the development of sequential Monte Carlo ABC is available in Beaumont (2010).

1.3. Model selection in SMC-ABC

Beyond the ability to fit models that would be otherwise computationally infeasible, ABC techniques provide a natural way to compare the relative evidence for different models. Informally, sets of parameters and models that produce better simulated data are more probable than others. This can be used to compare a set of candidate models, and in fact the ratio of acceptance rates between two models is an estimate of the Bayes Factor comparing the two (Beaumont, 2010).

In the SMC-ABC context, however, such comparisons are a bit more problematic. Care must be taken to employ comparable instantiations of the algorithm. For example, comparison between non-converged and converged