Identifying the Infectious Period Distribution for Stochastic Epidemic Models Using the Posterior Predictive Check

Muteb Alharthi[1], Philip O'Neill[1], Theodore Kypraios[1]

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

[1] School of Mathematical Sciences, University of Nottingham, UK
pmxma16@nottingham.ac.uk, Philip.ONeill@nottingham.ac.uk,
Theodore.Kypraios@nottingham.ac.uk

Abstract

Under the Bayesian framework, we develop a novel method to assess the goodness of fit for the SIR (Susceptible → Infective → Removed) stochastic epidemic model. This method seeks to determine whether or not one can identify the infectious period distribution based only on a set of partially observed data using a posterior predictive distribution approach. Our criterion for assessing the model's goodness of fit is based on the notion of Bayesian residuals.

Keywords: Epidemic models; Goodness of fit; Posterior predictive distribution; Bayesian residual.

1 Introduction

Poor fit of a statistical model to data can result in suspicious outcomes and misleading conclusions. Although the area of parameter estimation for stochastic epidemic models has been a subject of considerable research interest in recent years (see e.g.[1], [2] and [3]), more work is needed for the model assessment in terms of developing new methods and procedures to evaluate goodness of fit for epidemic models. Therefore, it is of importance to seek a method to assess the quality of fitting a stochastic epidemic model to a set of epidemiological data. The most well-known stochastic model for the transmission of infectious diseases is considered, that is the SIR (Susceptible - Infective - Removed) stochastic epidemic model. We recall methods of Bayesian inference using Markov chain Monte Carlo (MCMC) techniques for the SIR model where partial temporal data are available. Then, a new simulation-based goodness of fit method is presented. This method explores whether or not the infectious period distribution can be identified based on removal data using a posterior predictive model checking procedure.

2 Model, Data and Inference

We consider a Susceptible-Infective-Removed (SIR) stochastic epidemic model in which the infection rate at time t is given by $\beta n^{-1} X(t) Y(t)$, where $X(t)$ and $Y(t)$ represent the number of susceptible and infective individuals at t in a closed homogeneous population

of size $\mathcal{N} = n + 1$, which consists of n initial susceptibles and one initial infective, and β denotes the infection rate parameter. Following [4] and [5], let $f_{T_I}(\cdot)$ denote the probability density function of T_I (the length of infectious period, which is assumed to be a continuous random variable) and let θ indicate the parameter governing T_I. Also, define $\mathbf{I} = (I_1, ..., I_{n_I})$ and $\mathbf{R} = (R_1, ..., R_{n_R})$, where I_j and R_j are the infection and removal times of individual j and where it is assumed that the total number of infections and removals are equal, that is $n_I = n_R$. Assuming a fully observed epidemic (complete data) with the initial infective labelled κ such that $I_\kappa < I_j$ for all $j \neq \kappa$, the likelihood of the data given the model parameters is

$$L(\mathbf{I}, \mathbf{R}|\beta, \theta, \kappa) = \left(\prod_{j=1, j\neq\kappa}^{n_I} \beta n^{-1} Y(I_j-) \right) \cdot \exp\left(-\beta n^{-1} A\right) \cdot \prod_{j=1}^{n_R} f_{T_I}\left(R_j - I_j\right),$$

where $A = \sum_{j=1}^{n_I} \sum_{k=1}^{\mathcal{N}} (R_j \wedge I_k - I_k \wedge I_j)$ with $I_k = \infty$ for $k = n_I + 1, ..., \mathcal{N}$.

Unfortunately, incomplete data (where we observe only removal times) are the most common type of epidemic data. As a result, the likelihood of observing only the removal times given the model parameters is intractable. One solution to make the likelihood tractable is to use the data augmentation technique by treating the missing data as extra (unknown) parameters [1]. For instance, let $T_I \sim Exp(\gamma)$, where γ is referred to as the removal rate. By adopting a Bayesian framework and assigning conjugate gamma prior distributions for the model parameters [1], that are $\beta \sim Gamma(\lambda_\beta, \nu_\beta)$, (with mean $= \lambda_\beta/\nu_\beta$) and $\gamma \sim Gamma(\lambda_\gamma, \nu_\gamma)$, we get the following marginal posterior distributions:

$$\beta|\gamma, \mathbf{I}, \mathbf{R} \sim Gamma\left(\lambda_\beta + n_I - 1, \nu_\beta + n^{-1}A\right),$$

$$\gamma|\beta, \mathbf{I}, \mathbf{R} \sim Gamma\left(\lambda_\gamma + n_R, \nu_\gamma + \sum_{j=1}^{n_R} (R_j - I_j)\right),$$

as well as

$$\pi(\mathbf{I}|\beta, \gamma, \mathbf{R}) \propto \left(\prod_{j=1, j\neq\kappa}^{n_I} Y(I_j-) \right) \cdot \exp\left(-\beta n^{-1} A\right) \cdot \prod_{j=1}^{n_R} \exp\left(-\gamma(R_j - I_j)\right).$$

The model parameters β and γ can be updated using Gibbs sampling steps as they have closed form of the posterior distributions. However, the infection times need to be updated using a Metropolis-Hastings step. Having done that, we can obtain samples from the marginal posterior distributions of the model parameters.

3 Methodology

We are concerned with identifying the infectious period distribution of the SIR model based only on removal data. In the SIR stochastic epidemic model, regardless of the type of infectious period distribution (we consider Exponential, Gamma and Constant), the total population size is constant and satisfies $\mathcal{N} = X(t) + Y(t) + Z(t)$, where $Z(t)$ denotes the number of removed individuals at event time t with $X(0) \geq 1, Y(0) \geq 1$ and $Z(0) = 0$; note that $Z(s) \leq Z(t)$ for any $0 \leq s \leq t; s, t \geq 0$.
However, due to the fact that epidemic data are partially observed it is sufficient for our purpose to consider only the times when removals occur instead of looking at all event times. Assuming that all infected individuals are removed by the end of the epidemic,

the behaviour of the three models in terms of $Z(r_1), Z(r_2), ...$, differs, where r_j represents the j-th removal time.
We turn our attention to take advantage of this difference to distinguish between these three models when fitting them to data in the case of partial observations. Let $\mathbf{R}^{obs}$ and $\mathbf{R}^{rep}$ denote the observed and replicated removal times respectively, then our proposed method can be generally described by the following algorithm.

Algorithm 1 Generic algorithm for our method

1. Given $\mathbf{R}^{obs}$, fit an SIR model using MCMC to get $\pi(\beta|\mathbf{R}^{obs})$ and $\pi(\theta|\mathbf{R}^{obs})$.
2. Draw $\beta^i \sim \pi(\beta|\mathbf{R}^{obs})$ and $\theta^i \sim \pi(\theta|\mathbf{R}^{obs})$, $i = 1, ..., M$.
3. Use β^i and θ^i to draw samples from $\pi(\mathbf{R}^{rep\,i}|\mathbf{R}^{obs})$ conditioning on $n_R^{rep\,i} = n_R^{obs}$.
4. Compare $\mathbf{R}^{obs}$ to $\pi(\mathbf{R}^{rep\,i}|\mathbf{R}^{obs})$.

4 Illustration

To illustrate our method, 92 removal times were simulated from an SIR model in which $T_I \sim Exp(0.5)$ and $\beta = 1.5$ in population of size $\mathcal{N} = 100$, that consists of $n = 99$ initial susceptibles and one initial infective, then our procedure was applied (see Figure 1). By looking at Figure 1, it is clearly noticeable that the observed data fit very well within the predictive distribution of the exponential SIR model, the model that has generated the data.

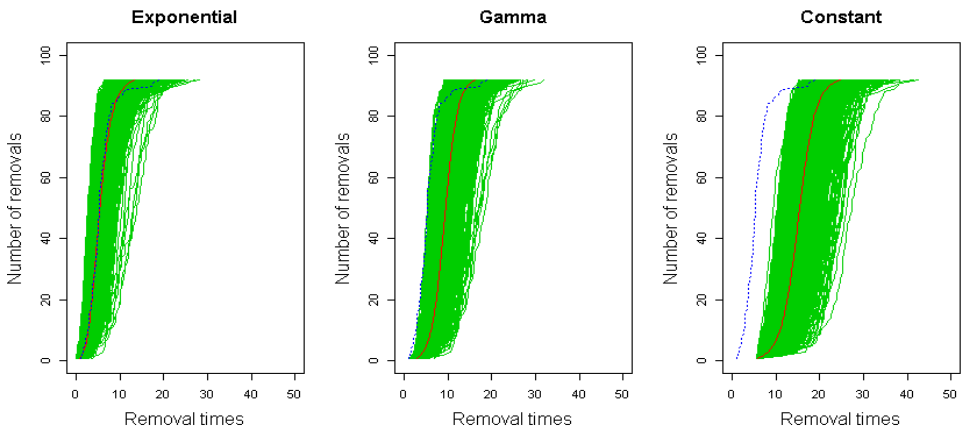

Figure 1: Comparison of the predictive distribution for the three models where dotted line indicates the observed data and solid line represents the predictive mean conditioning on the observed final size.

As mentioned above, our preferred criterion to measure the goodness of fit is the Bayesian residual [6], that is, conditioning on $n_R^{rep\,i} = n_R^{obs}$,

$$d_j = R_j^{obs} - E(R_j^{rep\,i}|\mathbf{R}^{obs}),\ j = 1, ..., n_R,$$

where $E(R_j^{rep\,i}|\mathbf{R}^{obs}) = \int R_j^{rep\,i}\, \pi(R_j^{rep\,i}|\mathbf{R}^{obs})\, dR_j^{rep\,i} \approx \frac{1}{M} \sum_{i=1}^{M} R_j^{rep\,i}$.

It is worth mentioning here that the quantity $\sum_{j=1}^{n_R} d_j^2$ could provide an overall measure of fit. Figure 2 shows the the Bayesian residual distributions for the three models in which it is qualitatively obvious that there is a high density accumulated near zero, coming from the exponential SIR model, compared to the other two models. On

the top of that, quantitatively, the sum of the squared Bayesian residuals $\sum_{j=1}^{n_R} d_j^2$ are $101.03, 1256.55$ and 8501.78 for the exponential, gamma and constant SIR models respectively. Therefore, as expected, the exponential SIR model, from which the data was generated, has the smallest value.

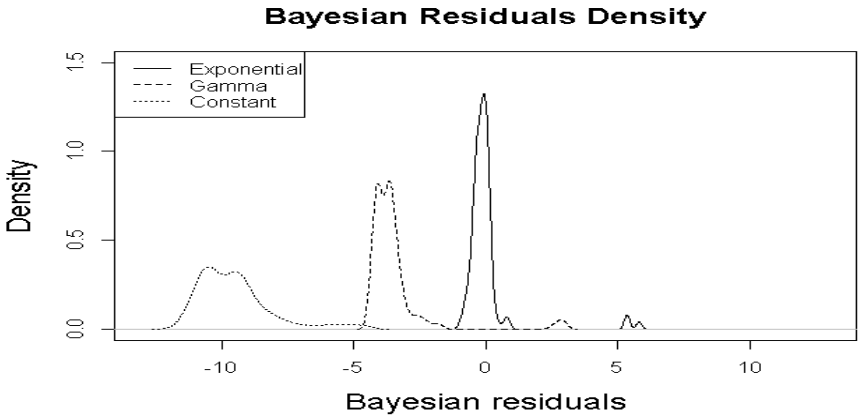

Figure 2: The Bayesian residual distributions of the three SIR models.

5 Conclusion

Bayesian inference for the SIR model has been introduced where the epidemic outbreak is partially observed. We have proposed a method to assess the goodness of fit for the SIR stochastic model based only on removal data. A simulation study has been performed to test the proposed method. Using the posterior predictive assessment for checking models, this diagnostic method is seen to identify the true model reasonably well.

References

[1] O'Neill, P.D., Roberts, G.O. (1999). "Bayesian inference for partially observed stochastic epidemics." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**(1), 121–129.

[2] Andersson, H.,& Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Lecture Notes in Statistics. Springer, New York.

[3] Kypraios, T., (2007). " Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models", PhD Thesis, Lancaster University.

[4] Britton, T., O'Neill, P.D. (2002). "Bayesian inference for stochastic epidemics in populations with random social structure." *Scandinavian Journal of Statistics*, **29**(3), 375–390.

[5] Neal, P., Roberts, G.O. (2005). "A case study in non-centering for data augmentation: stochastic epidemics." *Stat. Comput.*, **15**(4), 315–327.

[6] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. 2nd edition. CRC press.