

Chapter 9

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

Muteb Alharthi, Philip O’Neill, and Theodore Kypraios

Abstract Under the Bayesian framework, we develop a novel method for assessing 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.

Key words: SIR, Epidemic models, MCMC, Predictive distribution, Bayesian residual

9.1 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, 7, 9]), 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 for assessing 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

M. Alharthi (✉) • P. O’Neill • T. Kypraios
School of Mathematical Sciences, University of Nottingham, Nottingham, UK
e-mail: pmxma16@nottingham.ac.uk; philip.oneill@nottingham.ac.uk;
heodore.kypraios@nottingham.ac.uk

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.

9.2 Model, Data and Inference

We consider a SIR stochastic epidemic model [2] in which the rate of new infections 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 [3, 5], let $f_{T_I}(\cdot)$ denote the probability density function of T_I (the length of the infectious period, which is assumed to be a continuous random variable) and let θ indicate the parameter governing T_I . Also, define $\mathbf{I} = (I_1, \dots, I_{n_I})$ and $\mathbf{R} = (R_1, \dots, R_{n_R})$, where I_j and R_j are the infection and removal times of individual j and where we shall assume, for simplicity, that the total number of infections and removals are equal, that is $n_I = n_R = m$ (this assumption can be relaxed, see [8] for the details). Assuming a fully observed epidemic (complete data) with the initial infective labelled z such that $I_z < I_j$ for all $j \neq z$, the likelihood of the data given the model parameters is

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

where $A = \sum_{j=1}^m \sum_{k=1}^{\mathcal{N}} (R_j \wedge I_k - I_k \wedge I_j)$ with $I_k = \infty$ for $k = m + 1, \dots, \mathcal{N}$. Here, I_j^- denotes the time just prior to I_j and R_{j-} is defined similarly.

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 [8]. For instance, let $T_I \sim \text{Exp}(\gamma)$, where γ is referred to as the removal rate. By adopting a Bayesian framework and assigning conjugate gamma prior distributions to the model parameters [8] that are $\beta \sim \text{Gamma}(\lambda_\beta, \nu_\beta)$, (with mean = $\lambda_\beta / \nu_\beta$) and $\gamma \sim \text{Gamma}(\lambda_\gamma, \nu_\gamma)$, we get the following full conditional posterior distributions:

$$\begin{aligned} \beta | \gamma, \mathbf{I}, \mathbf{R} &\sim \text{Gamma}(\lambda_\beta + m - 1, \nu_\beta + n^{-1} A), \\ \gamma | \beta, \mathbf{I}, \mathbf{R} &\sim \text{Gamma}\left(\lambda_\gamma + m, \nu_\gamma + \sum_{j=1}^m (R_j - I_j)\right), \end{aligned}$$

as well as

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

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.

When the length of the infectious periods is assumed to be constant, we have two model parameters to be updated, namely the mean of the infectious period $E(T_I) = c$ and the infection rate parameter β . However, if we let the infectious periods to have a gamma distribution $Gamma(\alpha, \delta)$, in addition to estimating the infection rate parameter β , we shall assume for computational reasons that the gamma shape parameter α is known (although it can be considered as unknown parameter to be estimated from the data, see [6] for the details) and the scale parameter δ is unknown and has to be estimated using MCMC output.

9.3 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 $Z(r_1), Z(r_2), \dots$, differs, where r_j represents the j -th removal time.

We turn our attention to taking advantage of this difference to distinguish between these three models when fitting them to data in the case of partial observations. Let $\mathbf{R}^{obs} = (R_1^{obs}, \dots, R_m^{obs})$ and $\mathbf{R}^{rep} = (R_1^{rep}, \dots, R_m^{rep})$ denote the observed and replicated removal times, respectively, and also let $\pi(\mathbf{R}^{rep} | \mathbf{R}^{obs})$ represent the removal times predictive distribution. Then our proposed method can be generally described by the **Algorithm 1**.

Step 3 in the **Algorithm 1** can be done simply by keeping simulating (until the desired sample size is obtained) from the model using the model parameter posterior distributions while rejecting simulations that do not match the observed final size.

Algorithm 1 Generic algorithm for our method

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

9.4 Illustration

To illustrate our method, 93 removal times were simulated from an SIR model in which $T_I \sim Exp(0.5)$ and $\beta = 1.5$ in a population of size $\mathcal{N} = 100$, that consists of $n = 99$ initial susceptibles and one initial infective.

Throughout the analysis, uninformative gamma prior distributions with parameters $\lambda_\beta = \lambda_\gamma = \lambda_\delta = 1$ and $\nu_\beta = \nu_\gamma = \nu_\delta = 0.001$ were set to the parameters of the SIR models and it was assumed that the gamma shape parameter, when fitting the SIR model with gamma infectious period $T_I \sim Gamma(\alpha, \delta)$ is known ($\alpha = 10$).

By looking at Fig. 9.1, 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.

As mentioned above, our preferred criterion to measure the goodness of fit is the Bayesian residual [4], that is, conditioning on $m^{rep i} = m^{obs}$,

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

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 $\sum_{j=1}^m d_j^2$ could provide an overall measure of fit. Figure 9.2 shows 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 top of that, quantitatively, the sum of the squared Bayesian residuals $\sum_{j=1}^m d_j^2$ are 96.3, 354.7 and 812.6 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 of the sum of the squared Bayesian residuals.

9.5 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

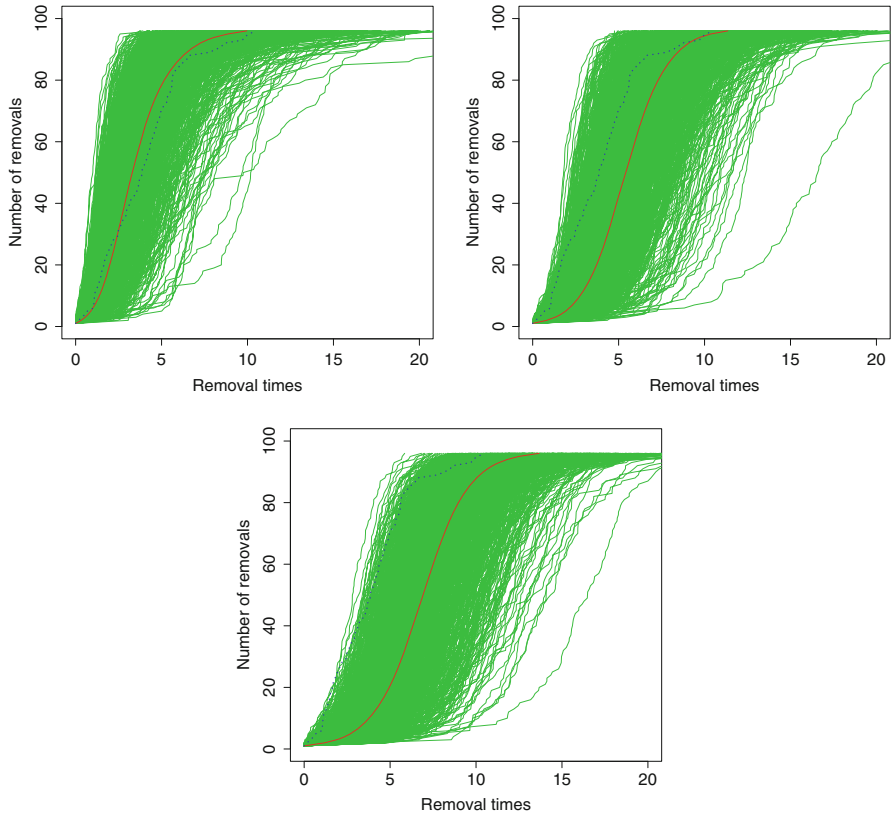


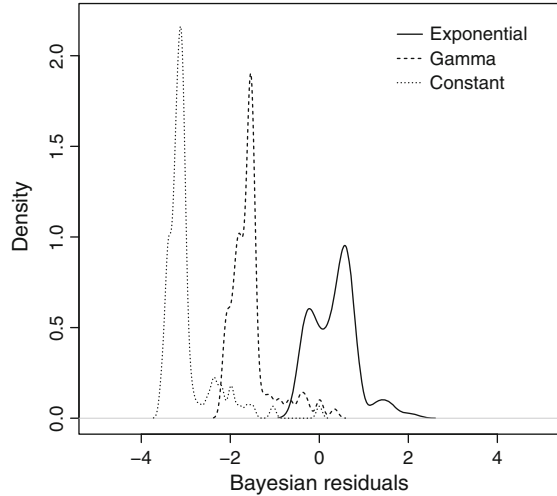
Fig. 9.1 Comparison of the removal times predictive distribution for the three SIR models (*top left*: Exponential, *top right*: Gamma, *bottom*: Constant) based on 1,000 realizations and conditioning on the observed final size, where the *dotted line* indicates the observed data and the *solid line* represents the predictive mean

has been performed to test the proposed method. Using the posterior predictive assessment for checking models, this diagnostic method is able to identify the true model reasonably well.

One advantage of this method is that it looks explicitly at the discrepancy between observed and predicted data, which avoids using unobserved quantities in the process of assessment, see [10] as an example. Furthermore, this method is still valid when including an extra state, the exposed period, to the SIR model in which individuals in this state are infected but not yet infectious.

Acknowledgements The first author is supported by a scholarship from Taif University, Taif, Saudi Arabia.

Fig. 9.2 The Bayesian residual distributions for each SIR model based on 1,000 samples from the conditioning predictive distribution for the three models



References

- [1] Andersson, H., Britton, T.: *Stochastic Epidemic Models and Their Statistical Analysis*, vol. 4. Springer, New York (2000)
- [2] Bailey, N.T.J.: *The Mathematical Theory of Infectious Diseases and Its Applications*. Charles Griffin & Company, London (1975)
- [3] Britton, T., O'Neill, P.D.: Bayesian inference for stochastic epidemics in populations with random social structure. *Scand. J. Stat.* **29**(3), 375–390 (2002)
- [4] Gelfand, A.E.: Model determination using sampling-based methods. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (eds.) *Markov Chain Monte Carlo in Practice*, pp. 145–161. Springer, New York (1996)
- [5] Kypraios, T.: *Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models*. Ph.D. thesis, Lancaster University (2007)
- [6] Neal, P., Roberts, G.O.: A case study in non-centering for data augmentation: stochastic epidemics. *Stat. Comput.* **15**(4), 315–327 (2005)
- [7] O'Neill, P.D.: Introduction and snapshot review: relating infectious disease transmission models to data. *Stat. Med.* **29**(20), 2069–2077 (2010)
- [8] O'Neill, P.D., Roberts, G.O.: Bayesian inference for partially observed stochastic epidemics. *J. Roy. Stat. Soc. Ser. A (Stat. Soc.)* **162**(1), 121–129 (1999)
- [9] Streftaris, G., Gibson, G.J.: Bayesian inference for stochastic epidemics in closed populations. *Stat. Model.* **4**(1), 63–75 (2004)
- [10] Streftaris, G., Gibson, G.J.: Non-exponential tolerance to infection in epidemic systems—modeling, inference, and assessment. *Biostatistics* **13**(4), 580–593 (2012)