Bayesian Analysis (2023) **18**, Number 4, pp. 1283–1310

Posterior Predictive Checking for Partially Observed Stochastic Epidemic Models

Georgios Aristotelous*, Theodore Kypraios† and Philip D. O'Neill‡

Abstract. We address the problem of assessing the fit of stochastic epidemic models to data. Two novel model assessment methods are developed, based on disease progression curves, namely the distance method and the position-time method. The methods are illustrated using SIR (susceptible-infective-removed) models. We assume a typical data observation setting in which case-detection times are observed while infection times are not. Both methods involve Bayesian posterior predictive checking, in which the observed data are compared to data generated from the posterior predictive distribution. The distance method does this by calculating distances between disease progression curves, while the position-time method does this pointwise at suitably selected time points. Both methods provide visual and quantitative outputs with meaningful interpretations. The performance of the methods benefits from the development and application of a time-shifting method that accounts for the random time delay until an epidemic takes off. Extensive simulation studies show that both methods can successfully be used to assess the choice of infectious period distribution and the choice of infection rate function.

Keywords: epidemic model, infectious disease data, posterior predictive p-value.

MSC2020 subject classifications: Primary 62F15, 62P10; secondary 62-08.

1 Introduction

Stochastic epidemic models are vital public health tools for understanding and predicting progression of an outbreak and formulating control strategies. However, such models have limited practical use unless they sufficiently represent key features of real-life outbreaks. Although there has been significant progress in the area of parameter estimation for stochastic epidemic models (e.g. O'Neill and Roberts, 1999; Kypraios, 2007; Streftaris and Gibson, 2012; Xiang and Neal, 2014; Nguyen-Van-Yen et al., 2021), model assessment methods remain less developed. This paper is concerned with the problem of assessing the fit of stochastic epidemic models, fitted to partially observed temporal outbreak data. We employ a Bayesian framework and develop methods based on the notion of posterior predictive checking, whereby one or more aspects of a model are examined for departure from their posterior predictive distribution (Gelman et al., 2013). Note that our sole focus is assessing whether or not a proposed epidemic model adequately describes the data to hand, rather than choosing between a collection of competing models as addressed by Alharthi et al. (2019).

*Department of Social Statistics and Demography, University of Southampton, UK, g.aristotelous@soton.ac.uk

†School of Mathematical Sciences, University of Nottingham, UK, theodore.kypraios@nottingham.ac.uk

‡School of Mathematical Sciences, University of Nottingham, UK, philip.oneill@nottingham.ac.uk

 https://doi.org/10.1214/22-BA1336

Analysing data from infectious disease outbreaks contains inherent challenges, since typically data are not independent, the infection process is unobserved, and an outbreak is only realized once. This in turn makes standard methods for assessing model fit inapplicable. An unsurprising consequence is that the literature on model assessment for stochastic epidemic models is rather limited. In the Bayesian setting, existing approaches can broadly be divided into two categories. In the first, the posterior distribution of a set of stochastic residuals is assessed for consistency against the reference sampling distribution (e.g. Jewell et al., 2009; Streftaris and Gibson, 2012). The second is based on posterior predictive checking (e.g. Gardner et al., 2011; Parry et al., 2014). Although these approaches have proved useful as a means of excluding models on the basis of their inability to reproduce key aspects of observed epidemics, they have considerable drawbacks and have failed to provide model assessment methods that are well-established among the epidemic community. Approaches based on residuals rely heavily on information imputed from the model itself, thus reinforcing the model being assessed (Gibson et al., 2018) and the choice of which residuals to use can be somewhat arbitrary (O'Neill, 2010). Posterior predictive checking approaches have not been employed to their full potential, often assessing only non-temporal features of a model (e.g. Lekone and Finkenstädt, 2006), or when assessing temporal features, failing to consider important peculiarities of the epidemic modelling setting, a point we return to in Section 3.

In this paper we focus on posterior predictive checking methods based on disease progression curves which represent the temporal features of an outbreak. We develop two novel computational methods, namely the distance method and the position-time method. Both methods provide visual and quantitative outputs with meaningful interpretations. The methods greatly benefit from the application of a time-shifting intervention which we also develop. This intervention adjusts for a peculiar feature of stochastic epidemics, related to the randomness of their take-off time, and facilitates better conditions for model assessment. Throughout this paper, we assume a partial observation setting that commonly occurs in practice, where case-detection times are observed while actual times of infection are unknown, though the methods we develop are in principle applicable to any temporal data setting. We demonstrate the performance of our methods via extensive simulation studies that address two specific problems, namely the choice of infectious period distribution and the choice of infection process mechanism in SIR (susceptible-infective-removed) models. In addition to simulated data, the methods are also illustrated on real data via two influenza outbreak examples.

This paper is structured as follows. Section 2 contains relevant preliminary information. Section 3 motivates and describes the time-shifting method we employ. In Sections 4 and 5, respectively, we develop and illustrate our two methods for model assessment, and present extensive simulation studies in Sections 6 and 7. In Section 8 we apply the methods to real data. Finally, in Section 9 we give some additional perspectives and commentary on the work in this paper. All computer code used to produce the results in this paper was conducted using the statistical programming language R Core Team (2019). The code to implement the methods has been deposited in Github (https://github.com/kypraios/post-pred-check-epi).

2 Preliminaries

2.1 Posterior predictive checking

Posterior predictive checking (Gelman et al., 2013) is an intuitive and natural way to assess model fit within a Bayesian framework. The key idea is that replicated data, generated under the model, should look similar to the observed data. Let $\boldsymbol{y}^{obs}$ denote the observed data, $\pi(\boldsymbol{y} \mid \boldsymbol{\theta})$ the sampling density of an assumed model with parameter $\boldsymbol{\theta}$ and $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}^{obs})$ the posterior density of $\boldsymbol{\theta}$. Then replicated data $\boldsymbol{y}^{rep}$ are data generated from the *posterior predictive density* of the model, denoted $\pi(\boldsymbol{y}^{rep} \mid \boldsymbol{y}^{obs})$, and given by

$$\pi(\boldsymbol{y}^{rep} \mid \boldsymbol{y}^{obs}) = \int \pi(\boldsymbol{y}^{rep} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y}^{obs})d\boldsymbol{\theta}. \tag{2.1}$$

Since direct comparison of $\boldsymbol{y}^{obs}$ and $\boldsymbol{y}^{rep}$ can be complicated unless both are low-dimensional, it is instead common to focus on real-valued test statistics. A test statistic T assumes its observed value $T^{obs} := T(\boldsymbol{y}^{obs})$ whereas $T^{rep} := T(\boldsymbol{y}^{rep})$ is a random variable. Assessment for the aspect of the data represented by T is conducted by comparing the posterior predictive distribution of T^{rep} to its observed value T^{obs}.

In practice, and for all models considered in this paper, the posterior and the posterior predictive distributions are not known analytically and so we instead conduct assessment via samples. A sample from the posterior predictive distribution of T^{rep} can be obtained as follows. First, draw a sample $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(S)}\}$ from the posterior density of the model, $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}^{obs})$, using a method such as Markov chain Monte Carlo (MCMC). For each posterior value $\boldsymbol{\theta}^{(s)}$, $s = 1, 2, \ldots, S$, simulate a replicated dataset $\boldsymbol{y}^{rep(s)}$ from the sampling density of the model $\pi(\boldsymbol{y} \mid \boldsymbol{\theta}^{(s)})$. It follows from equation (2.1) that $\{\boldsymbol{y}^{rep(1)}, \boldsymbol{y}^{rep(2)}, \ldots, \boldsymbol{y}^{rep(S)}\}$ is a sample from the posterior predictive distribution of the model. Finally, calculate $T^{rep(s)} := T(\boldsymbol{y}^{rep(s)})$, for each $s = 1, 2, \ldots, S$, to produce a sample, $\{T^{rep(1)}, T^{rep(2)}, \ldots, T^{rep(S)}\}$, from the posterior predictive distribution of T^{rep}.

Given $\{T^{rep(1)}, T^{rep(2)}, \ldots, T^{rep(S)}\}$, posterior predictive checking can be conducted both quantitatively and visually. Quantitatively, the quantity of interest is the *posterior predictive p-value* (ppp-value), which for continuous T is defined as ppp-value := $P(T^{rep} < T^{obs})$ and calculated as

$$\begin{aligned} \text{ppp-value} &= P(T^{rep} < T^{obs}) = \mathrm{E}(\mathbb{1}_{\{T^{rep} < T^{obs}\}}) \\ &= \int \mathbb{1}_{\{T^{rep} < T^{obs}\}} \pi(\boldsymbol{y}^{rep} \mid \boldsymbol{y}^{obs}) d\boldsymbol{y}^{rep} \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}_{\{T^{rep(s)} < T^{obs}\}}, \end{aligned} \tag{2.2}$$

where $\mathbb{1}_A$ denotes the indicator function of the event A. Extreme ppp-values, close to 0 or 1, imply evidence for lack of fit, in the sense that the discrepancy between model and data can not be reasonably explained by chance, whereas values near 0.5 indicate goodness of fit, suggesting that the model adequately captures the aspect of the data in question (Gilks et al., 1996; Gelman et al., 2013). Visually, the observed value T^{obs} can be imposed on a histogram of the sampled replicated values $\{T^{rep(1)}, T^{rep(2)}, \ldots, T^{rep(S)}\}$. An observed value near the middle of the histogram would indicate goodness of fit.

2.2 Standard and non-linear infection rate SIR models

Recall the *standard SIR model* (see e.g. Andersson and Britton, 2000), defined as follows. Consider a closed population consisting of individuals, each of which can, at any time t, be either susceptible (S), infective (I) or removed (R). Initially, N individuals are susceptible, m are infective and none are removed. A susceptible individual can potentially contract the disease. An infective individual has the disease and can transmit it to others. A removed individual can no longer transmit the disease and is also not susceptible. The removal state might correspond to different things in practice such as immunity, death, isolation or the appearance of symptoms which make the individual too ill to continue interacting with the population as usual. The common characteristic of all removed individuals is that they play no part in the spread of the epidemic. Individuals can only transition from susceptible to infective (S $\rightarrow$ I) and infective to removed (I $\rightarrow$ R). Let X_t and Y_t respectively be the numbers of susceptible and infective individuals in the population at time t. Transitions may occur as follows.

S $\rightarrow$ I: A given infective individual makes contacts with any other individual at the time points of a homogeneous Poisson process of rate β. If a contacted individual is susceptible they immediately become infective. All Poisson processes are assumed to be mutually independent. Thus the all-to-all infection rate is $\beta X_t Y_t$.

I $\rightarrow$ R: Upon infection an individual starts their infectious period, in which they remain until they become removed. The infectious periods of individuals are assumed to be independent and identically distributed according to a specified random variable T_D.

In the rest of this paper we assume for simplicity that $m = 1$ but this assumption can easily be relaxed. The epidemic ends when no infectives are left in the population. The total number of susceptibles that ever become infected is known as the *final size* of the epidemic.

A key parameter associated with the standard SIR model is the basic reproduction number R_0. This is loosely defined as the average number of new infections caused by a typical infective in a large susceptible population, and is given by

$$R_0 = N\beta \mathrm{E}(T_D). \tag{2.3}$$

The value of R_0 can be thought of as quantifying the infective potential of the disease, with more infections occurring overall as R_0 increases.

The *non-linear infection rate SIR model* (Severo, 1969; O'Neill and Wen, 2012) is an extension of the standard SIR model, obtained by modifying the all-to-all infection rate to be $\beta X_t Y_t^p$, for $p \in [0, 1]$. The parameter p controls the level of exposure of susceptibles to infectives, so the smaller the p the less the exposure.

In this paper, we consider two specific choices of infectious period distribution, namely Exponential, with rate parameter γ and probability density function (p.d.f.) $f(x; \gamma) = \gamma \exp(-\gamma x)$, $x \geq 0$; $\gamma > 0$, and Gamma, with shape parameter ν, rate parameter λ and p.d.f. $f(x;\ \nu, \lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} \exp(-\lambda x)$, $x \geq 0$; $\nu, \lambda > 0$. Both these choices appear frequently in the epidemic modelling literature. We will consider the standard SIR model using both Exponential and Gamma infectious periods and denote these

versions of the model by Exp-HM and Gamma-HM, respectively, where HM denotes Homogeneous Mixing. We also consider the non-linear infection rate SIR model with Exponential infectious period distribution, and denote this as Exp-NL.

2.3 Data assumptions

We shall assume we have data on a single epidemic outbreak in which case-detection times are observed but infection times are not. In addition, we assume that upon appearance of symptoms an individual stops interacting with the population so that case-detection times correspond to removal times. These assumptions are reasonable for many real-life settings, and are common in the modelling literature (e.g. O'Neill and Roberts, 1999; Neal and Roberts, 2005; Xiang and Neal, 2014).

We write $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$ to denote a vector of time-ordered removal times, so that $r_1 < r_2 < \ldots < r_n$. This may refer to observed removal times, or to replicated removal times generated under the posterior predictive distribution of a particular model. To distinguish these cases, we write $\boldsymbol{r}^{obs} = (r_1^{obs}, r_2^{obs}, \ldots, r_n^{obs})$ and $\boldsymbol{r}^{rep} = (r_1^{rep}, r_2^{rep}, \ldots, r_n^{rep})$ in the obvious manner, and write $\boldsymbol{r}^{rep(s)} = (r_1^{rep(s)}, r_2^{rep(s)}, \ldots, r_n^{rep(s)})$ to denote the realisation of $\boldsymbol{r}^{rep}$ corresponding to a specific replication s.

We impose a condition on replicated data, namely that they must be of equal dimension to the observed data, or equivalently, that they must have the same final size as the observed data. This is achieved by employing a rejection sampling algorithm (e.g. Ripley, 2009) via which we sample realisations from the fitted model and only keep those whose final size is the same as in the observed data. The ability of a model to reproduce the observed final size can be assessed separately, for example via posterior predictive checking as in Alharthi (2016, chapter 3). The motivation for this assumption is that it is more natural to compare vectors of the same dimension. It is however possible to relax this assumption, a point we return to in Section 9.

In what follows we assume that, given some data and a model, the model has already been fitted to the data and a sample from the posterior distribution has been obtained. A sample from the posterior predictive distribution of the model can be obtained as described in Section 2.1. We will consider a sample from the posterior predictive distribution of the model to be the starting point for the model assessment procedures we describe. To fit the models considered in this paper we use MCMC methods (O'Neill and Roberts, 1999). The MCMC algorithms used to fit these models, and their run conditions, are described in Aristotelous (2020, section 1.3.5) and Aristotelous (2020, section 2.7.1).

2.4 Removal curve

The *removal curve*, denoted z_t, is defined as the cumulative number of removals at time t, i.e. $z_t(\boldsymbol{r}) := \sum_{k=1}^{n} \mathbb{1}_{\{r_k \leq t\}}$. Removal curves are frequently used in the literature to assess disease progression dynamics, when only removal data are available (e.g. Gibson et al., 2018). In the context of posterior predictive checking, assessment based on removal

curves can be conducted visually by imposing the observed removal curve on a pack of replicated removal curves drawn from the posterior predictive distribution of a fitted model (e.g. Stockdale et al., 2017). In this paper we will additionally develop quantitative measures of fit based on removal curves.

Note that z_t is a function of time and is therefore potentially much more informative for model assessment purposes than scalar quantities such as the final size or duration of an outbreak. It is easy to see that, given $z_t(\boldsymbol{r})$ for each $t \in \mathbb{R}$, one can reconstruct the removal data vector $\boldsymbol{r}$, and thus, the removal curve statistic is another way of viewing the data. This property is what makes z_t the focus of the methods developed in this paper. An illustrative example, demonstrating the clear advantage that the removal curve has in assessing temporal dynamics, compared to the final size and the duration statistics, is provided in Aristotelous (2020, section 2.2.4). We use the notation z_t^{obs}, z_t^{rep} and $z_t^{rep(s)}$ to denote $z_t(\boldsymbol{r}^{obs})$, $z_t(\boldsymbol{r}^{rep})$ and $z_t(\boldsymbol{r}^{rep(s)})$, respectively.

3 Time-shifting of removal curves

3.1 Motivation

Many stochastic epidemic models, at least in large populations, have the feature of a "random time" until take-off, followed by a more-or-less deterministic phase (see e.g. Andersson and Britton, 2000; Aristotelous, 2020, section 2.4.1). In our setting, this means that different realisations of an SIR model may look like time-shifted versions of each other, which complicates any attempt to see if they appear similar to observed data. Moreover, the stochasticity of the take-off time can be large enough to make it hard to extract meaningful conclusions from any visual or quantitative assessment. To our knowledge, none of the existing approaches in the literature for model fit or assessment for epidemic models take random take-off time into account.

We demonstrate these issues visually in Figures 1a and 1c via two examples. For visual aid, the plots also show mean removal curves, defined formally in Section 4.1. The first example illustrates that high stochasticity in the take-off time can lead to a misspecified model being plausible for observed removal data. We generate event times from a homogeneous Poisson Process (HPP) with rate $\rho = 1$ during the time interval $[T_{\text{on}}, T_{\text{off}}] = [0, 170]$, treat these times as the removal data, and fit a Gamma-HM model (Figure 1a). Here the level of model misspecification is considerable, since the Gamma-HM model assumes that removals occur at a rate that varies as the epidemic progresses, while event times generated under a HPP occur at a constant rate. Despite this, the variability in take-off time makes the pack of replicated removal curves too wide for the misspecification to be apparent.

The second example shows undesired implications even when the epidemic model is correctly specified, even for large datasets. We fit the Gamma-HM model to data generated from itself (Figure 1c). The replicated removal curves have very similar shape to the observed curve, due to the correctly specified process dynamics. However, the observed curve is atypical with respect to the take-off time, and it therefore lies on the tails of the replicated pack, raising doubts for the fit of the model.

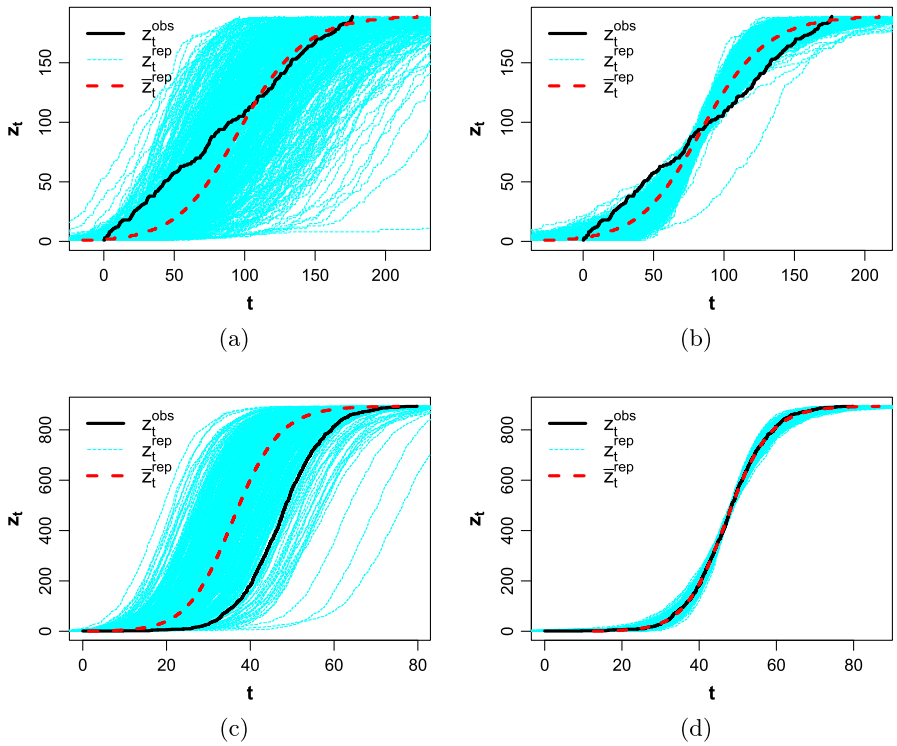

Figure 1: Plots of 500 replications from the posterior predictive distribution of the removal curve, z_t^{rep}, with its mean removal curve, $\bar{z}_t^{rep}$ (red, dashed line), and the observed removal curve, z_t^{obs} (black, solid line), imposed. In the top row Gamma-HM ($\nu = 10$) is fitted to data generated from a HPP ($\rho = 1$, $T_{\text{on}} = 0$, $T_{\text{off}} = 170$). In the bottom row Gamma-HM ($\nu = 10$) is fitted to data generated from the same model ($N = 1000$, $R_0 = 2.5$, $\nu = 10$, $\lambda = 1$). Left and right columns are without and with applying time-shifting, respectively.

3.2 Procedure and implementation

To deal with the issues related to the randomness of the take-off time, we propose an intervention, referred to as *time-shifting*. The purpose of this intervention is to remove the stochasticity of the take-off time, and thus allow a like-for-like comparison between removal curves. The idea is to shift each removal curve in order to minimize its distance from the observed removal curve, according to some specified distance function.

Given $c \in \mathbb{R}$, the removal vector $\boldsymbol{r} + c := (r_1 + c, r_2 + c, \ldots, r_n + c)$, is a shift by c time units of the removal vector $\boldsymbol{r}$, and the removal curve $z_t(\boldsymbol{r} + c)$ is the corresponding shift of the removal curve $z_t(\boldsymbol{r})$. Suppose that a model has been fitted to $\boldsymbol{r}^{obs}$ and that a sample $\{\boldsymbol{r}^{rep(1)}, \boldsymbol{r}^{rep(2)}, \ldots, \boldsymbol{r}^{rep(S)}\}$ has been obtained. The time-shifting is applied

as follows. For each replication $s = 1, 2, \ldots, S$, we calculate a shifting constant

$$c^{(s)} = \arg\min_{c \in \mathbb{R}} d(z_t(\boldsymbol{r}^{obs}), z_t(\boldsymbol{r}^{rep(s)} + c)),$$

where d is a distance function between removal curves, the choice of which is discussed in Section 4.1. We then replace each replication, $\boldsymbol{r}^{rep(s)}$, with its shifted counterpart, $\boldsymbol{r}^{rep(s)} + c^{(s)}$, $s = 1, 2, \ldots, S$. Algorithm 1 describes the steps in the time-shifting intervention.

Algorithm 1 Time-shifting.

Let $\boldsymbol{r}^{obs}$ be the observed removal data and $\{\boldsymbol{r}^{rep(1)}, \boldsymbol{r}^{rep(2)}, \ldots, \boldsymbol{r}^{rep(S)}\}$ a sample of replicated removal data, drawn from the posterior predictive density of the model, $\pi(\boldsymbol{r}^{rep} \mid \boldsymbol{r}^{obs})$.

1. Calculate $c^{(s)} = \arg\min_{c \in \mathbb{R}} d(z_t(\boldsymbol{r}^{obs}), z_t(\boldsymbol{r}^{rep(s)} + c))$, $s = 1, 2, \ldots, S$.
2. Apply the time-shifting to each replication by replacing $\boldsymbol{r}^{rep(s)}$ with $\boldsymbol{r}^{rep(s)} + c^{(s)}$, $s = 1, 2, \ldots, S$.

To illustrate the effect of time-shifting, we return to the two examples presented in Section 3.1. Figure 1b shows that for the example where the model is clearly misspecified, application of time-shifting gives increased power to detect the misspecification. After time-shifting, the pack of replicated removal curves becomes much narrower and the observed curve lies on the tails and even outside the pack. For the example of the correctly specified model, Figure 1d shows that time-shifting effectively removes the undesired noise around the take-off time of the observed removal curve and allows the ability of the model to reproduce the observed data dynamics to be revealed. Typically, when applying time-shifting, the observed removal curve is placed in the middle of the pack of replicated removal curves with the mean removal curve being on top of the observed. We note that, although our motivation for time-shifting of the removal curves is to overcome the random time until take-off, it also helps to visualise the ensemble of removal curves in the same way that alignment is typically used to visualise functional data in general (e.g. Ramsay and Silverman, 2005).

Moreover, an alternative method of time-shifting, which only considers the early stages of the outbreak, can be implemented and this is referred to as *theoretical shifting* in Aristotelous (2020). This approach is based on the result that the initial stage of a standard SIR model in a large population can be approximated by a branching process (Ball and Donnelly, 1995) and it works by essentially pinning all replicated curves so that the time until $\sqrt{N}$ individuals become removed is the same among them and that of the observed curve. Extensive simulation studies are presented in Aristotelous (2020, section 2.7.1) where the different approaches to time-shifting are compared under a range of different scenarios (e.g. when a model is correctly specified as well as when it is misspecified). The results reveal both approaches of time-shifting have the desirable effect under all cases of (mis)specification and that the theoretical shifting is inferior to the presented time-shifting above. Therefore, all methods developed in this paper make

use of the time-shifting method described above, while the theoretical shifting is not considered any further.

4 Distance method

4.1 Procedure

We now introduce the *distance method* for assessing model fit. Consider a real-valued statistic T_d on the space of removal curves L such that $T_d(z_t) = d(z_t, \mathbb{E}z_t^{rep})$, where d is a distance function on L and $\mathbb{E}z_t^{rep}$ is the mean of z_t^{rep}. The choice of d and definition of $\mathbb{E}z_t^{rep}$ are addressed below. Then $T_d(z_t^{obs}) = d(z_t^{obs}, \mathbb{E}z_t^{rep})$ is the distance of the observed removal curve z_t^{obs} from the mean $\mathbb{E}z_t^{rep}$. The test statistic $T_d(z_t^{rep}) = d(z_t^{rep}, \mathbb{E}z_t^{rep})$ is a random variable, having the posterior predictive distribution of replicated distance, which is the distance of z_t^{rep} from the mean $\mathbb{E}z_t^{rep}$. To simplify notation let $T_d^{obs} := T_d(z_t^{obs})$ and $T_d^{rep} := T_d(z_t^{rep})$. Model assessment can be conducted, quantitatively and visually, in the usual fashion of posterior predictive checking by calculating the tail-area probability $P(T_d^{rep} < T_d^{obs})$ and by overlaying T_d^{obs} on a histogram of sampled replicated values of T_d^{rep}. The idea is that if a model fits the data, z_t^{obs} must not be further from the mean $\mathbb{E}z_t^{rep}$ than z_t^{rep} is, i.e. T_d^{obs} must look plausible under T_d^{rep}.

In this paper, we set d as the Euclidean distance on L,

$$d(z_t, z_t^*) = \left(\int_{\min(r_1, r_1^*)}^{\max(r_n, r_n^*)} (z_t - z_t^*)^2 \, dt \right)^{\frac{1}{2}},$$

where $z_t, z_t^* \in L$, with corresponding time-ordered removal vectors $\boldsymbol{r}, \boldsymbol{r}^* \in \mathbb{R}^n$. Note that other choices of d can be considered, such as the area between removal curves or the Euclidean distance between removal vectors. A thorough comparison between these distances is conducted in Aristotelous (2020, section 2.7.1), with the distance method exhibiting similar performance under all choices.

To define the posterior predictive mean removal curve, $\mathbb{E}z_t^{rep}$, we follow Alharthi (2016). First, define $\mathbb{E}\boldsymbol{r}^{rep} := (\mathrm{E}(r_1^{rep}), \mathrm{E}(r_2^{rep}), \ldots, \mathrm{E}(r_n^{rep}))$, where

$$\mathrm{E}(r_k^{rep}) = \int r_k^{rep} \pi(r_k^{rep} \mid \boldsymbol{r}^{obs}) dr_k^{rep}, \quad k = 1, 2, \ldots, n.$$

Then define $\mathbb{E}z_t^{rep} := z_t(\mathbb{E}\boldsymbol{r}^{rep}) = \sum_{k=1}^{n} \mathbb{1}_{\{\mathrm{E}(r_k^{rep}) \leq t\}}$. It is easy to see that, as constructed above, $\mathbb{E}z_t^{rep} \in L$ and thus all distance calculations required to implement the distance method are well defined.

4.2 Folded ppp-value and the assumption of symmetry

The key idea in the distance method is to calculate distances from the posterior predictive mean removal curve $\mathbb{E}z_t^{rep}$. This section explains the precise involvement that $\mathbb{E}z_t^{rep}$ has in the procedure, the assumptions that are imposed, and how the procedure can be interpreted.

Consider a simplified version of our setting in which all random quantities are random variables. Suppose the observed data consist of a single real number $\boldsymbol{y}^{obs}$. Let $\boldsymbol{y}^{rep}$ be a random variable having the posterior predictive distribution of a proposed model that has been fitted to $\boldsymbol{y}^{obs}$. Then, using as test statistic the identity function $T(\boldsymbol{y}) = \boldsymbol{y}$, from equation (2.2) we have ppp-value $= P(\boldsymbol{y}^{rep} < \boldsymbol{y}^{obs})$. Next, define the *folded ppp-value* as fppp-value $:= P(|\boldsymbol{y}^{rep} - \mathrm{E}(\boldsymbol{y}^{rep})| < |\boldsymbol{y}^{obs} - \mathrm{E}(\boldsymbol{y}^{rep})|)$, its name coming from the fact that $|\boldsymbol{y}^{rep} - \mathrm{E}(\boldsymbol{y}^{rep})|$ is a fold of $\boldsymbol{y}^{rep}$ at its mean $\mathrm{E}(\boldsymbol{y}^{rep})$. Then, under the assumption that $\boldsymbol{y}^{rep}$ has a symmetric distribution, it is easy to see that fppp-value $= 2|$ppp-value $- 0.5|$. This simple relationship that connects the folded ppp-value with the ppp-value implies that the former can also be used as a sensible and interpretable measure of model fit. More specifically, folded ppp-values near 0 would indicate goodness of fit, since they correspond to ppp-values near 0.5, while extreme folded ppp-values near 1 would imply evidence of lack of fit, since they correspond to ppp-values near 0 or 1.

In settings such as those above, where all random quantities are random variables and the ppp-value is clearly defined by utilizing the natural order of $\mathbb{R}$, working with the folded ppp-value and requiring symmetry for the posterior predictive distribution is unnecessary. However, in spaces such as the space of removal curves L with no natural ordering, it is not immediately obvious how to define a ppp-value. Instead it appears more natural to extend the definition of the folded ppp-value, by replacing the absolute value distance on $\mathbb{R}$ with some distance function on that space, and by requiring the posterior predictive distribution to be symmetric, in some sense, around its mean. For the space L, the extended folded ppp-value is then $P\left(d\left(z_t^{rep}, \mathbb{E}z_t^{rep}\right) < d\left(z_t^{obs}, \mathbb{E}z_t^{rep}\right)\right)$ where, as in Section 4.1, d is our chosen distance function on L, and $\mathbb{E}z_t^{rep}$ a suitably defined mean of z_t^{rep}.

From the definition of T_d in Section 4.1, it is clear that the extended folded ppp-value coincides with the tail-area probability $P(T_d^{rep} < T_d^{obs})$ of a posterior predictive check using the statistic T_d. Thus this probability is actually a folded ppp-value and, under the assumption that the posterior predictive distribution of the removal curve is symmetric, it can be interpreted as explained above.

The connection between the ppp-value and the folded ppp-value in $\mathbb{R}$ arises in the context of observing a single realization, and the test statistic being the identity function. A subtle point to note is that when extending to L these conditions are not violated, since the observed data still consist of a single realization, and the test statistic is still the identity function, since the removal curve is equivalent to the observed data as explained in Section 2.4.

By construction, for the folded ppp-value $P(T_d^{rep} < T_d^{obs})$ to be interpretable, we require the posterior predictive distribution to be symmetric. In the space L, it is not obvious how to explicitly define or verify symmetry. Even if a definition was proposed, the posterior predictive distributions of our models do not have known closed forms and thus analytically checking for symmetry appears to be infeasible. For these reasons, we adopt a pragmatic approach and assess the assumption of symmetry by visually inspecting whether the mean removal curve lies in the center of the pack of sampled replicated removal curves.

4.3 Implementation

The implementation steps for the distance method are given in Algorithm 2. To illustrate the method, we use it to assess the fit of three models, the Exp-HM model, the Gamma-HM model with shape parameter fixed at $\nu = 10$, and the Exp-NL model with power parameter fixed at $p = 0.5$, all fitted to the same dataset, itself generated from the Exp-HM model with $N = 500$, $R_0 = 2.5$ and $\gamma = 0.1$. Our intention is to create three cases of model specification, namely correct specification, misspecification due to the infectious period, and misspecification due to the infection rate form. The Exp-HM model represents the case of correct specification, being fitted to data generated from itself. The Gamma-HM model has the same infection process as the Exp-HM but with a different infectious period. The Exp-NL model has the same infectious period specification as the Exp-HM, but a different infection process. Notice that, for $\nu = 1$ and for $p = 1$, the Gamma-HM model and the Exp-NL model, respectively, reduce to the Exp-HM model. Thus, fixing $\nu = 10$ and $p = 0.5$ when fitting these models, rather than

Algorithm 2 Distance method.

Let $\boldsymbol{r}^{obs}$ be the observed removal data and $\{\boldsymbol{r}^{rep(1)}, \boldsymbol{r}^{rep(2)}, \ldots, \boldsymbol{r}^{rep(S)}\}$ a sample of replicated removal data, drawn from the posterior predictive density of the model, $\pi(\boldsymbol{r}^{rep} \mid \boldsymbol{r}^{obs})$.

1. **Apply the time-shifting:** Apply Algorithm 1 and note that for all following steps $\{\boldsymbol{r}^{rep(1)}, \boldsymbol{r}^{rep(2)}, \ldots, \boldsymbol{r}^{rep(S)}\}$ refers to the sample of shifted replicated removal data and $\{z_t^{rep(1)}, z_t^{rep(2)}, \ldots, z_t^{rep(S)}\}$ refers to the corresponding sample of shifted replicated removal curves.
2. **Calculate the mean removal vector and the mean removal curve:** Calculate the expected value of each replicated removal time, under its marginal posterior predictive distribution, using Monte Carlo (MC) approximation as $\mathrm{E}(r_k^{rep}) = \int r_k^{rep} \pi(r_k^{rep} \mid \boldsymbol{r}^{obs}) dr_k^{rep} \approx \frac{1}{S}\sum_{s=1}^{S} r_k^{rep(s)} = \bar{r}_k^{rep}$, $k = 1, 2, \ldots, n$. Then the mean removal vector is approximated as $\mathbb{E}\boldsymbol{r}^{rep} := (\mathrm{E}(r_1^{rep}), \mathrm{E}(r_2^{rep}), \ldots, \mathrm{E}(r_n^{rep})) \approx (\bar{r}_1^{rep}, \bar{r}_2^{rep}, \ldots, \bar{r}_n^{rep}) =: \bar{\boldsymbol{r}}^{rep}$ and the mean removal curve as $\mathbb{E}z_t^{rep} := z_t(\mathbb{E}\boldsymbol{r}^{rep}) \approx z_t(\bar{\boldsymbol{r}}^{rep}) =: \bar{z}_t^{rep}$.
3. **Calculate the required distances:** Calculate the (approximate) observed value of T_d as $T_d^{obs} := T_d(z_t^{obs}) = d(z_t^{obs}, \mathbb{E}z_t^{rep}) \approx d(z_t^{obs}, \bar{z}_t^{rep})$ and obtain an (approximate) sample $\{T_d^{rep(1)}, T_d^{rep(2)}, \ldots, T_d^{rep(S)}\}$ from the posterior predictive distribution of $T_d^{rep} = T_d(z_t^{rep}) = d(z_t^{rep}, \mathbb{E}z_t^{rep})$, by calculating the (approximate) replicated distances as $T_d^{rep(s)} := d(z_t^{rep(s)}, \mathbb{E}z_t^{rep}) \approx d(z_t^{rep(s)}, \bar{z}_t^{rep})$, $s = 1, 2, \ldots, S$.
4. **Assess the model:** Assess the model quantitatively by calculating the folded ppp-value using MC approximation as $P(T_d^{rep} < T_d^{obs}) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}_{\{T_d^{rep(s)} < T_d^{obs}\}}$ and visually by inspecting the histogram of replicated distances $\{T_d^{rep(1)}, T_d^{rep(2)}, \ldots, T_d^{rep(S)}\}$ with T_d^{obs} imposed, as well as by inspecting the plot of replicated removal curves $\{z_t^{rep(1)}, z_t^{rep(2)}, \ldots, z_t^{rep(S)}\}$ with z_t^{obs} and $\bar{z}_t^{rep}$ imposed.

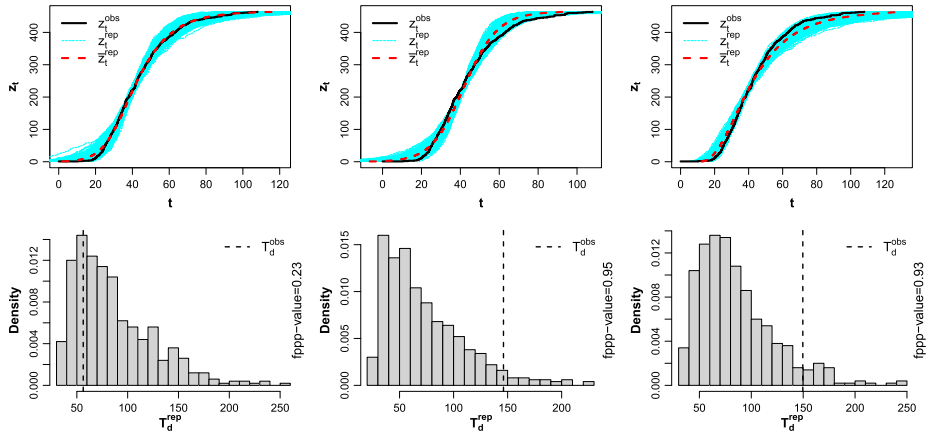

Figure 2: Example of posterior predictive checking using the distance method. Observed data are generated from an Exp-HM model ($N = 500$, $R_0 = 2.5$, $\gamma = 0.1$). Left, middle and right columns correspond to fitted model being the Exp-HM (correctly specified), the Gamma-HM ($\nu = 10$) (misspecified due to the infectious period) and the Exp-NL ($p = 0.5$) (misspecified due to the infection rate form), respectively. Top row are plots of 500 replications from the posterior predictive distribution of the removal curve, z_t^{rep}, with the mean removal curve, $\bar{z}_t^{rep}$ (red, dashed line), and the observed removal curve, z_t^{obs} (black, solid line), imposed. Bottom row are histograms of 500 replications from the posterior predictive distribution of the distance, T_d^{rep}, with the observed distance, T_d^{obs} (black, dashed line), overlaid and the corresponding folded ppp-value stated.

estimating these parameters, prevents the models from being reduced to the Exp-HM model, and thus creates the desired model misspecification.

Typical output of the assessment is given in Figure 2. The distance method behaves appropriately since it yields a small folded ppp-value, 0.23, for the Exp-HM model and large ones for the Gamma-HM and the Exp-NL models, 0.95 and 0.93 respectively. Based on its outputs, the distance method suggests that the Exp-HM fits the data, whereas the other two models are highly likely to be misspecified, which is indeed the case.

Looking at the removal curve plots (top row of Figure 2), it is evident that curves produced by the Gamma-HM model tend to peak faster, compared to curves produced by the Exp-HM, which in turn, tend to peak faster compared to curves produced by the Exp-NL model. This explains why in such cases, where data are generated from the Exp-HM model, the Gamma-HM and the Exp-NL models struggle to produce replicated curves that are similar to the observed curve. Based on the folded ppp-values, the distance method provides an accurate quantification of the differences in the removal curve behaviour of these models.

The results of this example are indicative of the distance method's ability to successfully assess both the infectious period assumption and the infection rate form as-

sumption of epidemic models. A much more detailed depiction of this ability is given in the extensive simulation studies of Sections 6 and 7.

5 Position-time method

5.1 Rationale and procedure

The distance method provides a quantitative measure of goodness of fit via the use of a distance function that integrates over time. In a stochastic process setting however, it is natural and informative to also consider methods that provide quantitative measures of fit that are functions of time. Below, we develop one such method, which we refer to as the *position-time method*. The method works by specifying the position of the observed curve, with respect to the pack of replicated curves, at the time points of an appropriately chosen time interval.

The procedure of the position-time method is as follows. Fix a time point t, such that $t \in [r_1^{obs}, r_n^{obs}]$. Specify the position of z_t^{obs}, with respect to the distribution of z_t^{rep}, by calculating the time-dependent (mid) ppp-value, given by ppp-value$(t) = P(z_t^{rep} < z_t^{obs}) + \frac{1}{2}P(z_t^{rep} = z_t^{obs})$. This modification from our definition of the ppp-value as given in equation (2.1) accounts for the fact that z_t^{rep} is a discrete random variable. Values closer to 0.5 provide indication for goodness of fit, as they correspond to the observed curve lying in the middle of the pack of replicated curves, and values near 0 or 1 provide evidence against the model's fit, as they correspond to the observed curve lying on the lower or upper tail of the pack of replicated curves, respectively.

Given the value of ppp-value(t) for every $t \in [r_1^{obs}, r_n^{obs}]$, there is flexibility for a range of visual and quantitative assessments. Visually, one can plot a histogram of time dependent ppp-values calculated at a collection of equally spaced time points of $[r_1^{obs}, r_n^{obs}]$; more mass near 0.5 would indicate better fit while more mass near 0 or 1 a bad fit. Another option, arguably more informative, is to plot the function ppp-value(t) against time; a good fit would be indicated in the cases that the function is consistently close to 0.5, while a lack of fit in the cases that the curve is consistently near 0 or 1. Quantitatively, statements for any interesting quantity, with respect to the posterior predictive distribution, can be made, by integrating the indicator function of the desired event over time. We particularly consider quantities such as the proportion of time that z_t^{obs} spends in a specified inverse quantile interval, $[p_1, p_2]$, of z_t^{rep}, where $0 \leq p_1 < p_2 \leq 1$. For example, choosing $[p_1, p_2] = [0.4, 0.6]$ gives the proportion of time that z_t^{obs} spends in the central interval $[0.4, 0.6]$. The analytic expression for calculating such proportions is $(r_n^{obs} - r_1^{obs})^{-1} \int_{r_1^{obs}}^{r_n^{obs}} \mathbb{1}_{\{\text{ppp-value}(t) \in [p_1, p_2]\}} dt$. Such information can be conveniently summarized by partitioning the space of inverse quantiles, $[0, 1]$, into intervals of length 0.1 and calculating the proportion of time that z_t^{obs} spends in each of these intervals.

5.2 Implementation

The position-time method can be implemented as described in Algorithm 3. To illustrate the method we use the same example dataset and the same fitted models as used to illustrate the distance method in Section 4.3. We also use $K = 500$ throughout the paper. Figure 3 and Tables 1 to 3 contain the results of the assessment. As for the distance method, the position-time method delivers appropriate assessments. For the Gamma-HM model, the method yields that the observed curve spends a proportion of 0.6 of its time at the lower tail interval $[0, 0.1]$ of the pack of replicated curves, giving strong reasons to doubt the adequacy of the model fit to the data (see Table 2). Similar doubts are raised for the fit of the Exp-NL model, for which the observed curve is found to spend a proportion of around 0.6 of its time at the upper tail interval $[0.9, 1]$ of replicated curves. Conversely, for the Exp-HM model, the observed curve spends a proportion of 0.8 of its time in the central interval $[0.2, 0.8]$ of replications (see Table 1), suggesting that the model is a good fit to the data.

As for the distance method, the position-time method appears to accurately capture the differences between the models, as far as the peaking tendency of their removal curves, as discussed in Section 4.3. This feature of the method is illustrated by looking

Algorithm 3 Position-time method.

Let $\boldsymbol{r}^{obs}$ be the observed removal data and $\{\boldsymbol{r}^{rep(1)}, \boldsymbol{r}^{rep(2)}, \ldots, \boldsymbol{r}^{rep(S)}\}$ a sample of replicated removal data, drawn from the posterior predictive density of the model, $\pi(\boldsymbol{r}^{rep} \mid \boldsymbol{r}^{obs})$.

1. **Apply the time-shifting:** Apply Algorithm 1 and note that for all following steps $\{\boldsymbol{r}^{rep(1)}, \boldsymbol{r}^{rep(2)}, \ldots, \boldsymbol{r}^{rep(S)}\}$ refers to the sample of shifted replicated removal data and $\{z_t^{rep(1)}, z_t^{rep(2)}, \ldots, z_t^{rep(S)}\}$ refers to the corresponding sample of shifted replicated removal curves.
2. **Calculate, pointwise, the position of the observed removal curve with respect to its posterior predictive distribution:** Discretize the interval $[r_1^{obs}, r_n^{obs}]$ by choosing a collection of equally spaced points as $r_1^{obs} = t_1 \leq t_2 \leq \cdots \leq t_K = r_n^{obs}$ with K large enough so that the numerical approximation is accurate. For each t_k calculate the time dependent ppp-value(t_k) using MC approximation as ppp-value$(t_k) = P(z_{t_k}^{rep} < z_{t_k}^{obs}) + \frac{1}{2}P(z_{t_k}^{rep} = z_{t_k}^{obs}) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}_{\{z_{t_k}^{rep(s)} < z_{t_k}^{obs}\}} + \frac{1}{2S}\sum_{s=1}^{S} \mathbb{1}_{\{z_{t_k}^{rep(s)} = z_{t_k}^{obs}\}}$, $k = 1, 2, \ldots K$.
3. **Assess the model:** Assess the model quantitatively by calculating the proportion of time that z_t^{obs} spends in specified (inverse) quantile intervals $[p_1, p_2]$ of z_t^{rep}, where $p_1, p_2 \in [0, 1], p_1 \leq p_2$, using numerical approximation as $\frac{1}{r_n^{obs} - r_1^{obs}} \int_{r_1^{obs}}^{r_n^{obs}} \mathbb{1}_{\{\text{ppp-value}(t) \in [p_1, p_2]\}} dt \approx \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_{\{\text{ppp-value}(t_k) \in [p_1, p_2]\}}$ and visually by plotting the function ppp-value(t) against time t (a ppp-value(t) history plot), at the time points t_k, $k = 1, 2, \ldots, K$, as well as by inspecting the plot of replicated removal curves $\{z_t^{rep(1)}, z_t^{rep(2)}, \ldots, z_t^{rep(S)}\}$ with z_t^{obs} imposed.

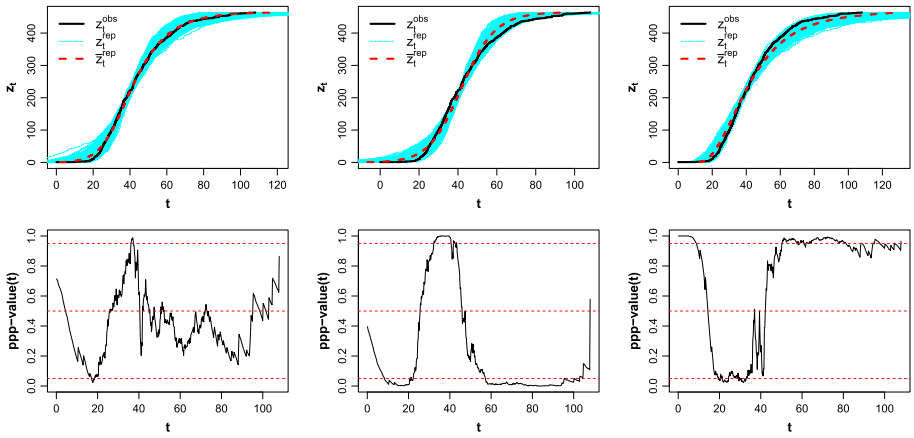

Figure 3: Example of posterior predictive checking using the position-time method. Left, middle and right columns correspond to fitted model being the Exp-HM (correctly specified), the Gamma-HM ($\nu = 10$) (misspecified due to the infectious period) and the Exp-NL ($p = 0.5$) (misspecified due to the infection rate form), respectively. Top row are plots of 500 replications from the posterior predictive distribution of the removal curve, z_t^{rep}, with the mean removal curve, $\bar{z}_t^{rep}$ (red, dashed line), and the observed removal curve, z_t^{obs} (black, solid line), imposed. Bottom row are history plots of the ppp-value(t) with the 0.05, 0.5 and 0.95 inverse quantiles (red, dashed lines) imposed. The proportion of time that z_t^{obs} spends at the inverse quantile intervals of z_t^{rep}, for each model, is given in Tables 1 to 3.

interval	[0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
time proportion	0.050	0.092	0.190	0.158	0.192
interval	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
time proportion	0.148	0.084	0.032	0.036	0.018

Table 1: Proportion of time that z_t^{obs} spends in the inverse quantile intervals of z_t^{rep} from the position-time method for the Exp-HM model (correctly specified), for the example dataset of Figure 3.

interval	[0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
time proportion	0.610	0.106	0.044	0.030	0.020
interval	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
time proportion	0.016	0.018	0.014	0.040	0.102

Table 2: Proportion of time that z_t^{obs} spends in the inverse quantile intervals of z_t^{rep} from the position-time method for the Gamma-HM ($\nu = 10$) model (misspecified due to the infectious period), for the example dataset of Figure 3.

interval	[0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
time proportion	0.166	0.028	0.026	0.022	0.010
interval	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
time proportion	0.014	0.012	0.028	0.100	0.594

Table 3: Proportion of time that z_t^{obs} spends in the inverse quantile intervals of z_t^{rep} from the position-time method for the Exp-NL ($p = 0.5$) model (misspecified due to the infection rate form), for the example dataset of Figure 3.

at the ppp-value(t) history plots (bottom row of Figure 3). Specifically, it can be seen that the observed curve, being generated from the Exp-HM model systematically lies on the lower (upper) tail interval of replicated curves produced by the Gamma-HM (Exp-NL) model, reflecting the fact that it peaks slower (faster) compared to most of the replicated curves. As for the distance method, the ability of the position-time method to assess the infectious period and the infection rate form assumptions of epidemic models is further illustrated in the simulation studies of Sections 6 and 7.

5.3 A scalar output for simulation studies

For the purposes of simulation studies in Sections 6 and 7, it is necessary in practice to restrict the output of the position-time method to a single scalar, in order to facilitate comparisons across multiple datasets. A natural choice for this output is the square root of the mean square error (MSE) of the collection of time-dependent ppp-values from the optimal value of 0.5, $\sqrt{\text{MSE}} = \left(\frac{1}{K}\sum_{k=1}^{K}\left(\text{ppp-value}(t_k) - 0.5\right)^2\right)^{\frac{1}{2}}$, where t_k and ppp-value(t_k) are defined and calculated, respectively, as in step 2 of Algorithm 3, $k = 1, 2, \ldots, K$. We note that we combine the pointwise ppp-values to produce this scalar only for manipulating the results from multiple datasets in the simulation studies and that is not a general suggestion. As we discuss in Section 9, we envisage the position-time method as a complementary type of assessment to the distance method and what makes it such is its capability to provide non-scalar, pointwise quantitative outputs.

Two useful reference points that can aid interpretation of $\sqrt{\text{MSE}}$ values are (i) $\sqrt{1/2} \approx 0.29$, the value of the $\sqrt{\text{MSE}}$ in the case that the collection of time-dependent ppp-values follows a uniform distribution in $[0, 1]$ (i.e. the observed curve assumes positions, with respect to the replicated, uniformly across time), and (ii) 0.5, the upper bound for $\sqrt{\text{MSE}}$ that occurs when the time dependent ppp-value is fixed at the least optimal value of 0 or 1 (i.e. the observed curve lies outside of the pack of replicated removal curves at all time points). For example, the $\sqrt{\text{MSE}}$ values for the example dataset of Section 5.2 are 0.22, 0.43 and 0.42 for the Exp-HM model (correctly specified), the Gamma-HM model (misspecified due to the infectious period), and the Exp-NL model (misspecified due to the infection rate form), respectively.

6 Simulation study A: Assessing the infectious period distribution assumption

6.1 Simulation and run conditions

For simulation study A, we consider the fit of the Exp-HM and Gamma-HM models under two scenarios. In scenario 1, data are generated from the Exp-HM model, and in scenario 2, data are generated from the Gamma-HM model. The simulation conditions are given in Table 4. Our intention is to create cases of model specification that are of interest, here being correct specification and misspecification due to the infectious period, and then check if the proposed methods perform appropriate assessments. To this end, we fix $\nu = 10$ whenever the Gamma-HM model is used to either generate data or be fitted to data. As explained in Section 4.3, this maintains distinction between the two models so that in scenario 1 the Exp-HM model is correctly specified and the Gamma-HM model is misspecified, and in scenario 2, the roles of the two models are reversed. Notice from Table 4 that the values of the parameters γ and λ are set so that the mean infectious period, $\mathrm{E}(T_D) = 10$, is the same under both scenarios. The same applies to the basic reproduction number R_0, which is set at 2.5 for both scenarios. Specifying the parameters in such a way ensures that datasets generated under the two scenarios only differ as a result of the different infectious period distribution.

	Data generating process	Parameter values
Scenario 1	Exp-HM	$R_0 = 2.5, \gamma = 0.1$
Scenario 2	Gamma-HM	$R_0 = 2.5, \nu = 10, \lambda = 1$

Table 4: Conditions for simulation study A. Each simulation scenario consists of 4 rounds, where the number of initial susceptibles N is set at 100, 200, 500 and 1000, respectively. For each round 24 datasets are generated.

Each simulation scenario consists of four rounds, with the number of initial susceptibles, N, being set at 100, 200, 500 and 1000, respectively. The number of initial susceptibles can be thought of as quantifying the dimension of the observed data insofar as the total number of removed individuals, n, is likely to increase with N. Thus the different rounds are used to examine if and how the performance of the methods changes, as the dimension of the observed data increases. For each round, we generate 24 datasets, to account for sampling variability.

To each generated dataset we fit both models and draw 500 replicated removal datasets from their respective posterior predictive distributions. As mentioned in Section 2.3, the replicated datasets are conditioned on having the same final size as the observed data. The runtime allowed to produce the required number of replications, under this condition, is set at 15 hours.

6.2 Results

First, from the 384 (2 scenarios $\times$ 4 rounds $\times$ 24 datasets $\times$ 2 fitted models) attempts to produce 500 replications of the same dimension as the observed data, only 2 failed

to be completed within the 15 hour limit. This is a reflection of the fact that standard SIR models accurately capture the final size when fitted to data produced from a standard SIR model, even if the infectious period distribution is not correctly specified (see Alharthi (2016, chapter 3) and Aristotelous (2020, chapter 2) for simulation studies supporting this claim).

For each model, we visualise the results by plotting medians and 95% quantile intervals for the folded ppp-value and the $\sqrt{\text{MSE}}$, for each value of N and model specification scenario (Figure 4). Plots of all folded ppp-values and all $\sqrt{\text{MSE}}$ values from the simulation study are provided in Appendix A (Aristotelous et al., 2022). Before turning to the results, recall that the folded ppp-value takes values in $[0, 1]$ and $\sqrt{\text{MSE}}$ takes values in $[0, 0.5]$, and in both cases a higher value means more evidence of lack of fit.

Under correct specification

We first consider results when the two models are correctly specified, namely scenario 1 for the Exp-HM model and scenario 2 for the Gamma-HM model. For the distance method, it can be seen from Figure 4 that the folded ppp-values are generally closer to 0 than 1 which, as appropriate, is an indication of goodness of fit. Furthermore, as N increases the folded ppp-values decrease and move closer to the optimal value of 0.

Results for the position-time method are similar to those for the distance method, as can be seen from Figure 4. The median (95% quantile interval) of the $\sqrt{\text{MSE}}$ values (pooling over N as trend was similar between rounds) was 0.21 (0.13, 0.35) and 0.19 (0.13, 0.32) for the Exp-HM and the Gamma-HM model, respectively, which is well below the least favourable value of 0.5. Unlike the distance method, it appears that N does not have an effect on the results, since the sampling distribution of the $\sqrt{\text{MSE}}$ is similar for all values of N. This in itself is not an issue; for example, the p-value in the classical setting has a uniform sampling distribution when the model is true, independently of the dimension of the data.

Under misspecification

We now discuss results when the two models are misspecified, namely scenario 2 for the Exp-HM model and scenario 1 for the Gamma-HM model. For the distance method, results appear reasonable in the sense that the folded ppp-values are generally higher under misspecification compared to correct specification (see Figure 4). For example, for the Gamma-HM model with $N = 500$, the median (95% quantile interval) folded ppp-value is 0.11 (0.01, 0.37) and 0.65 (0.10, 0.95), under correct specification and misspecification, respectively. Note however that particularly for smaller values of N ($N = 100$ and $N = 200$), the folded ppp-values are not high enough to systematically raise serious concerns for the fit of the models. However, for both models, as N increases, the folded ppp-values increase as well, implying that the lack of fit can be detected for larger values of N (see Figure 4). Notice that the values for the Exp-HM model are slightly lower than those of the Gamma-HM model. This is due to the fact that, in single realization settings such as ours, the more stochasticity a model has the harder

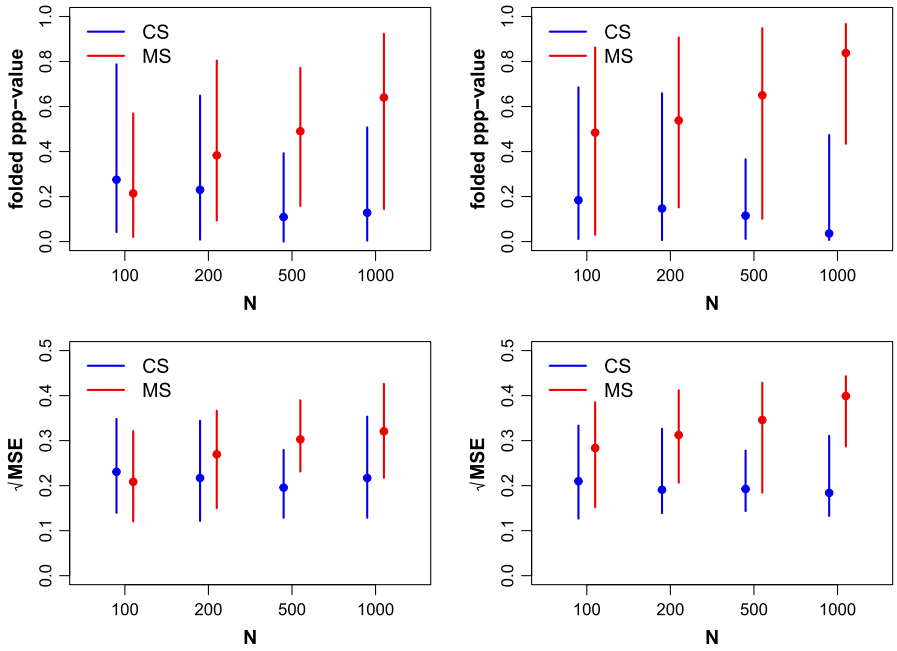

Figure 4: Folded ppp-values (top row) from the distance method and $\sqrt{\text{MSE}}$ values (bottom row) from the position-time method for the Exp-HM (left column) and Gamma-HM (right column) models, for simulation study A. The vertical lines represent 95% quantile intervals and the filled circles represent medians. Results under correct specification (CS) are coloured in blue and under misspecification (MS) in red. Simulation conditions are given in Table 4.

it is to detect misspecification. Here the Exp-HM model has more stochasticity than the Gamma-HM model due its infectious period distribution having a higher variance. Further discussion of this feature and an illustrative example are given in Aristotelous (2020, section 2.2.3).

For the position-time method, comments and conclusions are the same as for the distance method, both regarding results and the effect of N (see Figure 4).

7 Simulation study B: Assessing the infection rate form assumption

7.1 Simulation and run conditions

For simulation study B, we consider the Exp-HM and Exp-NL models. These two models have the same infectious period but different infection rate forms. We assess the fit of the two models under two simulation scenarios, scenario 1, where the data come from

	Data generating process	Parameter values
Scenario 1	Exp-HM	$R_0 = 2.5, \gamma = 0.1$
Scenario 2	Exp-NL	$p = 0.5, \gamma = 0.1,$ $\beta N = 0.85, 1.05, 1.35, 1.65$

Table 5: Conditions for simulation study B. Each simulation scenario consists of 4 rounds, where the number of initial susceptibles N is set at 100, 200, 500 and 1000, respectively. For each round 24 datasets are generated.

the Exp-HM model, and scenario 2, where data come from the Exp-NL model (see Table 5). We set the power parameter in the Exp-NL model as $p = 0.5$ to ensure a clear distinction from the Exp-HM model.

For both scenarios we set $\gamma = 0.1$ for both models so that the infectious period distribution is always the same. For scenario 1 we set $R_0 = 2.5$ as for simulation study A. The Exp-NL model does not have an equivalent reproduction number (O'Neill and Wen, 2012), so instead we set β so that the final size distribution peaks at similar values for different values of N.

We fit both the Exp-HM and Exp-NL models to each generated dataset, and produce 500 replicated removal datasets from their respective posterior predictive distributions. As in simulation study A, the replicated datasets are produced conditioned on having the same final size as the observed data, a procedure for which we set a runtime limit of 15 hours. If the required number of replications is not achieved within that time limit then the methods are not applied.

7.2 Results

For the rounds of $N = 100, 200$ and 500, the procedure of producing 500 replications within the required runtime was completed in nearly all instances, even when a model was misspecified. However, for $N = 1000$, under misspecification, the procedure was completed for only 2 out of 24 datasets for the Exp-NL model, and for 7 out of 24 datasets for the Exp-HM model. These findings show that when the infection rate form is misspecified, the ability of these models to reproduce the final size becomes increasingly low as the dimension of the observed data increases. For a further illustration of this pattern see Aristotelous (2020, section 2.8.1).

As for simulation study A, the performance of the methods is assessed via looking at medians and 95% quantile intervals of folded ppp-values and $\sqrt{\text{MSE}}$ values, while plots of all folded ppp-values and all $\sqrt{\text{MSE}}$ values are provided in Appendix B (Aristotelous et al., 2022).

Under correct specification

Results for the Exp-HM model under correct specification have already been discussed in simulation study A. When the Exp-NL model is correctly specified, Figure 5 shows that the folded ppp-values are typically closer to 0 than 1, indicating goodness of fit. It

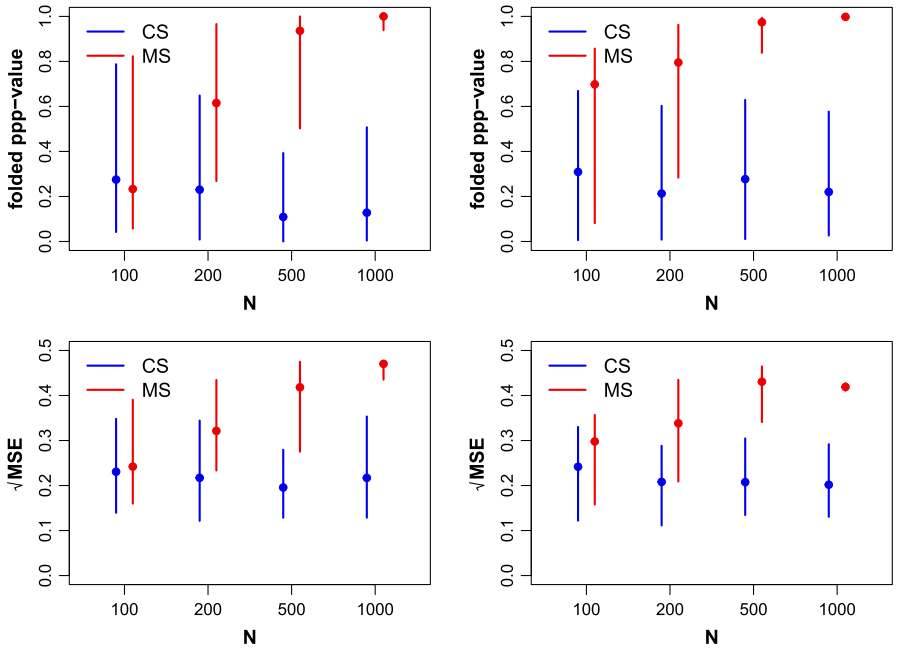

Figure 5: Folded ppp-values (top row) from the distance method and $\sqrt{\text{MSE}}$ values (bottom row) from the position-time method for the Exp-HM (left column) and Exp-NL (right column) models, for simulation study B. The vertical lines represent 95% quantile intervals and the filled circles represent medians. Results under correct specification (CS) are coloured in blue and under misspecification (MS) in red. Simulation conditions are given in Table 5.

appears that the dimension of the data does not have any apparent effect on the sampling distribution of the folded ppp-values, which is sensible under correct specification.

For the position-time method, results and the influence of N are very similar to the distance method, as can be seen from Figure 5.

Under misspecification

We now consider results when the fitted models are misspecified. For the distance method, Figure 5 shows that the folded ppp-values increase towards the optimal value of 1 for both models. This suggests that the methods are able to detect misspecification, at least for moderate sized outbreaks. As mentioned above, for $N = 1000$ only a limited number of datasets were available, but nevertheless the results were consistent enough for us to draw meaningful conclusions.

Conclusions regarding the performance of the position-time method under misspecification are essentially the same as for the distance method, for both models (see Figure 5).

8 Application of the methods to two real data examples

In this section we apply our methods to two real datasets, both from outbreaks of influenza A (H1N1) in schools. Our intention here is not to conduct a detailed and realistic analysis of these datasets but rather to provide an illustration of our methods on real data. The first outbreak (Smith et al., 2009) occurred in a boarding school in England and led to 102 cases among 1307 students. The second outbreak (Huai et al., 2010) took place in a primary school in China and resulted in 105 cases among 1314 students. The duration of the England outbreak was 33 days, while for the China one it was 17 days. For both of these outbreaks the schools were closed to contain the spread, at day 27 and at day 10, respectively. For both outbreaks data are provided in the form of new daily cases (see Smith et al. (2009, Figure 1) and Huai et al. (2010, Figure 1)) which we can transform to individual removal times, under our assumption that case-detections correspond to removals.

To each of these datasets we fitted the Exp-HM model and produced 500 replicated removal datasets from their respective posterior predictive distributions, conditioning on the final size being the same as for the observed data. We then applied the distance and the position-time methods, for which the results are given in Figure 6 and Tables 6 and 7. For the England outbreak, the two methods do not provide strong evidence for lack of fit. Neither the folded ppp-value nor the proportion of time that the observed curve spends at the lower or upper tail intervals of replications, $[0, 0.1]$ and $[0.9, 1]$, are considered too extreme, being 0.55 and 0.12 (see Table 6), respectively. The methods yield different results for the China outbreak, where there is a rather strong suggestion for lack of fit from both methods. For the distance method, the folded ppp-value is quite extreme, 0.96. Similarly, for the position-time method, the observed curve spends a proportion of 0.76 of its time at the lower or upper tail intervals of replications, $[0, 0.1]$ and $[0.9, 1]$ (see Table 7).

It is worth noting that the Exp-HM model does not account for the change in dynamics known to have taken place for both outbreaks with the schools closing. Interestingly

interval	[0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
time proportion	0.050	0.066	0.042	0.060	0.098
interval	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
time proportion	0.138	0.100	0.140	0.238	0.068

Table 6: Proportion of time that z_t^{obs} spends in the inverse quantile intervals of z_t^{rep} from the position-time method for the Exp-HM model, for the England outbreak.

interval	[0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
time proportion	0.316	0.044	0.032	0.028	0.020
interval	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
time proportion	0.026	0.016	0.038	0.038	0.442

Table 7: Proportion of time that z_t^{obs} spends in the inverse quantile intervals of z_t^{rep} from the position-time method for the Exp-HM model, for the China outbreak.

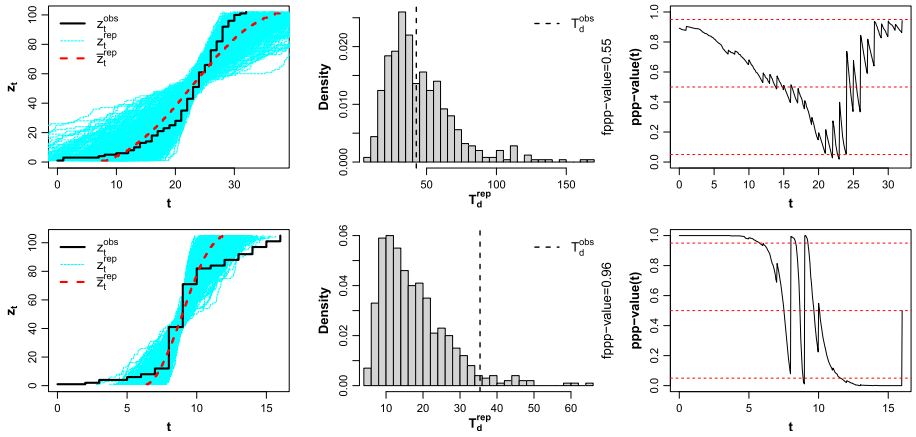

Figure 6: Posterior predictive checking using the distance and the position-time methods for the two influenza outbreak examples. Fitted model for both is the Exp-HM. Top row corresponds to the England outbreak and bottom row to the China outbreak. Left column are plots of 500 replications from the posterior predictive distribution of the removal curve, z_t^{rep}, with the mean removal curve, $\bar{z}_t^{rep}$ (red, dashed line), and the observed removal curve, z_t^{obs} (black, solid line), imposed. Middle column (distance method) are histograms of 500 replications from the posterior predictive distribution of the distance, T_d^{rep}, with the observed distance, T_d^{obs} (black, dashed line), overlaid and the corresponding folded ppp-value stated. Right column (position-time method) are history plots of the ppp-value(t) with the 0.05, 0.5 and 0.95 inverse quantiles (red, dashed lines) imposed.

though, the evidence against the fit of the model is evidently stronger for the China outbreak. This might be explained by the fact that for the England outbreak, school closure was applied at the last stages of the outbreak, day 27 out of 33, whereas for the China outbreak, it was applied relatively earlier, day 10 out of 17.

9 Discussion

The two model assessment methods we developed in this paper offer a much needed addition to the rather sparse currently existing toolkit. Both methods are based on removal data, which are typically observed in practice and thus they do not have the drawback of relying on information imputed from the model itself, like many existing stochastic residual approaches. Moreover, they effectively incorporate temporal dynamics into the assessment, an aspect which is not assessed by the commonly used final size and duration statistics. Finally, via the application of time-shifting, both methods acknowledge and account for a peculiar random take-off time feature of epidemic models, something which none of the currently existing methods do.

A key difference between the distance and the position-time methods is that for the latter no dimension reduction takes place. For the distance method, the information from the multidimensional space of removal curves L is compressed into the one-dimensional space $\mathbb{R}_{\geq 0}$, via the use of a distance statistic T_d. The effectiveness of the method relies on how effectively T_d can carry out this transfer of information and on the assumption that the posterior predictive distribution of the removal curve is symmetric around its mean. Conversely, the position-time method does not use a statistic nor does it require an assumption of symmetry. In fact it does not even need a mean removal curve to be defined. Also, it gives the possibility of determining whether the observed curve lies on the lower or the upper tail with respect to the distribution of predictive curves, corresponding to a ppp-value(t) near 0 or 1, respectively, and at which specific time points this happens. These types of information are not provided by the distance method.

Another important difference between the two methods is how the information from each replication is utilized. The distance method does not combine the information from different replications. Instead, a distance between each replicated curve and the mean curve is calculated, and then assessment is based on comparing these distances. Conversely, the position-time method is a pointwise approach and it gives the position of the observed curve, with respect to the pack of replicated curves, by combining the information from the replicated curves at each given time point. We perceive the fact that the two methods manage this information differently as useful, since our view is not to encourage choosing one of the methods over the other, but rather to employ both in order to provide complementary types of assessment regarding the fit of a model e.g. if the distance method provides an indication that the model might be lacking in fit, then the position-time method might be able to additionally pinpoint at which time points this happens.

The ability of the methods to assess the infectious period and infection rate form assumptions increases with the scale of the outbreak, as shown by the simulation studies. This implies that the methods are of practical utility, since in real-life applications interest is mainly in large-scale outbreaks. The methods could be used to assess other aspects of epidemic models as well.

We have focused solely on single-population SIR models. The main reason to do this was to facilitate an assessment of our proposed methods without possible confounding arising from other facets of the epidemic model. Nevertheless, our methods can in principle be applied to far more general models, such as those incorporating a latent period or those with structured populations.

Although in this paper we have restricted to the use of replications of the same final size as the observed data, it is possible to apply the methods using replications of varying final sizes. The benefit of using the latter type of replications, compared to using the former, is that the computational cost of producing replications is significantly less. Simulation studies in Aristotelous (2020, chapter 2), suggest that, when the dimension of the observed data is sufficiently large, the methods can still be successful in assessing the infectious period and the infection rate form assumptions.

The approach taken in this paper was to examine the performance and properties of our model assessment methods via simulation studies. Ideally, one would like to conduct such examinations theoretically, as well. However, the complexity of the epidemic setting, and the fact that the closed forms of the posterior and the posterior predictive distributions of the models are analytically intractable in practice make this task very challenging.

Furthermore, it is worth noting that some of the tools developed in this paper may prove fruitful in the context of approximate Bayesian computation (ABC) inference for epidemic models (Kypraios et al., 2017; McKinley et al., 2018). For example, it is worth investigating whether the Euclidean distance between removal curves, used in this paper, is a preferable alternative to the Euclidean distance between removal vectors, which is typically used in the context of stochastic epidemic modelling (e.g. Kypraios et al., 2017; McKinley et al., 2018), particularly because the former is computationally much cheaper, since, unlike the latter, it does not require for the removal data to be of the same dimension. It is also possible that the application of time-shifting would improve the efficiency of ABC algorithms. Finally, although the recent advances in constructing boxplot-type displays for functional data (Sun and Genton, 2011; Xie et al., 2017) were primarily developed for visualising variability in functional data, they appear to offer a promising framework via which one could perform model assessment, e.g. by being able to identify outliers from a sample of curves.

Supplementary Material

Supplementary material of "Posterior Predictive Checking for Partially Observed Stochastic Epidemic Models" (DOI: 10.1214/22-BA1336SUPP; .pdf). This consists of Appendices A and B, which contain plots of all folded ppp-values and all $\sqrt{\text{MSE}}$ values, for each model, round and scenario, for simulation studies A and B, respectively.

References

Alharthi, M. (2016). "Bayesian Model Assessment for Stochastic Epidemic Models." Ph.D. thesis, University of Nottingham. 1287, 1291, 1300

Alharthi, M., Kypraios, T., and O'Neill, P. D. (2019). "Bayes factors for partially observed stochastic epidemic models." *Bayesian Analysis*, 14(3). MR3960776. doi: https://doi.org/10.1214/18-BA1134. 1283

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Springer New York. arXiv:1411.2624. URL http://arxiv.org/pdf/1411.2624v2 MR1784822. doi: https://doi.org/10.1007/978-1-4612-1158-7. 1286, 1288

Aristotelous, G. (2020). "Topics in Bayesian Inference and Model Assessment for Partially Observed Stochastic Epidemic Models." Ph.D. thesis, University of Nottingham. URL http://eprints.nottingham.ac.uk/id/eprint/63384 MR3960776.

doi: https://doi.org/10.1214/18-BA1134. 1287, 1288, 1290, 1291, 1300, 1301, 1302, 1306

Aristotelous, G., Kypraios, T., and O'Neill, P. D. (2022). "Supplementary Material for "Posterior predictive checking for partially observed stochastic epidemic models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1336SUPP. 1300, 1302

Ball, F. and Donnelly, P. (1995). "Strong approximations for epidemic models." *Stochastic Processes and their Applications*, 55(1): 1–21. MR1312145. doi: https://doi.org/10.1016/0304-4149(94)00034-Q. 1290

Gardner, A., Deardon, R., and Darlington, G. (2011). "Goodness-of-fit measures for individual-level models of infectious disease in a Bayesian framework." *Spatial and Spatio-temporal Epidemiology*, 2(4): 273–281. 1284

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, third edition. MR3235677. 1283, 1285

Gibson, G. J., Streftaris, G., and Thong, D. (2018). "Comparison and assessment of epidemic models." *Statistical Science*, 33(1): 19–33. MR3757501. doi: https://doi.org/10.1214/17-STS615. 1284, 1287

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, first edition. MR1397966. doi: https://doi.org/10.1007/978-1-4899-4485-6. 1285

Huai, Y., Lin, J., Varma, J. K., Peng, Z., He, J., Cheng, C., Zhong, H., Chen, Y., Zheng, Y., Luo, Y., Liang, W., Wu, X., Huang, Z., McFarland, J., Feng, Z., Uyeki, T. M., and Yu, H. (2010). "Original Article: A primary school outbreak of pandemic 2009 influenza A (H1N1) in China." *Influenza and Other Respiratory Viruses*, 4(5): 259–266. MR2663325. doi: https://doi.org/10.1051/mmnp/20105312. 1304

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009). "Bayesian analysis for emerging infectious diseases." *Bayesian Analysis*, 4(3): 465–496. MR2551042. doi: https://doi.org/10.1214/09-BA417. 1284

Kypraios, T. (2007). "Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models." Ph.D. thesis, Lancaster University. 1283

Kypraios, T., Neal, P., and Prangle, D. (2017). "A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation." *Mathematical Biosciences*, 287: 42–53. MR3634152. doi: https://doi.org/10.1016/j.mbs.2016.07.001. 1307

Lekone, P. E. and Finkenstädt, B. F. (2006). "Statistical inference in a stochastic epidemic SEIR model with control intervention: ebola as a case study." *Biometrics*, 62(4): 1170–1177. MR2307442. doi: https://doi.org/10.1111/j.1541-0420.2006.00609.x. 1284

McKinley, T. J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J. E., Nsubuga,

R. N., Goldstein, M., and White, R. G. (2018). "Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models." *Statistical Science*, 33(1): 4–18. MR3757500. doi: https://doi.org/10.1214/17-STS618. 1307

Neal, P. and Roberts, G. (2005). "A case study in non-centering for data augmentation: Stochastic epidemics." *Statistics and Computing*, 15(4): 315–327. MR2205394. doi: https://doi.org/10.1007/s11222-005-4074-7. 1287

Nguyen-Van-Yen, B., Del Moral, P., and Cazelles, B. (2021). "Stochastic epidemic models inference and diagnosis with Poisson random measure data augmentation." *Mathematical Biosciences*, 335: 108583. MR4236034. doi: https://doi.org/10.1016/j.mbs.2021.108583. 1283

O'Neill, P. and Wen, C. (2012). "Modelling and inference for epidemic models featuring non-linear infection pressure." *Mathematical Biosciences*, 238(1): 38–48. MR2947082. doi: https://doi.org/10.1016/j.mbs.2012.03.007. 1286, 1302

O'Neill, P. D. (2010). "Introduction and snapshot review: Relating infectious disease transmission models to data." *Statistics in Medicine*, 29(20): 2069–2077. MR2756555. doi: https://doi.org/10.1002/sim.3968. 1284

O'Neill, P. D. and 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. MR0763809. doi: https://doi.org/10.2307/2981670. 1283, 1287

Parry, M., Gibson, G. J., Parnell, S., Gottwald, T. R., Irey, M. S., Gast, T. C., and Gilligan, C. A. (2014). "Bayesian inference for an emerging arboreal epidemic in the presence of control." *Proceedings of the National Academy of Sciences*, 111(17): 6258–6262. 1284

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ 1284

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer. MR2168993. 1290

Ripley, B. D. (2009). *Stochastic Simulation*. John Wiley & Sons. MR2299137. 1287

Severo, N. C. (1969). "Generalizations of some stochastic epidemic models." *Mathematical Biosciences*, 4(3-4): 395–402. MR0245166. doi: https://doi.org/10.1016/0025-5564(69)90019-4. 1286

Smith, A., Coles, S., Johnson, S., Saldana, L., Ihekweazu, C., and O'Moore, É. (2009). "An outbreak of influenza A(H1N1)v in a boarding school in South East England, May–June 2009." *Eurosurveillance*, 14(27). 1304

Stockdale, J. E., Kypraios, T., and O'Neill, P. D. (2017). "Modelling and Bayesian analysis of the Abakaliki smallpox data." *Epidemics*, 19: 13–23. 1288

Streftaris, G. and Gibson, G. J. (2012). "Non-exponential tolerance to infection in epidemic systems – modeling, inference, and assessment." *Biostatistics*, 13(4): 580–593. 1283, 1284

Sun, Y. and Genton, M. G. (2011). "Functional boxplots." *Journal of Computational and Graphical Statistics*, 20(2): 316–334. MR2847798. doi: https://doi.org/10.1198/jcgs.2011.09224. 1307

Xiang, F. and Neal, P. (2014). "Efficient MCMC for temporal epidemics via parameter reduction." *Computational Statistics & Data Analysis*, 80: 240–250. MR3240490. doi: https://doi.org/10.1016/j.csda.2014.07.002. 1283, 1287

Xie, W., Kurtek, S., Bharath, K., and Sun, Y. (2017). "A geometric approach to visualization of variability in functional data." *Journal of the American Statistical Association*, 112(519): 979–993. MR3735354. doi: https://doi.org/10.1080/01621459.2016.1256813. 1307

Acknowledgments

We thank the associate editor and the two anonymous referees for all their comments and suggestions which improved our manuscript.