

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 4, 918–929
https://doi.org/10.1080/10618600.2017.1328365

Taylor & Francis
Taylor & Francis Group

Efficient Data Augmentation for Fitting Stochastic Epidemic Models to Prevalence Data

Jonathan Fintzi[a], Xiang Cui[b], Jon Wakefield[a,b], and Vladimir N. Minin[b,c]

[a]Department of Biostatistics, University of Washington, Seattle, WA; [b]Department of Statistics, University of Washington, Seattle, WA; [c]Department of Biology, University of Washington, Seattle, WA

ABSTRACT

Stochastic epidemic models describe the dynamics of an epidemic as a disease spreads through a population. Typically, only a fraction of cases are observed at a set of discrete times. The absence of complete information about the time evolution of an epidemic gives rise to a complicated latent variable problem in which the state space size of the epidemic grows large as the population size increases. This makes analytically integrating over the missing data infeasible for populations of even moderate size. We present a data augmentation Markov chain Monte Carlo (MCMC) framework for Bayesian estimation of stochastic epidemic model parameters, in which measurements are augmented with subject-level disease histories. In our MCMC algorithm, we propose each new subject-level path, conditional on the data, using a time-inhomogenous continuous-time Markov process with rates determined by the infection histories of other individuals. The method is general, and may be applied to a broad class of epidemic models with only minimal modifications to the model dynamics and/or emission distribution. We present our algorithm in the context of multiple stochastic epidemic models in which the data are binomially sampled prevalence counts, and apply our method to data from an outbreak of influenza in a British boarding school. Supplementary material for this article is available online.

ARTICLE HISTORY
Received June 2016
Revised February 2017

KEYWORDS
Bayesian data augmentation; Continuous-time Markov chain; Epidemic count data; Hidden Markov model; Stochastic epidemic model

1. Introduction

Stochastic epidemic models (SEMs) are classical tools for modeling the spread of infectious diseases. An SEM represents the time evolution of an epidemic in terms of the disease histories of individuals as they transition through disease states. Incorporating stochasticity into epidemic models is important when the disease prevalence is low or when the population size is small. In both cases, the stochastic variability in the evolution of an epidemic greatly influences the probability and severity of an outbreak, as well as the conclusions we draw about its dynamics (Allen 2008; Keeling and Rohani 2008). Moreover, many questions—for example, what is the outbreak size distribution? What is the probability that a disease has been eradicated?—cannot be answered using deterministic methods (Britton 2010).

The task of fitting an SEM is typically complicated by the limited extent of epidemiological data, which are recorded at discrete observation times, commonly describe just one aspect of the disease process, for example, infections, and usually capture only a fraction of cases. Complete subject–level data, which would consist of the exact times at which individuals transition through disease states, are often unavailable (O'Neill 2010). Fitting SEMs in the absence of complete subject-level data presents a complicated latent variable problem since it is usually impossible to analytically integrate over the missing data (O'Neill 2002). This makes the observed data likelihood for an SEM intractable.

Existing approaches to fitting SEMs with intractable likelihoods have largely fallen into four groups: martingale methods, approximation methods, simulation-based methods, and data augmentation (DA) methods (O'Neill 2010). Martingale methods estimate the parameters of interest using estimating equations based on martingales for the counting processes within the SEM, for example, infections and recoveries (Becker 1977; Watson 1981; Sudbury 1985; Andersson and Britton 2000; Lindenstrand and Svensson 2013). These methods are not easily implemented for SEMs with complex dynamics fit to partially observed count data. Approximation methods replace the SEM, typically represented as a Markov jump process, with a simpler model whose likelihood is more tractable. For example, Roberts and Stramer (2001) and Cauchemez and Ferguson (2008) used diffusion processes that approximate the SEM dynamics, while Jandarov et al. (2014) used a Gaussian process approximation of a related gravity model. Another typical simplification is to discretize time and to construct a transition model for the population flow between model compartments at successive times (Longini Jr. and Koopman 1982; Held, Höhle, and Hofmann 2005; Lekone and Finkenstädt 2006; Held and Paul 2012). These methods are computationally efficient and in many cases yield sensible estimates. However, the simplifying assumptions used in the various approximations are not always appropriate. For instance, the diffusion approximation may not be valid in small populations where the system is far from its deterministic limit (Andersson and Britton 2000), while the discretization of time makes it awkward to approximate systems in which the observation times are not evenly spaced or the rates of transition events

CONTACT Vladimir N. Minin *vminin@uw.edu* Department of Statistics; Department of Biology, University of Washington, Seattle, WA 98105.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

span several orders of magnitude (Glass, Xia, and Grenfell 2003; Shelton and Ciardo 2014). Simulation-based methods use the underlying SEM to generate epidemic paths that serve as the basis for inference. This class of methods includes approximate Bayesian computation (ABC) methods (McKinley, Cook, and Deardon 2009; Toni et al. 2009), pseudo–marginal methods (McKinley et al. 2014), and sequential Monte Carlo (or particle filter) methods (Toni et al. 2009; Andrieu, Doucet, and Holenstein 2010; Ionides et al. 2011; Dukic, Lopes, and Polson 2012; Koepke et al. 2016). Within this class of methods, the particle marginal Metropolis–Hastings algorithm of Andrieu, Doucet, and Holenstein (2010) stands out in being a general method for Bayesian inference and is used as a benchmark method in this article. Although simulation-based methods have been used to fit complex models, they are computationally intensive and suffer from well-known pitfalls. ABC methods are sensitive to the choice of summary statistic, rejection threshold, and prior (Toni et al. 2009). Sequential Monte Carlo methods, on which pseudo-marginal methods often rely, are prone to "particle impoverishment" problems (Cappé, Moulines, and Ryden 2006; Dukic, Lopes, and Polson 2012).

Traditional agent-based DA methods for fitting SEMs, first presented by O'Neill and Roberts (1999) and Gibson and Renshaw (1998), target the joint posterior distribution of the missing data and model parameters to obtain a tractable complete data likelihood. That the augmentation is agent-based refers to the fact that subject-level disease histories, rather than population-level epidemic paths, are introduced as latent variables in the model. The advantage of the agent-based approach is that household structure and subject-level covariates may be incorporated into the model (Auranen et al. 2000; Höhle and Jørgensen 2002; Cauchemez et al. 2004; Neal and Roberts 2004; O'Neill 2009). Development of DA methods for SEMs is of continuing interest, and recent works by Pooley, Bishop, and Marion (2015), Qin and Shelton (2015), and Shestopaloff and Neal (2016) have presented methods that could possibly be applied to epidemic count data. However, their algorithms forgo the flexibility of agent-based DA, and in the case of the latter two articles have not been applied to SEMs.

We present an agent-based DA Markov chain Monte Carlo (MCMC) framework for fitting SEMs to time series count data. We obtain a tractable complete data likelihood by augmenting the data with subject-level disease histories. Our MCMC targets the joint posterior distribution of the latent epidemic process and the model parameters as we alternate between updating subject-level paths and model parameters. We propose each new subject-path, conditionally on the data, using a time-inhomogenous continuous-time Markov chain (CTMC) with rates determined by the disease histories of the other individuals. These data-driven path proposals result in highly efficient perturbations to the latent epidemic path, and enable us to analyze epidemic count data in the absence of any subject-level information. In contrast, traditional agent-based DA MCMC algorithms rely on data-agnostic transdimensional proposals and suffer from convergence issues as the fraction of missing information becomes large (Roberts and Stramer 2001; McKinley et al. 2014; Pooley, Bishop, and Marion 2015). The de facto need for some subject-level data has precluded the use of classical DA machinery in many settings. Thus, our MCMC algorithm enables exact Bayesian inference for SEMs fit to datasets that would have been impossible to study with existing agent-based DA methods. Finally, our algorithm is not specific to any particular SEM dynamics or measurement process, and may be applied, with minimal modifications, to a broad class of SEMs.

2. The Data Augmentation Algorithm for an SIR Model

For concreteness and clarity of exposition, we present our Bayesian DA algorithm (BDA) in the context of fitting a stochastic susceptible-infected-recovered (SIR) model to binomially distributed prevalence counts. We also use our algorithm to fit susceptible-exposed-infected-recovered (SEIR) and susceptible-infected-recovered-susceptible (SIRS) models in Sections 3.1, 3.2, and 4, and outline the minimal adaptations required for these models in Section S6 of supplementary materials.

The SIR model describes the time evolution of an epidemic in terms of the disease histories of individuals as they transition through three states—susceptible (S), infected/infectious (I), and recovered (R). Under simple SIR dynamics, each individual becomes infectious immediately upon becoming infected, and acquires lifelong immunity upon recovery. For simplicity, we assume that the population is closed and mixes homogenously, and that there is no external force of infection. Therefore, the epidemic ceases once the pool of infectious individuals is depleted.

2.1. Measurement Process and Data

Our data, $\mathbf{Y} = \{Y_1, \ldots, Y_L\}$, are disease prevalence counts recorded at times $t_1, \ldots, t_L \in [t_1, t_L]$. It should not beggar belief that the data could be subject to measurement error, for example, underreporting in settings where asymptomatic individuals escape detection. Let S_τ, I_τ, and R_τ denote the total susceptible, infected, and recovered people at time τ. We model the observed prevalence as a binomial sample, with constant detection probability ρ, of the true prevalence at each observation time. Thus,

$$Y_\ell | I_{t_\ell}, \rho \sim \text{Binomial}(I_{t_\ell}, \rho). \tag{1}$$

2.2. Latent Epidemic Process

The data are sampled from a latent epidemic process, $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, that evolves continuously in time as individuals become infected and recover. The state space of this process is $\mathcal{S} = \{S, I, R\}^N$, the Cartesian product of N state labels taking values in $\{S, I, R\}$. The state space of a single subject, $\mathbf{X}_j$, is $\mathcal{S}_j = \{S, I, R\}$, and a realized subject-path is of the form

$$\mathbf{x}_j(\tau) = \begin{cases} S, & \tau < \tau_{\text{I}}^{(j)}, \\ I, & \tau_{\text{I}}^{(j)} \le \tau < \tau_{\text{R}}^{(j)}, \\ R, & \tau_{\text{R}}^{(j)} \le \tau, \end{cases} \tag{2}$$

where $\tau_{\text{I}}^{(j)}$ and $\tau_{\text{R}}^{(j)}$ are the infection and recovery times for subject j (though subject j may also never become infected or

recover, or may become infected or recover outside of the observation period $[t_1, t_L]$). We write the configuration of $\mathbf{X}$ at time τ as $\mathbf{X}(\tau) = (\mathbf{X}_1(\tau), \ldots, \mathbf{X}_N(\tau))$, and adopt the convention that $\mathbf{X}(\tau)$ and derived quantities, for example, I_τ, depend on the configuration just before τ. We use τ^+ for quantities evaluated just after a particular time. The waiting times between transition events are taken to be exponentially distributed, and we denote by β and μ the per-contact infectivity and recovery rates. Thus, the latent epidemic process evolves according to a time-homogenous CTMC, with transition rate from configuration $\mathbf{X}$ to $\mathbf{X}'$ given by

$$\lambda_{\mathbf{X},\mathbf{X}'} = \begin{cases} \beta I, & \text{if } \mathbf{X} \text{ and } \mathbf{X}' \text{ differ only in subject } j, \\ & \text{with } \mathbf{X}_j = S, \text{ and } \mathbf{X}'_j = I, \\ \mu, & \text{if } \mathbf{X} \text{ and } \mathbf{X}' \text{ differ only in subject } j, \\ & \text{with } \mathbf{X}_j = I, \text{ and } \mathbf{X}'_j = R, \\ 0, & \text{for all other configurations } \mathbf{X} \text{ and } \mathbf{X}'. \end{cases} \quad (3)$$

At the first observation time, we let $\mathbf{X}(t_1)|\mathbf{p}_{t_1} \sim \text{Categorical}(\{S, I, R\}, \mathbf{p}_{t_1})$, where $\mathbf{p}_{t_1} = (p_S, p_I, p_R)$ are the probabilities that an individual is susceptible, infected, or recovered. Let $\boldsymbol{\tau} = \{\tau_0, \ldots, \tau_{K+1}\}$, where $t_1 \equiv \tau_0$ and $t_L \equiv \tau_{K+1}$, be the (ordered) set of K infection and recovery times of all individuals along with the endpoints of the observation period $[t_1, t_L]$. Let $\mathbb{I}(\tau_k \hat{=} I)$ and $\mathbb{I}(\tau_k \hat{=} R)$ indicate whether τ_k is an infection or recovery time, and let $\boldsymbol{\theta} = (\beta, \mu, \rho, \mathbf{p}_{t_1})$ denote the vector of unknown parameters.

The complete data likelihood is

$$\begin{aligned} L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) &= \Pr(\mathbf{Y}|\mathbf{X}, \rho) \times \Pr(\mathbf{X}(t_1)|\mathbf{p}_{t_1}) \times \pi(\mathbf{X}|\mathbf{X}(t_1), \beta, \mu) \\ &= \left[\prod_{l=1}^{L} \binom{I_{t_\ell}}{Y_\ell} \rho^{Y_\ell} (1-\rho)^{I_{t_\ell} - Y_\ell}\right] \times \left[p_S^{S_{t_1}} p_I^{I_{t_1}} p_R^{R_{t_1}}\right] \\ &\quad \times \prod_{k=1}^{K} \left\{\left[\beta I_{\tau_k} \times \mathbb{I}(\tau_k \hat{=} I) + \mu \times \mathbb{I}(\tau_k \hat{=} R)\right] \right. \\ &\quad \left. \times \exp\left[-(\tau_k - \tau_{k-1})\left(\beta I_{\tau_k} S_{\tau_k} + \mu I_{\tau_k}\right)\right]\right\} \\ &\quad \times \exp\left[-\left(t_L - \tau_K\right)\left(\beta I_{\tau_K^+} S_{\tau_K^+} + \mu I_{\tau_K^+}\right)\right]. \end{aligned} \quad (4)$$

We briefly reconcile what might seem like a discrepancy between the SIR model presented above and the lumped construction of the SIR model (see Andersson and Britton 2000), which, for a number of computational and analytical reasons, is somewhat more common. Our model describes the time evolution of the subject-level collection of disease histories, and thus evolves on the state space of individual disease labels. The lumped SIR model describes the time evolution of the vector of compartment counts, the state space of which is defined as the partition of the original state space obtained by aggregating the individuals in each of the model compartments. The lumped construction would have been appropriate had we chosen to augment the data with the compartment counts (e.g., as in Pooley, Bishop, and Marion 2015). Nonetheless, inference based on the full subject-level model will exactly match inference based on the lumped model. We discuss this further in Section S1 of supplementary materials.

2.3. Subject-Path Proposal Framework

The observed data likelihood in the posterior $\pi(\boldsymbol{\theta}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \int L(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})\pi(\mathbf{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\pi(\mathbf{X})$ is analytically intractable for even moderately sized N as it involves an extremely high-dimensional integral over the collection of subject-paths, $\mathbf{X}$. The strategy employed in DA methods is to introduce the subject-paths, $\mathbf{X}$, as latent variables in the model. This enables us to work with the tractable complete data likelihood given by (4). The joint posterior distribution is

$$\begin{aligned} \pi(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}) \ &\propto \Pr(\mathbf{Y}|\mathbf{X}, \rho) \times \pi(\mathbf{X}|\mathbf{X}(t_1), \beta, \mu) \\ &\quad \times \Pr(\mathbf{X}(t_1)|\mathbf{p}_{t_1}) \times \pi(\beta)\pi(\mu)\pi(\rho)\pi(\mathbf{p}_{t_1}), \end{aligned} \quad (5)$$

where $\pi(\beta)$, $\pi(\mu)$, $\pi(\rho)$, and $\pi(\mathbf{p}_{t_1})$ are prior densities. Our MCMC targets the joint posterior distribution, given by (5), as we alternate between updating $\mathbf{X}|\boldsymbol{\theta}, \mathbf{Y}$ and $\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}$.

Given the current collection of subject-paths, $\mathbf{x}^{\text{cur}}$, we propose $\mathbf{x}^{\text{new}}$ by sampling the path of a single subject $\mathbf{X}_j$, conditionally on the data, using a time-inhomogenous CTMC with state space $\mathcal{S}_j$ and rates conditioned on the collection of disease histories of the other individuals, $\mathbf{x}_{(-j)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_N\}$. The proposed collection of paths is accepted or rejected in a Metropolis–Hastings step.

Let $\boldsymbol{\tau}^{(j)} = \{\tau_{\text{I}}^{(j)}, \tau_{\text{R}}^{(j)}\}$ be the (possibly empty) set of infection and recovery times for subject j, and define $\boldsymbol{\tau}^{(-j)} = \{\boldsymbol{\tau} \setminus \boldsymbol{\tau}^{(j)}\} = \{\tau_0^{(-j)}, \tau_1^{(-j)}, \ldots, \tau_M^{(-j)}, \tau_{M+1}^{(-j)}\}$, where $t_1 \equiv \tau_0^{(-j)}$ and $t_L \equiv \tau_{M+1}^{(-j)}$, to be the set of $M \leq K$ (ordered) times at which other subjects become infected or recover, along with t_1 and t_L. Let $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_{M+1}\}$ be the intervals that partition $[t_1, t_L]$. That is, $\mathcal{I}_1 = [\tau_0^{(-j)}, \tau_1^{(-j)})$, $\mathcal{I}_2 = [\tau_1^{(-j)}, \tau_2^{(-j)}), \ldots,$ $\mathcal{I}_{M+1} = [\tau_M^{(-j)}, \tau_{M+1}^{(-j)})$. Let $I_\tau^{(-j)} = \sum_{i \neq j} \mathbb{I}(\mathbf{X}_i(\tau) = I)$ be the prevalence at time τ, excluding subject j. Let $\boldsymbol{\Lambda}^{(-j)}(\boldsymbol{\theta}) = \{\boldsymbol{\Lambda}_1^{(-j)}(\boldsymbol{\theta}), \ldots, \boldsymbol{\Lambda}_{M+1}^{(-j)}(\boldsymbol{\theta})\}$ be the sequence of rate matrices corresponding to each interval in $\mathcal{I}$, where for $m = 1, \ldots, M+1$,

$$\boldsymbol{\Lambda}_m^{(-j)}(\boldsymbol{\theta}) = \begin{array}{c} \\ S \\ I \\ R \end{array} \begin{array}{c} \begin{array}{ccc} S & I & R \end{array} \\ \begin{pmatrix} -\beta I_{\tau_m}^{(-j)} & \beta I_{\tau_m}^{(-j)} & 0 \\ 0 & -\mu & \mu \\ 0 & 0 & 0 \end{pmatrix} \end{array}. \quad (6)$$

We can construct the transition probability matrix for subject j over interval I_m,

$$\mathbf{P}^{(j)}(\tau_{m-1}, \tau_m) = \left(p_{a,b}^{(j)}(\tau_{m-1}, \tau_m)\right)_{a,b \in \mathcal{S}_j},$$

where $p_{a,b}^{(j)}(\tau_{m-1}, \tau_m) = \Pr(\mathbf{X}_j(\tau_m) = b|\mathbf{X}_j(\tau_{m-1}) = a, \boldsymbol{\theta})$, using the matrix exponential

$$\mathbf{P}^{(j)}(\tau_{m-1}, \tau_m) = \exp\left[(\tau_m - \tau_{m-1})\boldsymbol{\Lambda}_m^{(-j)}(\boldsymbol{\theta})\right].$$

This computation requires an eigen-decomposition of each rate matrix. We may reduce the total computational burden by computing the eigen-decompositions analytically, and by caching the decompositions to avoid duplicate computations. One additional point is that while the eigen–values of any SIR rate matrix are always real valued, this is not generally true, for example, it is possible for the rate matrix of an SIRS model to have complex eigenvalues. In this case, we obtain a real-valued

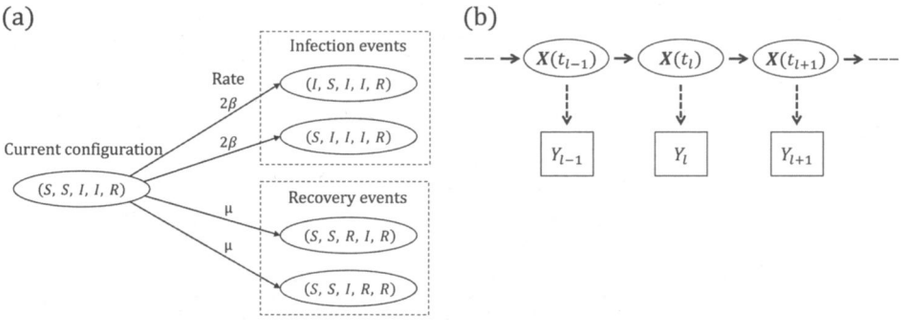


Figure 1. (a) SIR dynamics in a population of five subjects. The number of infecteds can increase from two to three via an infection of the first or second subject, reaching each of those configurations at rate 2β. The number of recovered individuals can increase from one to two via a recovery of the third or fourth subject, reaching each of those configurations at rate μ. (b) Hidden Markov model for the joint distribution of the latent epidemic process and the data. The observations, $\mathbf{Y}_\ell$, $\ell = 1, \ldots, L$, are conditionally independent given $\mathbf{X}(t)$, and $\mathbf{Y}_\ell | I_{t_\ell}, \rho \sim \text{Binomial}(I_{t_\ell}, \rho)$.

transition probability matrix by first applying a rotation to each rate matrix with complex eigenvalues to obtain its real canonical form (Hirsch, Smale, and Devaney 2013). This is discussed in Section S2 of supplementary materials.

By the Markov property, the time-inhomogenous CTMC density over the observation period $[t_1, t_L]$, denoted $\pi(\mathbf{X}_j|\mathbf{x}_{(-j)}, \boldsymbol{\theta}) \equiv \pi(\mathbf{X}_j|\boldsymbol{\Lambda}^{(-j)}(\boldsymbol{\theta}); \mathcal{I})$, can be written as a product of time-homogenous CTMC densities over the interevent intervals $\mathcal{I}_1, \ldots, \mathcal{I}_M$. Thus,

$$\pi\left(\mathbf{X}_j|\boldsymbol{\Lambda}^{(-j)}; \mathcal{I}\right) = \Pr(\mathbf{X}_j(t_1)|\mathbf{p}_{t_1}) \times \prod_{m=1}^{M} \pi\left(\mathbf{X}_j|\mathbf{x}_j(\tau_{m-1}), \boldsymbol{\Lambda}_m^{(-j)}(\boldsymbol{\theta}); \mathcal{I}_m\right). \quad (7)$$

Similarly, the transition probability matrix over an interval $\mathcal{I}_\ell = [t_{\ell-1}, t_\ell]$ can be written as the product of transition probability matrices over the subintervals in $\mathcal{I}_\ell$, within which the subject-level CTMC is time-homogenous. Thus, the transition probability matrix over an interobservation interval, $\mathcal{I}_\ell = [t_{\ell-1}, t_\ell]$, partitioned by S transition events that define interevent intervals with endpoints given by times $t_{\ell-1} \equiv \tau_{\ell,0}^{(-j)} < \tau_{\ell,1}^{(-j)} < \cdots < \tau_{\ell,S-1}^{(-j)} < \tau_{\ell,S}^{(-j)} \equiv t_\ell$, is constructed as

$$\mathbf{P}^{(j)}(t_{\ell-1}, t_\ell) = \prod_{s=1}^{S} \mathbf{P}^{(j)}\left(\tau_{\ell,s-1}^{(-j)}, \tau_{\ell,s}^{(-j)}\right).$$

The MCMC algorithm for constructing a subject-path proposal proceeds in three steps (Figure 2):

1. *HMM step*: sample the disease state of the subject under consideration at the observation times, conditional on the data and disease histories of other subjects.
2. *Discrete time skeleton step*: sample the state at times when the time-inhomogenous CTMC rates change, conditional on the states sampled in the HMM step.
3. *Event time step*: sample the exact times of transition events conditional on the sequence of states sampled in the previous steps.

2.3.1. HMM Step

The key to sampling a sequence of disease states at the observation times is to rewrite the emission probability, given by (1), as

$$Y_\ell|X_j(t_\ell), I_{t_\ell}^{(-j)}, \rho \sim \text{Binomial}\left(\mathbb{I}\left(X_j(t_\ell) = I\right) + I_{t_\ell}^{(-j)}, \rho\right). \quad (8)$$

The emission probability in (8) only depends on whether subject j is infected at time t_ℓ, since we treat the paths of all other subjects, and the parameters, as fixed. Furthermore, the data are conditionally independent of one another, given $\mathbf{x}$ and $\boldsymbol{\theta}$, which induces a hidden Markov model (HMM) over the joint distribution $\mathbf{X}$ and $\mathbf{Y}$ (Figure 1(b)).

We sample the discrete path of $\mathbf{X}_j$ at times $t_1, \ldots, t_L$ from the conditional distribution of $\mathbf{X}_j$, denoted $\pi(\mathbf{X}_j|\mathbf{Y}, \mathbf{x}_{(-j)}, \boldsymbol{\theta}; t_1, \ldots, t_L)$, using the stochastic forward–backward algorithm (Scott 2002). The algorithm efficiently computes the conditional probabilities of the paths that $\mathbf{X}_j$ can take through $\mathcal{S}_j$ in the forward recursion. A discrete path is then sampled in the backward recursion. We provide details about the HMM sampling step in supplementary material Section S3.

2.3.2. Discrete-Time Skeleton Step

It would be straightforward to sample the exact infection and recovery times of subject j, conditional on the sequence of states at times $t_1, \ldots, t_L$, if the subject-level CTMC rates did not possibly vary over each interobservation interval. We may reduce our problem to the time-homogenous case by first sampling the disease state at the intermediate event times when the CTMC rates change, and then sampling the full path within each interevent interval. Consider an interobservation interval, $\mathcal{I}_\ell = [t_{\ell-1}, t_\ell]$, containing interevent intervals whose endpoints are given by times $t_{\ell-1} \equiv \tau_{\ell,0}^{(-j)} < \tau_{\ell,1}^{(-j)} < \cdots < \tau_{\ell,n-1}^{(-j)} < \tau_{\ell,n}^{(-j)} \equiv t_\ell$. Let $\widetilde{\tau}_i = \tau_{l,i}^{(-j)}$ and $x_i = \mathbf{x}_j(\tau_{\ell,i}^{(-j)})$. We recursively sample $\mathbf{X}_j$ at each intermediate event time, beginning at $\widetilde{\tau}_1$, from the discrete distribution with masses

$$\begin{aligned}
&\Pr(\mathbf{X}_j(\widetilde{\tau}_i) = x_i \mid \mathbf{X}_j(\widetilde{\tau}_{i-1}) = x_{i-1}, \mathbf{X}_j(\widetilde{\tau}_n) = x_n) \\
&= \frac{\Pr\left(\mathbf{X}_j\left(\widetilde{\tau}_i\right) = x_i, \mathbf{X}_j\left(\widetilde{\tau}_{i-1}\right) = x_{i-1}, \mathbf{X}_j\left(\widetilde{\tau}_n\right) = x_n\right)}{\Pr\left(\mathbf{X}_j\left(\widetilde{\tau}_{i-1}\right) = x_{i-1}, \mathbf{X}_j\left(\widetilde{\tau}_n\right) = x_n\right)} \\
&= \frac{\Pr\left(\mathbf{X}_j\left(\widetilde{\tau}_i\right) = x_i \mid \mathbf{X}_j\left(\widetilde{\tau}_{i-1}\right) = x_{i-1}\right) \Pr\left(\mathbf{X}_j\left(\widetilde{\tau}_n\right) = x_n \mid \mathbf{X}_j\left(\widetilde{\tau}_i\right) = x_i\right)}{\Pr\left(\mathbf{X}_j\left(\widetilde{\tau}_n\right) = x_n | \mathbf{X}_j\left(\widetilde{\tau}_{i-1}\right) = x_{i-1}\right)} \\
&= \frac{\left[\mathbf{P}^{(j)}\left(\widetilde{\tau}_{i-1}, \widetilde{\tau}_i\right)\right]_{x_{i-1}, x_i} \left[\prod_{k=i}^{n-1} \mathbf{P}^{(j)}\left(\widetilde{\tau}_k, \widetilde{\tau}_{k+1}\right)\right]_{x_i, x_n}}{\left[\prod_{k=i-1}^{n-1} \mathbf{P}^{(j)}\left(\widetilde{\tau}_k, \widetilde{\tau}_{k+1}\right)\right]_{x_{i-1}, x_n}}.
\end{aligned} \quad (9)$$

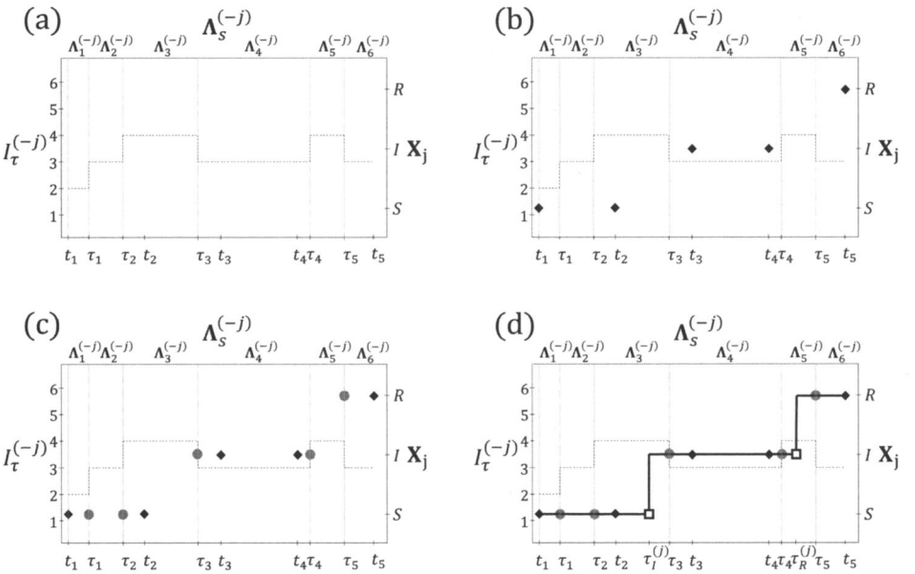


Figure 2. Procedure for constructing a subject-path proposal with SIR dynamics. (a) The dashed line depicts the number of infected individuals, excluding $\mathbf{X}_j$, the subject whose path is being sampled. The observation times, $t_1, \ldots, t_5$, and times at which other subjects change disease states, $\tau_1, \ldots, \tau_5$, are shown on the bottom axis. Rate matrices of the time-inhomogenous CTMC (top axis) are constant within interevent intervals (vertical lines). The state space of the subject-level process, $\mathbf{X}_j$, is shown on the right axis. (b) *HMM step*: Sample the state of $\mathbf{X}_j$ at $t_1, \ldots, t_5$, conditional on the data and on the disease histories of other subjects. (c) *Discrete time skeleton step*: Sample the infection status at $\tau_1, \ldots, \tau_5$, conditional on the sequence of states sampled in the HMM step. (d) *Event time step*: Sample the infection and recovery times from endpoint-conditioned time-homogenous CTMC distributions, conditional on the sequence of disease states sampled in the HMM and discrete time skeleton steps.

2.3.3. Event Time Step

The final step in constructing a subject-path is to sample the exact infection and recovery times given the discrete sequence of states obtained in the previous two steps. This amounts to simulating the path of an endpoint-conditioned time-homogeneous CTMC, a task for which there exists a variety of efficient methods (Hobolth and Stone 2009). When fitting the SIR model, we chose to use modified rejection sampling, a modification of Gillespie's direct algorithm (Gillespie 1976) that explicitly avoids simulating constant paths. This method is known to be efficient when the states differ at the endpoints of small time intervals. We used uniformization-based sampling (Hobolth and Stone 2009) when fitting SEIR and SIRS models, which was more robust when sampling paths in intervals with multiple transitions. Fast implementations of these methods are available in the `ECctmc` package in R (Fintzi 2016). We briefly summarize the algorithms in Section S4 of supplementary materials.

2.3.4. Metropolis–Hastings Step

Having constructed a complete subject-path proposal, we decide whether to accept or reject the proposal via a Metropolis–Hastings step. It is important to understand that the true distribution of $\mathbf{X}_j|\mathbf{x}_{(-j)}, \boldsymbol{\theta}$ is neither Markovian nor analytically tractable, and therefore, does not match the time-inhomogenous CTMC in our proposal. Suppressing the dependence on $\boldsymbol{\theta}$, the target distribution of the subject-path proposal is $\pi(\mathbf{X}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\mathbf{X})\pi(\mathbf{X})$. Thus, we accept a proposed subject-path with probability

$$a_{\mathbf{x}^{\text{cur}} \longrightarrow \mathbf{x}^{\text{new}}} = \min\left\{\frac{\pi(\mathbf{x}^{\text{new}}|\mathbf{y})}{\pi(\mathbf{x}^{\text{cur}}|\mathbf{y})}\frac{q(\mathbf{x}^{\text{cur}}|\mathbf{x}^{\text{new}}, \mathbf{y})}{q(\mathbf{x}^{\text{new}}|\mathbf{x}^{\text{cur}}, \mathbf{y})}, 1\right\}$$
$$= \min\left\{\frac{\pi(\mathbf{x}^{\text{new}})}{\pi(\mathbf{x}^{\text{cur}})}\frac{\pi\left(\mathbf{x}_j^{\text{cur}}|\boldsymbol{\Lambda}^{(-j)}; \mathcal{I}\right)}{\pi\left(\mathbf{x}_j^{\text{new}}|\boldsymbol{\Lambda}^{(-j)}; \mathcal{I}\right)}, 1\right\}. \quad (10)$$

Hence, the Metropolis–Hastings ratio is equal to the ratio of population-level time-homogenous CTMC densities, multiplied by the ratio of time-inhomogenous CTMC proposal densities (see supplementary material Section S5 for the derivation).

2.3.5. Initializing the Collection of Subject-Paths

We initialize the collection of subject-paths at the start of our MCMC by simulating paths using Gillespie's direct algorithm (Gillespie 1976) until we have found one under which the data have nonzero probability. A sufficient condition for this under the binomial sampling model is that the number of infected individuals is greater than the observed prevalence at each observation time.

2.4. Parameter Updates

One MCMC iteration includes a number of subject-path updates, followed by a set of parameter updates. The optimal number of subject-path updates per MCMC iteration is specific to the dynamics of the SEM and the epidemic setting (e.g., endemic vs. epidemic, high vs. low escape probability), but ultimately boils down to the cost of subject-path updates vis-a-vis parameter updates. We discuss this further in Section S7 of supplementary materials. In the case of the SIR model, as well as the other models, we will fit in subsequent sections, conjugate priors are available for all our model parameters. Thus, we use Gibbs sampling to draw new parameter values from their univariate full conditional distributions (see supplementary material Section S8).

2.5. Implementation

We provide the R and C++ code base for this article, along with examples and the code for reproducing the results we present

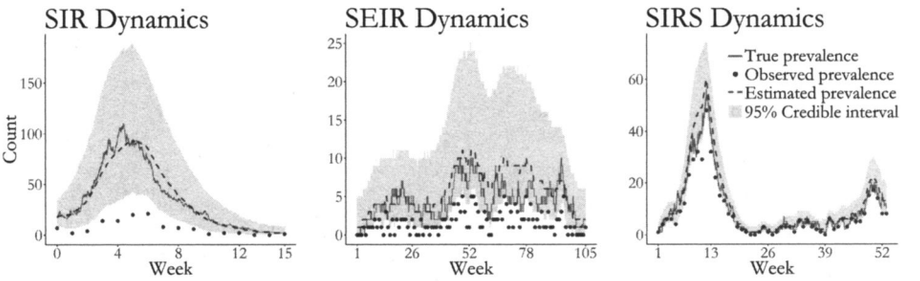


Figure 3. Estimated latent posterior distributions of disease prevalence in outbreaks simulated under SIR (left), SEIR (middle), and SIRS (right) dynamics. Depicted are the true unobserved prevalence (solid line), observed data (dots), pointwise posterior median prevalence (dashed line), and pointwise 95% credible intervals (shaded region). Latent posterior estimates are based on a thinned sample, with every 250th sample retained.

in the following sections, in the form of an R package in a stable GitHub repository (*https://github.com/fintzij/BDAepimodel*). Future implementations, including extensions to the algorithm presented in this article, along with improvements to the implementation, will be incorporated into the `stemr` package (*https://github.com/fintzij/stemr*).

3. Simulation Results

3.1. Inference Under Various Epidemic Dynamics

We fit SIR, SEIR, and SIRS dynamics to binomially distributed prevalence counts sampled from epidemics simulated under corresponding dynamics in populations of 750, 500, and 200 individuals (details provided in supplementary material Section S9). Priors for the rate parameters and binomial sampling probability were chosen so that the priors spanned reasonable ranges of values (e.g., recovery durations ranging from days to weeks/months rather than seconds to eons under extremely diffuse priors), but were otherwise only mildly informative, while the initial distribution parameters were assigned informative priors (see supplementary Tables S4, S6, and S8). The three datasets, depicted in Figure 3 along with the estimated pointwise posterior prevalence, presented a range of challenges. The SIR example was arguably the most "standard" example as the observation period captured the exponential growth and decline of the epidemic. Thus, much of the curvature in the latent path was reflected in the data. In contrast, data from the outbreak simulated under near-endemic SEIR dynamics contained very little information about the shape of the epidemic curve. The task of disentangling whether the data were sampled with low probability from a high-prevalence outbreak, or vice versa, was further complicated by the inclusion of an additional disease state—the exposed state—that was not directly observed. Finally, the SIRS model was more computationally challenging for two reasons. First, the recurrent nature of the disease process demanded that the disease state at each event time, and the path within each interevent interval, be sampled in the subject-path proposal. Second, it was possible for CTMC rate matrices to have complex eigen-decompositions, which made computing transition probability matrices more expensive. This affected the optimal number of subject-path updates per MCMC iteration (see supplementary material Section S7 for further discussion of this point). Simulation details, along with minor adaptations to our algorithm for fitting the SEIR and SIRS models, are presented in supplementary material Section S6.

The true epidemic paths and parameter values fell well within the 95% Bayesian credible intervals in all three simulations (Figure 3 presents the estimated latent posterior prevalence; Figure 4 presents posterior estimates of model parameters; supplementary material Figure S12 presents estimated latent posterior distributions and true epidemic paths for all model compartments). The acceptance rates for subject-path

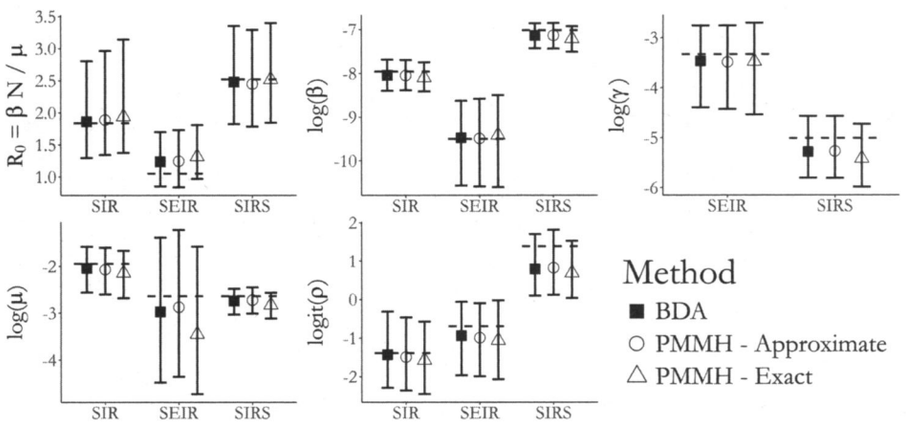


Figure 4. Posterior medians and 95% credible intervals of parameters in the SIR, SEIR, and SIRS model fits with Bayesian data augmentation (BDA) and particle marginal Metropolis–Hastings (PMMH) with particle paths simulated approximately (using τ-leaping) and exactly (using Gillespie's direct algorithm). Displayed are estimates of the basic reproductive number, R_0, the rate parameters, and the binomial sampling probability. In all models, β is the per-contact infectivity rate, μ is the recovery rate, and ρ is the binomial sampling probability. In the SEIR model, γ denotes the rate at which an exposed individual becomes infectious, while in the SIRS model γ denotes the rate at which immunity is lost.

proposals were roughly 92% for the SIR model, 91% for the SEIR model, and 77% for the SIRS model. Posterior estimates of model parameters obtained using our data augmentation algorithm also closely match estimates obtained using the particle marginal Metropolis–Hastings (PMMH) algorithm of Andrieu, Doucet, and Holenstein (2010), implemented using the pomp package in R (King, Nguyen, and Ionides 2016). We simulated particle paths in the PMMH algorithm in two ways; exactly using Gillespie's direct algorithm (Gillespie 1976), and approximately using a multinomial modification of τ-leaping (Bretó and Ionides 2011). In these small population examples, the exact algorithm is arguably more appropriate, as the leap conditions for τ-leaping may not be met in small populations, but it is also substantially slower. In these simple settings, PMMH tended to outperform our algorithm in terms of log-posterior effective sample size (ESS) per CPU time. When PMMH particle paths were simulated by τ-leaping, the average ESS per CPU compared to BDA was roughly 350× greater for the SIR model, 4.4× greater for the SEIR model, and 13× greater for the SIRS model. Exact simulation of PMMH particle paths reduced the computational advantage of PMMH substantially. In this case, the average log-posterior ESS per CPU time was 10.5× greater for PMMH in fitting the SIR model, 2× for the SEIR model, and 0.7× for the SIRS model. These comparisons did not include the time required to tune the MCMC for PMMH, which was nontrivial. In contrast, our algorithm required no tuning beyond selecting the number of subject-paths to update per MCMC iteration. We also note that in fitting the models using PMMH, we were required to make several implementation decisions to prevent particle degeneracy and to balance speed with precision. These included selecting the number of particles and the time-step in the approximate τ-leaping algorithm. For example, when using τ-leaping to simulate particle paths, the number of particles required to obtain good mixing for the SIRS model fit with PMMH was much higher than for the other two models. Details of the PMMH implementations and further results are discussed in supplementary material Section S9.

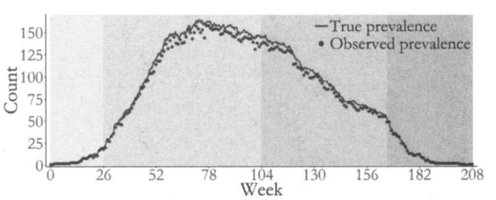
3.2. Inference Under Model Misspecification

In practice, every stochastic epidemic model is misspecified with respect to the real world epidemic process from which the data arise, and the malignancy of the model misspecification is often impossible to diagnose a priori. We can build up an understanding of an epidemic's dynamics by fitting SEMS under a range of dynamics, beginning with simple, easily interpretable models. The results of each model are interpreted counterfactually—for example, "If the true epidemic followed SIR dynamics, our best guess of the dynamics that gave rise to the data would be..." The iterative nature of epidemic modeling suggests that some minimal criteria for the usefulness of any computational algorithm would be that, for a reasonable model, the MCMC should converge to the posterior of the model parameters, and that the estimated latent posterior distribution under the hypothetical dynamics should reflect the true epidemic.

However, it is precisely the inherent misspecification of SEMs that leads simulation-based methods to struggle in many instances, and it is here that we highlight a critical advantage of our DA algorithm. Our subject-path proposals are driven, not just by the SEM dynamics, but also by the data. This enables us to overcome model misspecification in situations in which simulation-based methods degenerate due to their reliance on an adequately accurate model for simulating epidemic paths. We demonstrate this in a simple example in which we fit SIR and SEIR models to 4 years of weekly prevalence data sampled from an epidemic simulated under time-varying SEIR dynamics, where the latent period, infectious period, and per-contact infectivity rate were modulated over four discrete epochs (depicted in Figure 5, details presented in supplementary material Section S10).

We fit SIR and SEIR models to the data using our DA algorithm, and using PMMH with 2500 particles, the paths for which were simulated approximately via τ-leaping with a time-step of 1 day. We assigned weakly informative priors for the rate parameters governing the epidemic dynamics in both models, and informative priors for the binomial sampling probability and the initial state probabilities (supplementary material Table S11). The MCMC chains for models fit with PMMH suffered from severe particle degeneracy and did not converge (see supplementary material Figures S13 and S15).

Both models fit via DA yield reasonable estimates for the within-subject disease dynamics (i.e., the infectious period, as well as the latent period in the case of the SEIR model). The posterior median average infectious period duration was estimated to be 292 days (95% BCI: 263 days, 323 days) under SIR dynamics, and 287 days (95% BCI: 260 days, 318 days) under SEIR dynamics. The posterior median average latent period under SEIR dynamics was 211 days (95% BCI: 165 days, 260 days). The posterior median estimate of R_0 under SIR dynamics was 4.05 (95% BCI: 3.40, 4.81), while under SEIR dynamics, the posterior median estimate of R_0 was 23.8 (95% BCI: 15.1, 37.0). While the true prevalence fell well within the pointwise 95% credible interval for both models (Figure 6), we notice that the degree of model misspecification drastically affected our ability to estimate the history of the numbers of noninfectious people over the course of the epidemic. Under SIR dynamics, we



	Epoch			
Parameter	1	2	3	4
R_0^{Eff}	14.9	9.2	0.1	0
$1/\gamma$ (days)	210	210	90	180
$1/\mu$ (days)	150	330	300	70

Figure 5. Simulated outbreak with SEIR dynamics that varied over four epochs (shaded regions). Weekly prevalence counts (points) were binomially sampled with sampling probability $\rho = 0.95$ from the true unobserved prevalence (solid line). The table presents the effective reproductive number computed based on the number of susceptibles at the beginning of each epoch, $R_0^{\text{Eff}} = \beta(\tau)S(\tau)/\mu(\tau)$, the mean latent period, $1/\gamma$, and the mean infectious period, $1/\mu$.

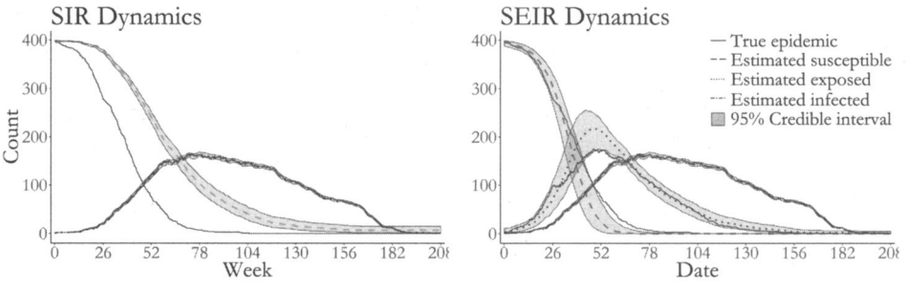


Figure 6. True epidemic path (solid lines), pointwise posterior median estimate of the numbers of susceptibles (dashed line), exposed (dotted line), and infected individuals (dash-dotted line), and pointwise 95% credible intervals (shaded regions) under SIR and SEIR dynamics.

drastically overestimate the number of susceptible individuals. The SEIR model much more closely resembles the time-varying SEIR model used to simulate the epidemic. Although the true path for the number of susceptible still falls outside the 95% credible interval at times, we are still able to reconstruct a reasonable range of paths for the number of exposed individuals. This contrasts with the models fit in Section 3.1, which were not misspecified with respect to the true epidemic dynamics. In that case, the complete path of the epidemic fell well within the estimated credible intervals for all disease states for all three models (supplementary material Figure S12). Therefore, we advise caution in reconstructing the epidemic history for disease states that were not measured, particularly when severe model misspecification is suspected.

3.3. *Inference Under Population Size Misspecification*

Model misspecification often extends not only to the SEM dynamics, but also to the assumed population size. This is often the case in settings where subject-level data are unavailable, for example, in resource limited settings or surveillance settings, and may result in biased estimates of the SEM dynamics. This bias is the result of a mismatch between the intensive dynamics of the epidemic process, which are a function of the fractions of people in the population in each disease state, and the extensive scale of prevalence counts, which are not normalized by the population size. Without knowing the true population size, it is difficult to know whether the scale of the observed counts reflects a high prevalence/low detection rate setting, or vice versa. Moreover, wrongly assuming too large, or too small, of a population size could bias posterior inference of the epidemic dynamics.

We simulated weekly prevalence counts under a binomial measurement process with detection probability $\rho = 0.3$ from an epidemic with SIR dynamics in a population of $N = 1250$ individuals. We then fit SIR models using a series of assumed population sizes under a flat prior for the binomial sampling probability and diffuse priors for the epidemic dynamics (see supplementary material Section S11 for complete simulation details and prior specifications), and compared the resulting scaled parameter estimates. The per-contact infectivity rate, β, was rescaled by the population size, N, so that it could be interpreted as the rate of disease transmission. We computed R_0 using the assumed population size. Finally, we scaled the binomial sampling probability by the assumed population size to give the expected number of observed infections in a completely infected population.

We are able to obtain approximately valid inference under moderate misspecification of the population size. However, estimates of the epidemic dynamics and the case detection probability become severely biased as the magnitude of the misspecification increases. Furthermore, the widths of the credible intervals for the model parameters shrink as misspecification of the population size becomes more severe (Figure 7). The constrained ranges of model dynamics also manifest in a narrowing of the widths of the pointwise credible intervals for disease

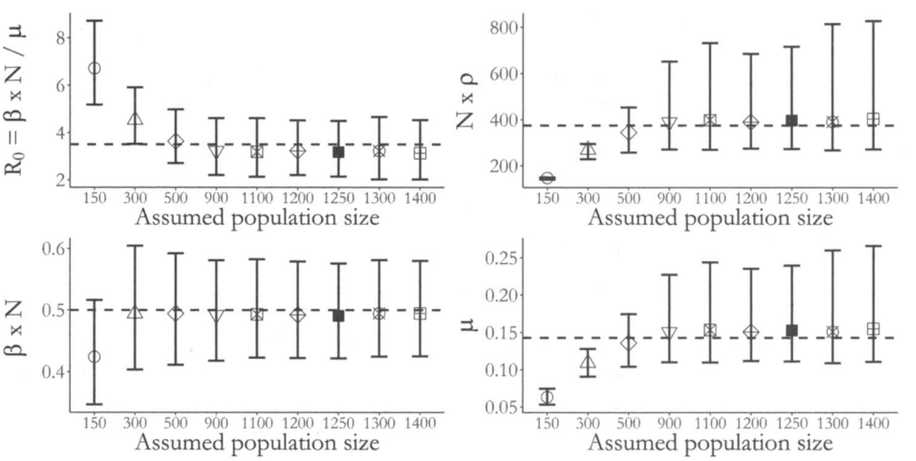


Figure 7. Posterior medians and 95% credible intervals for the basic reproductive number, R_0, infectivity rate, recovery rate, and binomial sampling probability scaled by the assumed population size. The dashed lines indicate the true values in the population of size 1250. The population size, N, indicates the assumed population size used in fitting the model.

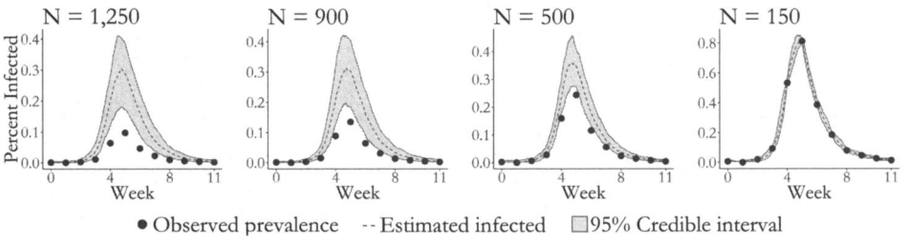


Figure 8. Estimated latent posterior distributions of disease prevalence under SIR dynamics. The true population size is 1250. Depicted are the observed prevalence (dots), pointwise posterior median prevalence (dashed line), and pointwise 95% credible intervals (shaded region) all scaled by the assumed population size. Latent posterior estimates are based on a thinned sample, with every 250th sample retained.

prevalence (Figure 8). Under severe misspecification of the population size ($N = 150$), the latent posterior distribution has 95% of its mass within only a narrow band of epidemic paths. In contrast, under moderate misspecification of the population size, the widths of the latent posterior credible intervals are quite similar to the estimated range using the true population size.

There are two final points that we wish to make based on this simulation. The first is that it might be possible to deliberately misspecify the true population size to speed up computation time and still obtain approximately valid inference. The average run time using the true population size of 1250 individuals was roughly 2× and 7× longer than the average run times in populations of 900 and 500 individuals. Yet, posterior inferences about the epidemic dynamics were not substantially affected. Longer run times in large populations result from having to sample more subject-paths per MCMC iteration at a relatively higher cost per subject-path. The second point is that in situations where the true population size is unknown, SEM likelihood-based inference has some robustness to misspecification of the population size, at least in a neighborhood of population sizes around the true number of individuals. Thus, comparing posterior inferences under a range of population sizes could be a useful heuristic diagnostic for population size misspecification.

3.4. *Effect of Prior Specification on Posterior Inference*

Given the relatively limited extent of aggregated prevalence counts compared to a setting in which subject-level data are available, we must consider how our choices of prior distributions influence our posterior inferences. We simulated an outbreak with SIR dynamics in a population of 750 individuals for which $R_0 = \beta \times 763/\mu \approx 1.84$ and the mean infectious period was $1/\mu = 7$ days. We fit SIR models to binomially distributed weekly prevalence data, sampled with detection probability $\rho = 0.2$, under the following four prior regimes: Regime 1—informative priors for all model parameters; Regime 2—vague priors for the rate parameters and an informative prior for the sampling probability; Regime 3—informative priors for the rate parameters and a flat prior for the sampling probability; Regime 4—vague priors for the rate parameters and a flat prior for the sampling probability. The same prior for the initial state probabilities was used in all four regimes. Complete simulation details and convergence diagnostics are supplied in Section S12.

The true values for all model parameters fell within the 95% credible intervals under all four prior regimes. Unsurprisingly, informative priors tended to result in narrower credible intervals for the parameters (Figure 9) as well as for the latent process (Figure 10). The strength of prior information about the sampling probability affected the widths of credible intervals to a much greater extent than the priors for the rate parameters. Strong prior information about the sampling probability also resulted in substantially narrower credible intervals for disease prevalence under each of the prior regimes for the rate parameters. In contrast, informative priors for the rate parameters yielded only slightly narrower credible intervals for disease prevalence when holding constant the strength of the sampling probability prior. The effects on the initial state probability parameters seem to reverse this pattern, although we caution against overinterpretion given the paucity of data available for estimating those parameters. MCMC chains with strong priors for the binomial sampling probability also appeared to mix somewhat better than chains with diffuse priors for the sampling probability (see traceplots in supplementary material Section S12.

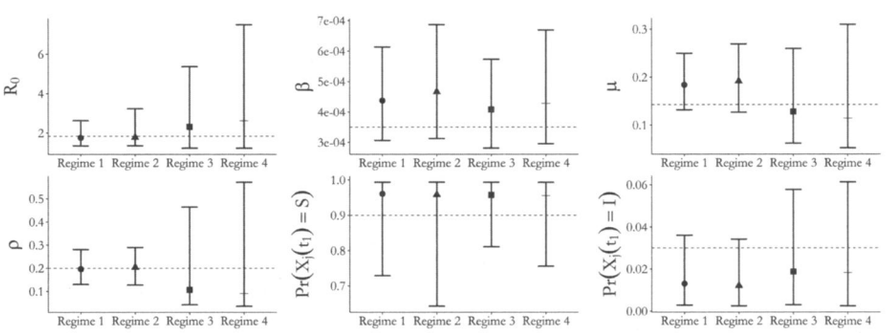


Figure 9. Posterior median estimates and 95% credible intervals for all SIR model parameters under four different prior regimes (Table S14). Regimes 1 and 3 set informative priors for the per-contact infectivity and recovery rates. Regimes 1 and 2 set informative priors for the binomial sampling probability. The same mildly informative prior for the initial state probabilities was used in all four regimes.

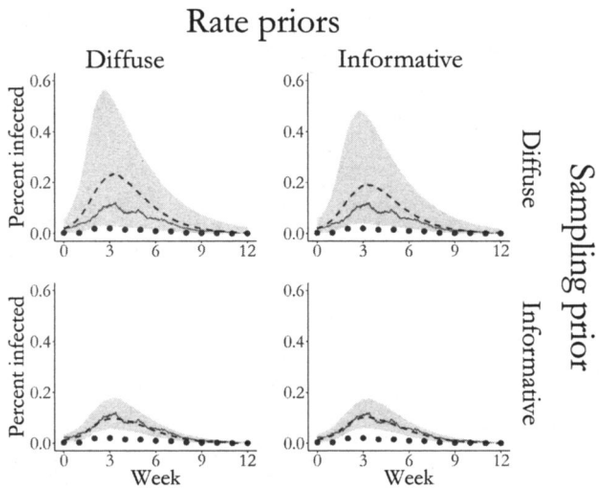


Figure 10. Estimated latent posterior distributions of disease prevalence in outbreaks simulated under four prior regimes for SIR model rate parameters and the binomial sampling probability. Depicted are the true unobserved prevalence (solid line), observed data (dots), pointwise posterior median prevalence (dashed line), and pointwise 95% credible intervals (shaded region). Latent posterior estimates are based on a thinned sample, with every 250th sample retained.

4. Influenza in a British Boarding School

As an application, we analyze data from an outbreak of influenza in a British boarding school (Anon. 1978; Davies et al. 1982). This outbreak took place shortly after the Easter term began in January 1978, and was estimated to eventually infect roughly 90% of the 763 boys aged 10–18. Daily counts of the boys who were confined to the infirmary from January 22nd through February 4th were accessed via the pomp package in R (King, Nguyen, and Ionides 2016), and are displayed in Figure 11.

We used our DA algorithm and PMMH to fit SIR and SEIR models with a binomial emission distribution to the data (see supplementary material Section S13 of the supplement for complete details). All of the parameters were assigned diffuse priors, which are plotted over the posterior ranges in Figure 12. The PMMH algorithm failed to converge for both models, which we suspect was due to a combination of model misspecification and the constrained state space of the binomial measurement process. We also fit a set of supplementary SIR and SEIR models in Section S13.2, in which we assumed a negative-binomial emission distribution. This was done to facilitate comparison with PMMH, although we feel that a negative binomial emission distribution is not appropriate in such a closely monitored outbreak setting since it does not rule out over-reporting of cases.

Together, the SIR and SEIR models suggest that cases were detected with high probability and that the outbreak, though aggressive, was not atypical given the closed environment in which it occurred. The posterior median estimates of the detection probability, roughly 0.98 for both models (SIR 95% BCI: 0.92, 1.00; SEIR 95% BCI: 0.91, 1.00), suggested that while almost all of the infectious boys were detected, a handful of cases went unnoticed. The posterior median recovery rate under SIR dynamics corresponds to an average period of 2.16 days (95%

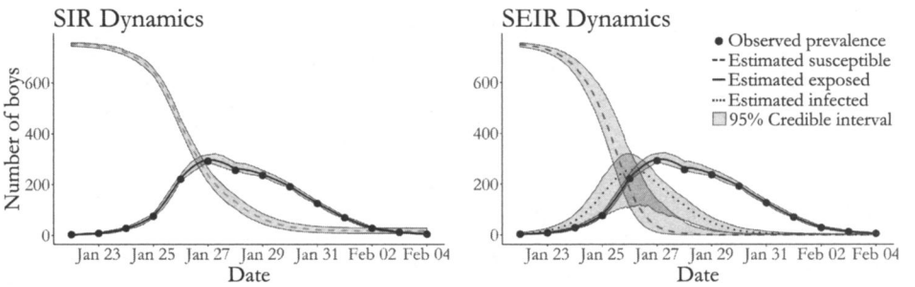


Figure 11. Boarding school data, pointwise posterior median estimates, and pointwise 95% credible intervals (gray shaded areas) under SIR and SEIR dynamics of the numbers of susceptible boys (dashed line), exposed boys (dotted line), and infected boys (solid line). Posterior estimates based on a thinned sample, with every 250th configuration retained.

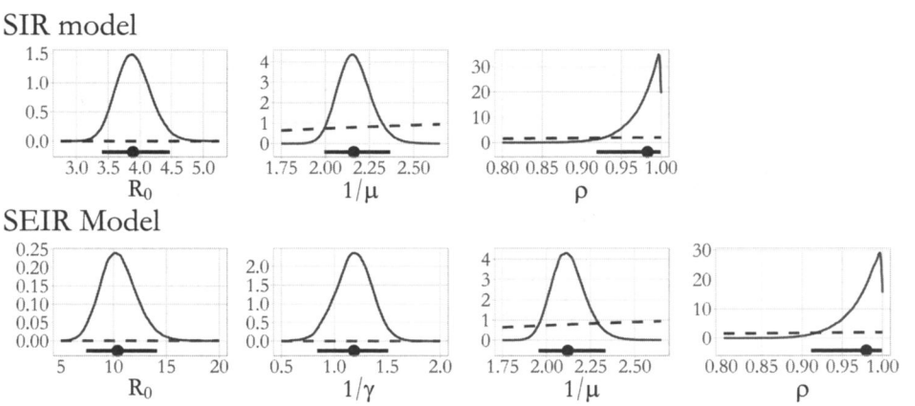


Figure 12. Posterior density estimates for $R_0 = \beta N/\mu$, the mean latent and infectious periods, $1/\gamma$ and $1/\mu$, and the binomial sampling probability, ρ, from SIR and SEIR model parameters fit to the British boarding school data (solid lines). The posterior median and 95% Bayesian credible intervals are drawn below the density plots (solid lines with circles). The implied prior densities (dashed lines) for R_0 and the latent and infectious periods, and the prior density for the binomial sampling probability, are plotted over the posterior ranges.

BCI: 1.99, 2.37) during which an infectious boy could transmit an infection to other boys before being confined to the infirmary. Under SEIR dynamics, the posterior median average infectious period was 2.12 days (95% BCI: 1.95, 2.33), and the posterior median average latent period was 1.19 days (95% BCI: 0.84, 1.51). These results are consistent with the typical progression of influenza, in which individuals typically incubate for between 1 and 4 days before symptoms manifest, and are typically infectious for 1 day before, and up to a week after, symptom onset (Centers for Disease Control and Prevention 2014). The posterior median estimates of R_0 were 3.89 (95% BCI: 3.40, 4.47) under SIR dynamics, and 10.38 (95% BCI: 7.40, 14.11) under SEIR dynamics. Previous analyses of this dataset with trajectory matching estimate R_0 to be roughly 3.7 for the SIR model and 35.9 for the SEIR model (Wearing, Rohani, and Keeling 2005; Keeling and Rohani 2008), though we note that these estimates are based on deterministic models that do not properly account for distributional properties of the data. Our results for both models are also in agreement with estimates of SIR and SEIR model dynamics under a negative binomial emission distribution (see Section S13.2).

5. Conclusion

We have presented an agent-based Bayesian DA algorithm for fitting SEMs to disease prevalence time series counts. This was previously difficult, if not computationally infeasible, to carry out using traditional agent-based DA methods in the absence of subject-level data. Although we outlined the algorithm in the context of fitting an SIR model to binomially distributed prevalence data, our algorithm represents a general solution for fitting SEMs to prevalence counts. In simulations and the applied example, we fit SEIR and SIRS models to prevalence data, and in the supplement also fit SIR and SEIR models with a negative binomial emission distribution to the British boarding school data. We have demonstrated that our algorithm yields approximately valid inference when the population size is misspecified. Moreover, our algorithm is usable in settings in which simulation-based methods, such as PMMH, break down due to misspecification of the SEM. Finally, our DA algorithm is carried out entirely at the subject level, making it possible to also incorporate subject-level covariates and household structure, or to fit models to subject-level data.

There are two fundamental limitations of agent-based DA methods from which our algorithm is not excepted. First, the bookkeeping required to track the collection of subject-paths increases in size and complexity as the number of events grows large. Attempts to fit stochastic epidemic models in large populations using agent-based DA may be thwarted by prohibitive computational overhead. MCMC run times using our implementation, which was coded for reliability rather than speed, substantially degraded once the assumed population size was greater than a few thousand people. Second, we suspect that MCMC mixing in large populations could eventually become too slow for agent-based DA to be of practical use, even if solutions could be found for the computational bottlenecks. As the population size gets large, perturbations to the likelihood from resampling one subject at a time become relatively less significant. For this reason, we view extensions for jointly sampling multiple subject-paths as a critical step in mitigating slow MCMC mixing in large populations.

Finally, we would like to comment on directions for future work that we intend to pursue. The DA algorithm in this article addresses the problem of fitting SEMs to prevalence data. This type of data summarizes total number of infections in the population at a particular time. However, outbreak data often consist of incidence counts, which are the number of new cases accumulated in each interobservation interval. Extending our DA algorithm to accommodate incidence data is an important next step and should be straightforward in situations where the state space for the subject level process is finite—for instance, if a subject cannot become reinfected more than once or twice in a given interobservation interval. We also believe it is important to investigate whether there is a way to make our DA algorithm more efficient by selecting the subjects whose paths are resampled in each iteration in a way that maximizes the perturbation to the population-level path and does not invalidate the MCMC. Designing an optimal schedule of subject-path updates could be critical to being able to use our algorithm in fitting more complex models to data from epidemics in large, structured populations.

Supplementary Materials

S1-S5: Additional mathematical and algorithmic details. S6-12: MCMC specification, supplementary results, and convergence diagnostics for simulations. S13: MCMC specification, additional results, and supplementary analysis of the British boarding school data.

Acknowledgments

J.F., J.W., and V.N.M. were supported by the NIH grant U54 GM111274. J.W. was supported by the NIH grant R01 CA095994. V.N.M. was supported by the NIH grant R01 AI107034. The authors thank Aaron King and the rest of the authors of the pomp package for their help with the PMMH algorithm that served as a benchmark for the methods presented in this article.

References

Allen, L. J. S. (2008), "An Introduction to Stochastic Epidemic Models," in *Mathematical Epidemiology*, eds. F. Brauer, P. van den Driessche, and J. Wu, New York: Springer, pp. 81–130. [918]

Andersson, H., and Britton, T. (2000), *Stochastic Epidemic Models and Their Statistical Analysis* (*Lecture Notes in Statistics*), New York: Springer. [918,920]

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Series B, 72, 269–342. [919,924]

Anon. (1978), "Influenza in a Boarding School," *The British Medical Journal*, 1, 587. [927]

Auranen, K., Arjas, E., Leino, T., and Takala, A. K. (2000), "Transmission of Pneumococcal Carriage in Families: A Latent Markov Process Model for Binary Longitudinal Data," *Journal of the American Statistical Association*, 95, 1044–1053. [919]

Becker, N. G. (1977), "On A General Stochastic Epidemic Model," *Theoretical Population Biology*, 11, 23–36. [918]

Bretó, C., and Ionides, E. L. (2011), "Compound Markov Counting Processes and their Applications to Modeling Infinitesimally Over-Dispersed Systems," *Stochastic Processes and their Applications*, 121, 2571–2591. [924]

Britton, T. (2010), "Stochastic Epidemic Models: A Survey," *Mathematical Biosciences*, 225, 24–35. [918]

Cappé, O., Moulines, E., and Ryden, T. (2006), *Inference in Hidden Markov Models* (*Springer Series in Statistics*), New York: Springer. [919]

Cauchemez, S., Carrat, F., Viboud, C., and Valleron, A. J. (2004), "A Bayesian MCMC Approach to Study Transmission of Influenza: Application to Household Longitudinal Data," *Statistics in Medicine*, 23, 3469–3487. [919]

Cauchemez, S., and Ferguson, N. M. (2008), "Likelihood-Based Estimation of Continuous-Time Epidemic Models from Time-Series Data: Application to Measles Transmission in London," *Journal of the Royal Society Interface*, 5, 885–897. [918]

Centers for Disease Control and Prevention (2014), "How Flu Spreads," available at *http://www.cdc.gov/flu/about/disease/spread.htm* [928]

Davies, J. R., Smith, A. J., Grilli, E. A., and Hoskins, T. W. (1982), "Christ's Hospital 1978–79: An Account of Two Outbreaks of Influenza A H1N1," *Journal of Infection*, 5, 151–156. [927]

Dukic, V., Lopes, H. F., and Polson, N. G. (2012), "Tracking Epidemics with Google Flu Trends Data and a State-Space SEIR Model," *Journal of the American Statistical Association*, 107, 1410–1426. [919]

Fintzi, J. (2016), "ECctmc: Simulation from Endpoint-Conditioned Continuous Time Markov Chains," R Package, Version 0.2.2, available at *https://github.com/fintzij/ECctmc* [922]

Gibson, G. J., and Renshaw, E. (1998), "Estimating Parameters in Stochastic Compartmental Models using Markov Chain Methods," *Mathematical Medicine and Biology*, 15, 19–40. [919]

Gillespie, D. T. (1976), "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions," *Journal of Computational Physics*, 22, 403–434. [922,924]

Glass, K., Xia, Y., and Grenfell, B. (2003), "Interpreting Time-Series Analyses for Continuous-Time Biological Models—Measles as a Case Study," *Journal of Theoretical Biology*, 223, 19–25. [919]

Held, L., Höhle, M., and Hofmann, M. (2005), "A Statistical Framework for the Analysis of Multivariate Infectious Disease Surveillance Counts," *Statistical Modelling*, 5, 187–199. [918]

Held, L., and Paul, M. (2012), "Modeling Seasonality in Space-Time Infectious Disease Surveillance Data," *Biometrical Journal*, 54, 824–843. [918]

Hirsch, M. W., Smale, S., and Devaney, R. L. (2013), *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, Waltham: Academic Press. [921]

Hobolth, A., and Stone, E. A. (2009), "Simulation from Endpoint-Conditioned, Continuous-Time Markov Chains on a Finite State Space, with Applications to Molecular Evolution," *The Annals of Applied Statistics*, 3, 1204–1231. [922]

Höhle, M., and Jørgensen, E. (2002), "Estimating Parameters for Stochastic Epidemics," Technical Report 102, The Royal Veterinary and Agricultural University. [919]

Ionides, E. L., Bhadra, A., Atchadé, Y., and King, A. A. (2011), "Iterated Filtering," *The Annals of Statistics*, 39, 1776–1802. [919]

Jandarov, R., Haran, M., Bjørnstad, O., and Grenfell, B. (2014), "Emulating a Gravity Model to Infer the Spatiotemporal Dynamics of an Infectious Disease," *Journal of the Royal Statistical Society*, Series C, 63, 423–444. [918]

Keeling, M. J., and Rohani, P. (2008), *Modeling Infectious Diseases in Humans and Animals*, Princeton: Princeton University Press. [918,928]

King, A. A., Nguyen, D., and Ionides, E. L. (2016), "Statistical Inference for Partially Observed Markov Processes Via the R Package pomp," *Journal of Statistical Software*, 69, 1–43. [924,927]

Koepke, A. A., Longini Jr, I. M., Halloran, M. E., Wakefield, J., and Minin, V. N. (2016), "Predictive Modeling of Cholera Outbreaks in Bangladesh," *The Annals of Applied Statistics*, 10, 575–595. [919]

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, 1170–1177. [918]

Lindenstrand, D., and Svensson, Å. (2013), "Estimation of the Malthusian Parameter in an Stochastic Epidemic Model Using Martingale Methods," *Mathematical Biosciences*, 246, 272–279. [918]

Longini Jr, I. M., and Koopman, J. S. (1982), "Household and Community Transmission Parameters from Final Distributions of Infections in Households," *Biometrics*, 38, 115–126. [918]

McKinley, T., Cook, A. R., and Deardon, R. (2009), "Inference in Epidemic Models without Likelihoods," *The International Journal of Biostatistics*, 5, 1–40. [919]

McKinley, T. J., Ross, J. V., Deardon, R., and Cook, A. R. (2014), "Simulation-Based Bayesian Inference for Epidemic Models," *Computational Statistics & Data Analysis*, 71, 434–447. [919]

Neal, P. J., and Roberts, G. O. (2004), "Statistical Inference and Model Selection for the 1861 Hagelloch Measles Epidemic," *Biostatistics*, 5, 249–261. [919]

O'Neill, P. D. (2002), "A Tutorial Introduction to Bayesian Inference for Stochastic Epidemic Models using Markov Chain Monte Carlo Methods," *Mathematical Biosciences*, 180, 103–114. [918]

—— (2009), "Bayesian Inference for Stochastic Multitype Epidemics in Structured Populations using Sample Data," *Biostatistics*, 10, 779–791. [919]

—— (2010), "Introduction and Snapshot Review: Relating Infectious Disease Transmission Models to Data," *Statistics in Medicine*, 29, 2069–2077. [918]

O'Neill, P. D., and Roberts, G. O. (1999), "Bayesian Inference for Partially Observed Stochastic Epidemics," *Journal of the Royal Statistical Society*, Series A, 162, 121–129. [919]

Pooley, C. M., Bishop, S. C., and Marion, G. (2015), "Using Model-Based Proposals for Fast Parameter Inference on Discrete State Space, Continuous-Time Markov Processes," *Journal of The Royal Society Interface*, 12, 20150225. [919,920]

Qin, Z., and Shelton, C. R. (2015), "Auxiliary Gibbs Sampling for Inference in Piecewise-Constant Conditional Intensity Models," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 722–731. [919]

Roberts, G. O., and Stramer, O. (2001), "On Inference for Partially Observed Nonlinear Diffusion Models using the Metropolis-Hastings Algorithm," *Biometrika*, 88, 603–621. [918,919]

Scott, S. L. (2002), "Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century," *Journal of the American Statistical Association*, 97, 337–351. [921]

Shelton, C. R., and Ciardo, G. (2014), "Tutorial on Structured Continuous-Time Markov Processes," *Journal of Artificial Intelligence Research*, 51, 725–778. [919]

Shestopaloff, A. Y., and Neal, R. M. (2016), "Sampling Latent States for High-Dimensional Non-Linear State Space Models with the Embedded HMM Method," arXiv preprint arXiv:1602.06030v2. [919]

Sudbury, A. (1985), "The Proportion of the Population Never Hearing a Rumour," *Journal of Applied Probability*, 22, 443–446. [918]

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009), "Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems," *Journal of the Royal Society Interface*, 6, 187–202. [919]

Watson, R. (1981), "An Application of a Martingale Central Limit Theorem to the Standard Epidemic Model," *Stochastic Processes and Their Applications*, 11, 79–89. [918]

Wearing, H. J., Rohani, P., and Keeling, M. J. (2005), "Appropriate Models for the Management of Infectious Diseases," *PLOS Medicine*, 2, e174. [928]