Efficient Data Augmentation for Fitting Stochastic Epidemic Models to Prevalence Data

Jonathan Fintzi[1], Jon Wakefield[1,2], and Vladimir Minin[2,3]

[1]Department of Biostatistics, University of Washington, Seattle
[2]Department of Statistics, University of Washington, Seattle
[3]Department of Biology, University of Washington, Seattle

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 unobserved 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 trajectories. In our MCMC algorithm, we propose each new subject–level path, conditional on the data, using a time–inhomogeneous continuous–time Markov process with rates determined by the infection histories of other individuals. The method is general, and may be applied, with minimal modifications, to a broad class of stochastic epidemic models. We present our algorithm in the context of a general stochastic epidemic model 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.

Keywords: Bayesian data augmentation, continuous–time Markov chain, epidemic count data, hidden Markov model, stochastic compartmental model

1 Introduction

Stochastic epidemic models (SEMs) are classic tools for modeling the spread of infectious diseases. A 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, along with conclusions we draw about its dynamics (Keeling and Rohani, 2008). Moreover, many questions — e.g., what is the final size distribution? What is the probability that a disease has been eradicated? — cannot be answered using deterministic methods (Britton, 2010).

The task of fitting a SEM is typically complicated by the limited extent of epidemiological data, which are recorded at discrete observation times and commonly describe just one aspect of the disease process, e.g. infections, and usually capture only a fraction of cases. Subject–level data, such as infection and recovery times, are rarely available (O'Neill, 2010). Fitting SEMs in the absence of complete subject–level data represents a complicated latent variable problem since it is usually impossible to analytically integrate over the missing data when an epidemic is not fully observed (ONeill, 2002). This makes the likelihood for a 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 that reflect events within the SEM, e.g. infections and recoveries (Becker, 1977, Watson, 1981, Sudbury, 1985, Andersson and Britton, 2000, Lindenstrand and Svensson, 2013). The resulting estimates are specific to the SEM dynamics and are not easily implemented for more complex dynamics. Approximation methods replace the epidemic model with a simpler model whose likelihood is analytically tractable. For example, Roberts and Stramer (2001) and Cauchemez and Ferguson (2008) use diffusion processes that approximate the SEM dynamics, while Jandarov et al. (2014) use a Gaussian process approximation of a related gravity model. Another typical simplification is to discretize time into "generational units", and to construct a transition model for the population flow at each generation time (Longini Jr. and Koopman, 1982, Held et al., 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 realistic. For instance, the discretization of time makes it awkward to approximate systems in which the observation times are not evenly spaced or the rates of events span several orders of magnitude (Glass et al., 2003, Shelton and Ciardo, 2014). Finally, simulation based methods use the underlying model to generate trajectories that serve as the basis for inference. This class of methods includes approximate Bayesian computation (ABC) methods (McKinley et al., 2009), pseudo–marginal methods (McKinley et al., 2014), and sequential Monte Carlo (or particle filter) methods (Toni et al., 2009, Ionides et al., 2011, Dukic et al., 2012, Koepke et al., In press). 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 et al., 2006, Dukic et al., 2012).

Traditional DA methods for fitting SEMs, first presented by ONeill 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. These methods have been used fruitfully in analyzing epidemics occurring in small to moderate sized populations in settings where some subject–level data is available. A significant advantage to agent–based DA is that household structure and subject–level covariates may be incorporated into the model (Auranen et al., 2000, Hohle

and Jorgensen, 2002, Cauchemez et al., 2004, Neal and Roberts, 2004, Jewell et al., 2009, O'Neill, 2009). However, existing DA methods suffer from convergence issues as the observed information becomes small relative to the missing data (Roberts and Stramer, 2001, McKinley et al., 2014, Pooley et al., 2015). The *de facto* need for some subject–level data has precluded the use of classical DA machinery in many settings. Development of DA methods for SEMs is of continuing interest, and recent works by Pooley et al. (2015) and Qin and Shelton (2015) have presented methods that do not rely on subject–level data, although their algorithms forgo the flexibility of agent–based DA.

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 missing data and the model parameters as we alternate between updating subject–level paths and model parameters. We propose subject–paths, conditionally on the data, using a time–inhomogeneous 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 make our method practical for analyzing epidemic count data, even in moderately large population settings. Our MCMC algorithm requires no tuning and converges quickly. Furthermore, our algorithm does not require subject–level data, and thus enables exact Bayesian inference for SEMs fit to datasets that would have been impossible to study with existing DA methods. Finally, our algorithm is not specific to any particular SEM dynamics or measurement process, and thus may be applied, with minimal modifications, to a broad class of SEMs.

2 SIR Model and Data Augmentation Algorithm

For concreteness and clarity of exposition, we present our DA algorithm in the context of fitting a stochastic Susceptible–Infected–Recovered (SIR) model to disease prevalence data. This model describes the time evolution of an epidemic in terms of the disease histories of the individuals as they transition through three disease states — susceptible (S), infected/infectious (I), and recovered (R). For simplicity, we assume a closed, homogeneously mixing population in which each individual becomes infected, and hence infectious, immediately upon coming into contact with an infected person. We also assume that recovery confers lifelong immunity 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, if 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 detection probability ρ, of the true prevalence at each observation time. Thus,

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

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 in continuous–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

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

where $\tau_{\mathrm{I}}^{(j)}$ and $\tau_{\mathrm{R}}^{(j)}$ are the infection and recovery times for subject j (if they occur, possibly not in $[t_1, t_L]$). 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, e.g. 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–homogeneous CTMC, with transition rate from configuration $\mathbf{x}$ to $\mathbf{x}'$ given by

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

At the first observation time, we let $\mathbf{X}(t_1)|\mathbf{p}_{t_1} \sim \text{Categorical}\left(\{S, I, R\}, \mathbf{p}_{t_1}\right)$, where $\mathbf{p}_{t_1} = (p_S, p_I, p_R)$ are the probabilities that an individual is susceptible, infected, or recovered. Let $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_K\}$ be the (ordered) infection and recovery times, and let $\mathbb{I}\left(\tau_k \triangleq I\right)$ and $\mathbb{I}\left(\tau_k \triangleq R\right)$ indicate if τ_k is an infection or recovery time. Let $\boldsymbol{\theta} = (\beta, \mu, \rho, \mathbf{p}_{t_1})$. The complete data likelihood is

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

We briefly reconcile what might seem like a discrepancy between our SIR model and the canonical construction of the model (see Andersson and Britton (2000)). 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 canonical SIR model describes the time evolution of the compartment counts, and thus evolves on the lumped state space of counts. The canonical construction would have been appropriate had we chosen to perform DA in terms of counts (for example, as in Pooley

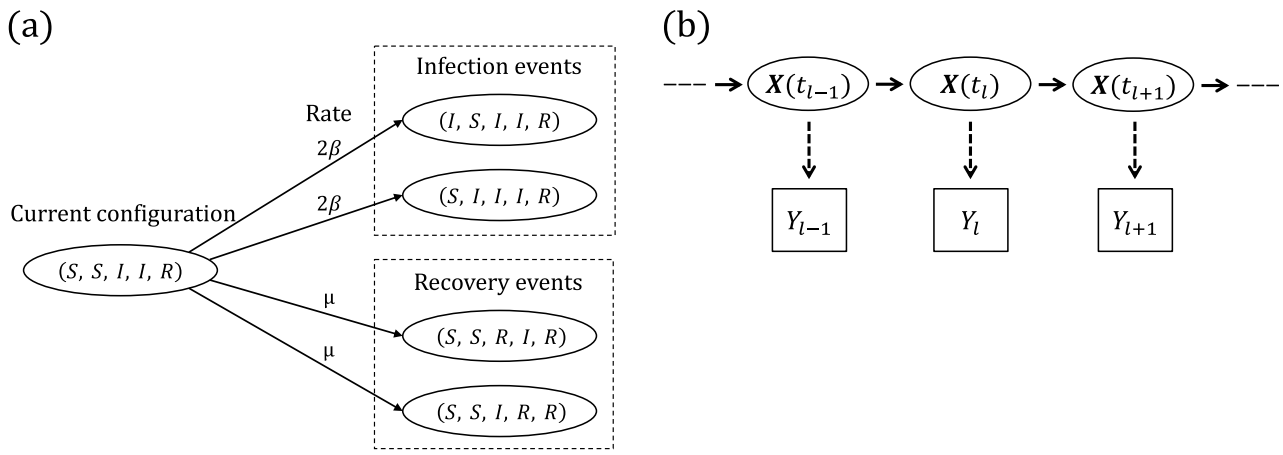

Figure 1: (a) SIR dynamics in a population of five subjects. The number of infecteds can increase from two to three via an infection of the first or second subject, reaching each of those configurations at rate 2β. The number of recovered individuals can increase from one to two via a recovery of the 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 $\mathbf{Y}_\ell | I_{t_\ell}, \rho \sim \text{Binomial}(I_{t_\ell}, \rho)$.

et al. (2015)). However, the Markov process in the canonical model is a lumping of our process with respect to the partition induced by aggregating the individuals in each model compartment. Therefore, inference made on the full subject–level state space will exactly match inference based on the canonical model. We discuss this further in Section S1 of the Supplement.

2.3 Subject–path proposal framework

The observed data likelihood term in the posterior $\pi(\boldsymbol{\theta}|\mathbf{Y}) \propto \int L(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \pi(\mathbf{X}|\boldsymbol{\theta}) \pi(\theta) \mathrm{d}\pi(\mathbf{X})$, is analytically and numerically intractable for even moderately sized N. We can obtain the tractable complete data likelihood in (4) by introducing the latent epidemic process, $\mathbf{X}$. The now tractable joint posterior distribution is

$$\pi(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}) \propto \Pr(\mathbf{Y}|\mathbf{X}, \rho) \times \Pr\left(\mathbf{X}(t_1)|\mathbf{p}_{t_1}\right) \times \pi(\mathbf{X}|\mathbf{X}(t_1), \beta, \mu) \times \pi(\beta)\pi(\mu)\pi(\rho)\pi(\mathbf{p}_{t_1}), \tag{5}$$

where $\pi(\beta)$, $\pi(\mu)$, $\pi(\rho)$, and $\pi(\mathbf{p}_{t_1})$ are prior densities. Our MCMC targets the joint posterior distribution (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 path of a single subject $\mathbf{X}_j$, conditionally on the data, using a time–inhomogeneous CTMC on the state space $\mathcal{S}_j$ with rates by the disease histories of all other individuals, $\mathbf{x}_{(-j)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_N\}$. The updated 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 $\boldsymbol{\tau}^{(-j)} = \{t_1, t_L\} \cup \{\boldsymbol{\tau} \backslash \boldsymbol{\tau}^{(j)}\} = \left\{\tau_0^{(-j)}, \tau_1^{(-j)}, \ldots, \tau_{M-1}^{(-j)}, \tau_M^{(-j)}\right\}$, where $t_1 = \tau_0^{(-j)}$ and

$t_L = \tau_M^{(-j)}$, to be the (ordered) times when other subjects become infected or recover, along with t_1 and t_L. Let $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$ be the intervals that partition $[t_1, t_L]$, i.e. $\mathcal{I}_1 = \left[\tau_0^{(-j)}, \tau_1^{(-j)}\right)$, $\mathcal{I}_2 = \left[\tau_1^{(-j)}, \tau_2^{(-j)}\right), \ldots,$ $\mathcal{I}_M = \left[\tau_{M-1}^{(-j)}, \tau_M^{(-j)}\right)$. Our Metropolis-Hastings proposal assumes that the CTMC for $\mathbf{X}_j$ is homogeneous within inter–event intervals. Let $I_\tau^{(-j)} = \sum_{i \neq j} \mathbb{I}\left(\mathbf{X}_i(\tau) = I\right)$ be the prevalence at time τ, excluding subject j. Define the rate matrices $\mathbf{\Lambda}^{(-j)} = \left\{\mathbf{\Lambda}_1^{(-j)}(\boldsymbol{\theta}), \ldots, \mathbf{\Lambda}_M^{(-j)}(\boldsymbol{\theta})\right\}$ corresponding to each interval in $\mathcal{I}$, where for $m = 1, \ldots, M$, the rate matrix for subject j is

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

We can construct the transition probability matrix for each interval,

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

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

$$\mathbf{P}^{(j)}(\tau_{m-1}, \tau_m) = \exp\left[(\tau_m - \tau_{m-1})\mathbf{\Lambda}_s^{(-j)}(\boldsymbol{\theta})\right]. \tag{7}$$

This computation requires an eigen-decomposition of each rate matrix, the parts of which are cached. By the Markov property, the time–inhomogeneous CTMC density over the observation period $[t_1, t_L]$, denoted $\pi(\mathbf{X}_j | \mathbf{x}_{(-j)}, \boldsymbol{\theta}) = \pi\left(\mathbf{X}_j | \mathbf{\Lambda}^{(-j)}; \mathcal{I}\right)$, can be written as the product of time–homogeneous CTMC densities over the inter–event intervals $\mathcal{I}_1, \ldots, \mathcal{I}_M$. Thus,

$$\pi\left(\mathbf{X}_j | \mathbf{\Lambda}^{(-j)}; \mathcal{I}\right) = \Pr(\mathbf{X}_j(t_1) | \mathbf{p}_{t_1}) \prod_{m=1}^{M} \pi\left(\mathbf{X}_j | \mathbf{X}_j(\tau_{m-1}), \mathbf{\Lambda}_m^{(-j)}(\boldsymbol{\theta}); \mathcal{I}_m\right). \tag{8}$$

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 sub–intervals in $\mathcal{I}_\ell$, within which the CTMC is time–homogeneous. Thus, the transition probability matrix over an inter–observation interval, $\mathcal{I}_\ell = [t_{\ell-1}, t_\ell]$, containing inter–event intervals whose endpoints are given by times $t_{\ell-1} = \tau_{\ell,0}^{(-j)} < \tau_{\ell,1}^{(-j)} < \cdots < \tau_{\ell,S-1}^{(-j)} <_{\ell,S}^{(-j)} = t_\ell$, is constructed as

$$\mathbf{P}^{(j)}(t_{\ell-1}, t_\ell) = \prod_{s=1}^{S} \mathbf{P}^{(j)}(\tau_{\ell,s-1}^{(-j)}, \tau_{\ell,s}^{(-j)}). \tag{9}$$

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

1. *HMM step*: sample the disease state at 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–inhomogeneous CTMC rates change, conditional on the states sampled in the HMM step.
3. *Event time step*: sample the exact transition times conditional on the discrete sequence of states drawn in the previous steps.

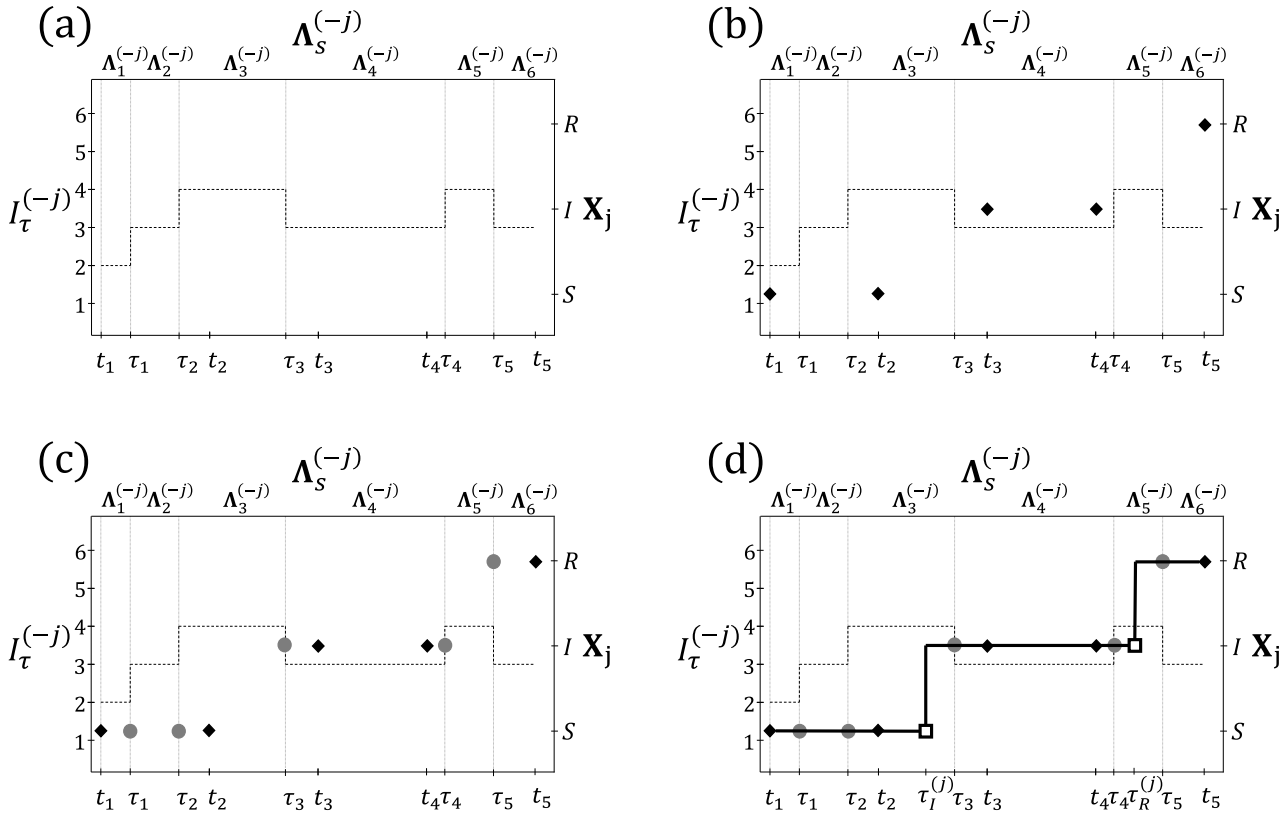

Figure 2: Procedure for constructing a subject–path proposal. (a) The dashed line depicts the number of infected individuals, excluding $\mathbf{X}_j$, the subject whose path is being sampled. The observation times, $t_1, \ldots, t_5$, and times at which other subjects change disease states, $\tau_1,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

2.3.1 HMM step

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

$$Y_\ell | X_j(t_\ell), I_{t_\ell}^{(-j)}, \rho \sim \text{Binomial}\left(\mathbb{I}\left(X_j(t_\ell) = I\right) + I_{t_\ell}^{(-j)}, \rho\right). \tag{10}$$

The emission probability in (10) only depends on whether subject j is infected at time t_ℓ, since we treat the parameters and other subjects as fixed. Furthermore, the observations 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 1b).

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 standard 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 Section S2 of the Supplement.

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 inter–observation interval. We may reduce our problem to the time–homogeneous 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 inter–event interval. Consider an inter–observation interval, $\mathcal{I}_\ell = [t_{\ell-1}, t_\ell]$, containing inter–event intervals whose endpoints are given by times $t_{\ell-1} = \tau_{\ell,0}^{(-j)} < \tau_{\ell,1}^{(-j)} < \cdots < \tau_{\ell,n-1}^{(-j)} <_{\ell,n}^{(-j)} = t_\ell$, and let $x_i = \mathbf{x}_j(\tau_i^\ell)$. We recursively sample $\mathbf{X}_j$ at each intermediate event time, beginning at τ_1^ℓ, from the discrete distribution with masses

$$\begin{aligned}
\Pr\left(\mathbf{X}_j(\tau_i^\ell) = x_i | \mathbf{X}_j(\tau_{i-1}^\ell) = x_{i-1}, \mathbf{X}_j(\tau_n^\ell) = x_n\right) &= \frac{\Pr\left(\mathbf{X}_j(\tau_i^\ell) = x_i, \mathbf{X}_j(\tau_{i-1}^\ell) = x_{i-1}, \mathbf{X}_j(\tau_n^\ell) = x_n\right)}{\Pr\left(\mathbf{X}_j(\tau_{i-1}^\ell) = x_{i-1}, \mathbf{X}_j(\tau_n^\ell) = x_n\right)} \\
&= \frac{\Pr\left(\mathbf{X}_j(\tau_i^\ell) = x_i | \mathbf{X}_j(\tau_{i-1}^\ell) = x_{i-1}\right) \Pr\left(\mathbf{X}_j(\tau_n^\ell) = x_n | \mathbf{X}_j(\tau_i^\ell) = x_i\right)}{\Pr\left(\mathbf{X}_j(\tau_n^\ell) = x_n | \mathbf{X}_j(\tau_{i-1}^\ell) = x_{i-1}\right)} \\
&= \frac{\left[\mathbf{P}^{(j)}(\tau_{i-1}^\ell, \tau_i^\ell)\right]_{x_{i-1}, x_i} \left[\prod_{k=i}^{n-1} \mathbf{P}^{(j)}(\tau_k^\ell, \tau_{k+1}^\ell)\right]_{x_i, x_n}}{\left[\prod_{k=i-1}^{n-1} \mathbf{P}^{(j)}(\tau_k^\ell, \tau_{k+1}^\ell)\right]_{x_{i-1}, x_n}}.
\end{aligned} \tag{11}$$

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 exist a variety of efficient methods (Hobolth and Stone, 2009). 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. A fast implementation is available through the `ECctmc` package in `R` (Fintzi, 2016). We briefly summarize the modified rejection sampling algorithm in section S3 of the supplement, and refer the reader to Hobolth and Stone (2009) for an excellent discussion of methods for simulating paths of endpoint–conditioned time–homogeneous CTMCs.

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–inhomogeneous CTMC in our proposal. 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 path proposal with probability

$$\begin{aligned} a_{\mathbf{x}^{\text{cur}} \longrightarrow \mathbf{x}^{\text{new}}} &= \min\left\{ \frac{\pi(\mathbf{x}^{\text{new}}|\mathbf{y})}{\pi(\mathbf{x}^{\text{cur}}|\mathbf{y})} \frac{q(\mathbf{x}^{\text{cur}}|\mathbf{x}^{\text{new}}, \mathbf{y})}{q(\mathbf{x}^{\text{new}}|\mathbf{x}^{\text{cur}}, \mathbf{y})},\ 1 \right\} \\ &= \min\left\{ \frac{\pi(\mathbf{x}^{\text{new}})}{\pi(\mathbf{x}^{\text{cur}})} \frac{\pi\left(\mathbf{x}_j^{\text{cur}}|\boldsymbol{\Lambda}^{(-j)}; \mathcal{I}\right)}{\pi\left(\mathbf{x}_j^{\text{new}}|\boldsymbol{\Lambda}^{(-j)}; \mathcal{I}\right)}, 1 \right\}, \end{aligned} \tag{12}$$

where we have suppressed the dependence on $\boldsymbol{\theta}$. Hence, the Metropolis–Hastings ratio is equal to the ratio of population-level time–homogeneous CTMC densities, multiplied by the ratio of time–inhomogeneous CTMC proposal densities (see Section S4 of the Supplement for 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 non–zero probability. A sufficient condition for this is that the number of infected individuals is greater than the observed prevalence at each observation time. For the sake of efficiency, we simulate paths using the canonical SIR model on the lumped state space of compartment counts (Andersson and Britton, 2000). A valid population–level path may then be mapped to a collection of subject–level paths by selecting the subject for each infection or recovery uniformly at random from among the subjects who are at risk for that event. Thus, the individual associated with a particular infection event is sampled uniformly at random from among the subjects who are susceptible just prior to that infection time, while the subject associated with a given recovery event is sampled uniformly at random from among the individuals who are infected just prior to that recovery time.

2.4 Parameter updates

One MCMC iteration includes a number of subject–path updates, followed by a set of parameter updates. We will discuss the choice of how many subject–path updates to undertake per set of parameter updates in a subsequent section. Conjugate priors are available for all our model parameters. Thus, we use Gibbs sampling to draw new parameter values from their full conditional distributions (shown in Table 1). We refer the reader to (Hohle and Jorgensen, 2002) for a derivation of the full conditional distributions.

Parameter	Conjugate Prior Dist.	Prior Hyperparameters	Full Conditional Hyperparameters
β	Gamma	$a_\beta,\ b_\beta$	$a_\beta + \sum_{j=1}^{M} \mathbb{I}\left(I_{\tau_j}\right),\ b_\beta + \sum_{j=1}^{M} S_{\tau_{j-1}} I_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
μ	Gamma	$a_\mu,\ b_\mu$	$a_\mu + \sum_{j=1}^{M} \mathbb{I}\left(R_{\tau_j}\right),\ b_\mu + \sum_{j=1}^{M} I_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
ρ	Beta	$a_\rho,\ b_\rho$	$a_\rho + \sum_{j=1}^{L} Y_{t_j},\ b_\rho + \sum_{j=1}^{L} (I_{t_j} - Y_{t_j})$
$\mathbf{p}_{t_1}$	Dirichlet	$a_S,\ b_I,\ c_R$	$a_S + S_{t_1},\ b_I + I_{t_1},\ c_R + R_{t_1}$

Table 1: Prior and full conditional distributions for SIR model parameters. Gamma priors are parameterized with rates, so a Gamma(a, b) distribution has mean a/b.

2.5 Implementation

We provide the `R` and `C++` code base for this paper, along with examples, 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 here along with improvements to the implementation, will be incorporated into the `stemr` package (`https://github.com/fintzij/stemr`).

3 Simulation Results

3.1 Inference for the posterior distribution of the latent process

We simulated an epidemic in a population of size $N = 750$ with proportions of initially susceptible, infected, and recovered individuals of 0.9, 0.03, and 0.07. The per–contact infectivity rate was $\beta = 0.00185$ and the recovery rate was $\mu = 0.5$, which together correspond to a basic reproduction number of $R_0 = \beta N/\mu = 2.78$ and a mean infectious duration of two days. Binomially sampled prevalence counts were drawn at observation times 0, 1,..., 15 with sampling probability $\rho = 0.2$. We ran three MCMC chains, each for 250,000 iterations, updating the paths of ten subjects, chosen

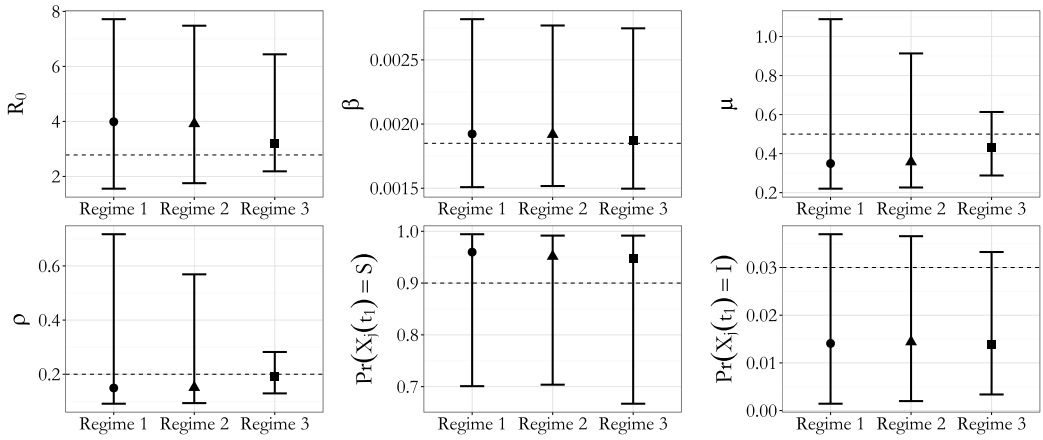

Figure 3: Posterior median estimates and 95% credible intervals for all model parameters under three different prior regimes (Table 2). Regime 1 used diffuse priors for all model parameters. Regime 2 used informative priors for the rate parameters but retained a diffuse prior for the binomial sampling parameter. Regime 3 had informative priors for the rate and sampling parameters.

uniformly at random, per MCMC iteration. Traceplots of the log–likelihood and parameters were monitored for convergence and are shown in section S5 of the supplement.

	Prior Distribution		
Parameter	Regime 1	Regime 2	Regime 3
$R_0 = 2.78$	Beta′(0.001, 0.001)	$\frac{1}{4}\times$Beta′(0.002, 2.1)	$\frac{1}{4}\times$Beta′(0.002, 2.1)
$\beta = 0.00185$	Gamma(0.001, 0.001)	Gamma(0.002, 1)	Gamma(0.002, 1)
$\mu = 0.5$	Gamma(0.001, 0.001)	Gamma(2.1, 4)	Gamma(2.1, 4)
$\rho = 0.2$	Beta(1, 1)	Beta(0.667, 1)	Beta(21, 75)

Table 2: True parameter values and prior distributions under three different prior regimes. The implied prior shown for R_0 is a scaled Beta Prime distribution, based on the priors for β and μ. Regime 1: diffuse priors for all parameters. Regime 2: informative priors for the rate parameters, diffuse priors for the sampling probability. Regime 3: informative priors all parameters. The same Dirichlet(9, 0.2, 0.5) prior for $\mathbf{p}_{t_1}$ was used in all three regimes.

We estimated the SIR model parameters under three regimes of model priors (summarized along with the implied prior distribution for R_0 in Table 2). Under the first regime, we assumed vague prior distributions over the rate parameters and the binomial sampling probability. The second regime reflected knowledge over a plausible range of disease dynamics, but remained vague regarding the binomial sampling probability. The third prior regime specified informative priors for all model parameters. The prior distribution of the initial state probabilities under all three prior regimes was taken to be Dirichlet(9, 0.2, 0.5).

The true values for all model parameters fell within the 95% credible intervals under all three prior regimes. The choice of prior regime did not affect the acceptance rate for subject–path proposals, and was roughly 93% acceptance under all three prior regimes. However, the widths of Bayesian credible intervals (Figure 3), and the widths of pointwise posterior distributions for the

latent process (Figure 4) were sensitive to the prior regime. In particular, additional information about the detection probability improved inference for both the model parameters and the latent process. Finally, posterior samples of the recovery rate and binomial sampling probability were highly correlated (supplementary Figures S2, S4, and S6).

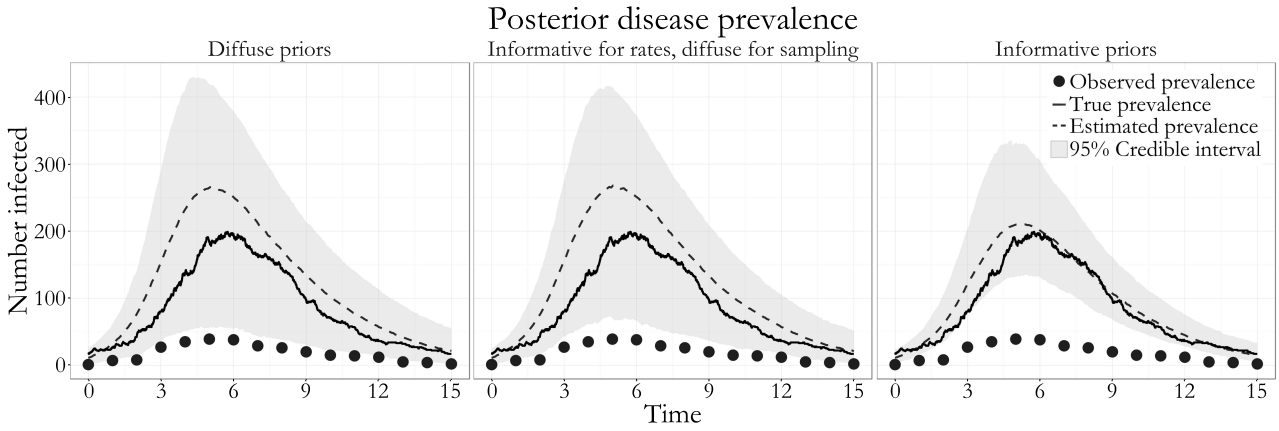

Figure 4: Estimated posterior distributions of the number of infected individuals under the three prior regimes. Depicted are the true unobserved prevalence (solid line), observed data (dots), pointwise posterior median prevalence (dashed line), and pointwise 95% credible intervals (shaded region). Latent posterior estimates are based on a thinned sample, with every 250^{th} sample retained.

3.2 Selecting the proportion of subject–paths to sample in each MCMC iteration

There is no need to re–sample the path of every subject within each MCMC iteration, and in the last section we sampled 10 subject–paths per iteration. Indeed, we might suspect that the efficiency of our MCMC could be improved by sampling only a few subject–paths between parameter updates. Subject–path proposals could result in high autocorrelation, as is the case for traditional DA methods (Roberts and Stramer, 2001), and frequently updating model parameters may help to break this correlation. Parameter updates also tend to produce high autocorrelation. However, subject–path proposals are costly compared to updates of model parameters. Therefore, it is reasonable to suspect that the effective sample size (ESS) per CPU time might be improved by sampling only a handful of subject–paths per MCMC iteration.

Many factors, including the SEM dynamics, population size, and efficiency of the implementation, could affect the optimal number subject–path updates per MCMC iteration. Rather than attempt to disentangle all the variables that could be involved, we performed a simulation to determine how prior immunity in the population might affect the MCMC efficiency. In populations with high immunity, the escape probability is high. Therefore, we might need to sample more subjects in order to meaningfully perturb the collection of subject–paths.

We simulated epidemics with SIR dynamics in two populations of size 500 that differed in their levels of prior immunity. In the first population, where prior immunity was only 5%, we found

that updating two subject paths per MCMC iteration maximized the ESS per CPU time and the estimated relative run length time required to estimate all posterior medians with accuracy ± 0.025 with probability 0.9, calculated using the Raftery-Lewis diagnostic in the `coda` package in `R` (Plummer et al., 2006). In the second population, where 35% of the individuals were immune *a priori*, the choice of ten subjects per iteration was optimal in terms of ESS per CPU time, while the estimated relative run length time required to achieve convergence was fairly flat until we began re-sampling tens of subjects. Thus, updating only a small fraction of the subject–paths per MCMC iteration yielded optimal perturbations to the latent process, but the optimal number of subject–paths per iteration also depended on the escape probability. This simulation also suggests that identifying an efficient number of subject–paths to update in each MCMC iteration could be accomplished relatively quickly, as we need only search over small numbers of subjects.

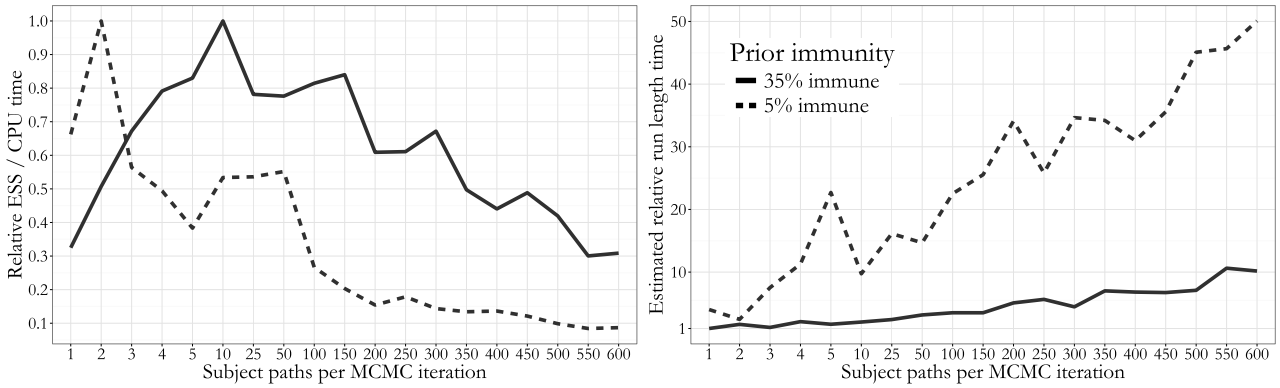

Figure 5: Performance of the MCMC algorithm versus the number of subject-level path updates per iteration in fitting SIR models to epidemics in populations with low prior immunity (dashed line) and moderate prior immunity (solid line). (a) Relative ESS of the log posterior per CPU time. (b) Estimated relative time to complete an MCMC run of sufficient length to estimate all posterior medians with accuracy $\pm$ 0.025 with probability 0.9, based on the Raftery-Lewis diagnostic computed by the `coda` package in R (Plummer et al., 2006).

3.3 Robustness to population size misspecification

There are two basic limitations of agent-based DA methods. 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 substantially degraded once the assumed population size was greater than a few thousand people, though more efficient implementation could help to alleviate some computational bottlenecks. 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 very large, perturbations to the likelihood from re-sampling one subject at a time become relatively less significant. It is possible that jointly sampling multiple subject–paths may help to mitigate slow MCMC mixing in large populations.

To speed up computation, it may be desirable to assume a smaller population size than is actually the case, and to rescale parameter estimates. We investigated the effects of population size misspecification on the posterior parameter estimates and credible intervals by fitting SIR models using a sequence of assumed population sizes, 1,500 (the truth), 1,000, 500, 300, and 150, to data from two epidemics simulated under different dynamic regimes: $R_0 = 1.47$ and $R_0 = 3.675$, with $\mu = 1/7$ corresponding to a mean recovery time of one week, and an initial state distribution of $\mathbf{p}_{t_1} = (0.948, 0.002, 0.05)$. The data were recorded until the end of each epidemic, and were binomially distributed with sampling probability $\rho = 0.2$. We set informative priors for the β, μ, and $\mathbf{p}_{t_1}$, but specified a flat prior for ρ. A posterior sample of the model parameters under each assumed population size was obtained by combining the results from five chains, each run for 250,000 iterations of which the first 1,000 were discarded as burn-in. We updated subject–paths for 1% of the assumed population size, rounded up, in each MCMC iteration.

We compared scaled parameter estimates from the models fit under different assumed population sizes. The per–contact infectivity rate, β, was rescaled by the ratio of assumed and true population sizes, $N/1500$, so that it could be interpreted as the per-contact infectivity rate times the relative contact rate. We computed R_0 using the assumed population size. Finally, we multiplied the binomial sampling probability by the assumed population size so the resulting product could be interpreted as the expected number of observed infections in a completely infected population in which each infection contributes to one observation.

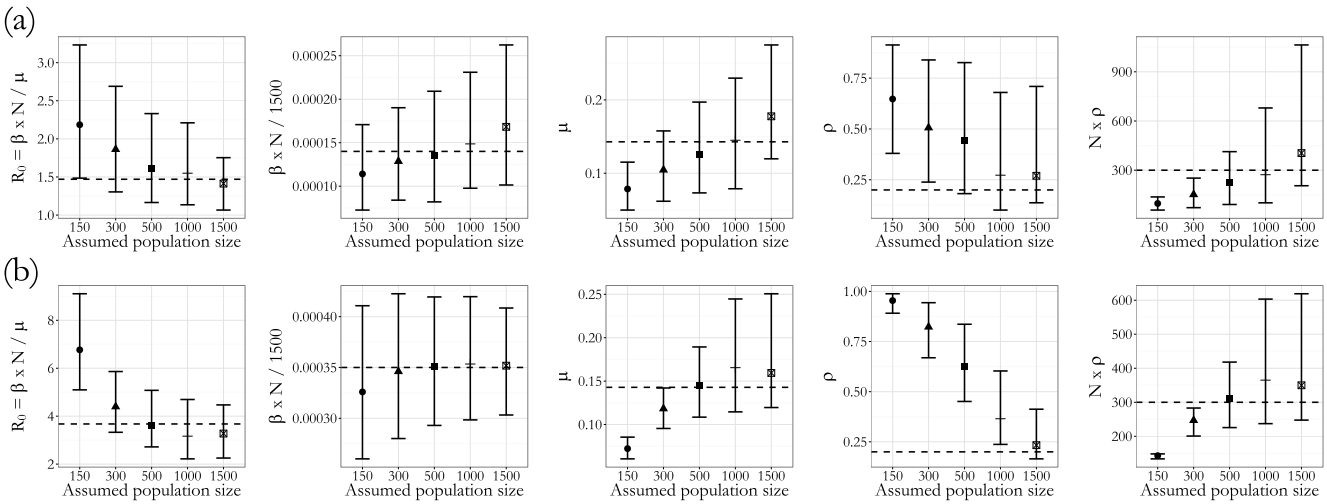

Figure 6: Posterior medians and 95% credible intervals for the basic reproductive number, R_0, infectivity rate, recovery rate, binomial sampling probability, and binomial sampling probability scaled by the assumed population size. The dashed lines indicate the true values in the population of size 1500. The population size, N, indicates the assumed population size used in fitting the model. (a) Estimates obtained for a slowly progressing epidemic, where $R_0 = 1.47$. (b) Estimates obtained for an epidemic where $R_0 = 3.675$.

Although we may achieve approximately valid inference under moderate population size misspecification, drastically understating the population size can lead to severe biases in parameter estimates (Figure 6). Estimates of R_0 exhibit severe upward bias, while estimates of the recovery rate are depressed. We conjecture that in a small assumed population, infected durations are longer since

latent prevalence must never be less than observed prevalence. This also forces the estimated detection probability close to one as the population size decreases. Finally, in the true population of size 1,500, we would expect to observe nearly 300 infections, since nearly the entire population was infected. For the assumed population sizes greater than 300, we notice that the credible interval for the expected number of observed cases still covers the true expected number. But, as the assumed population size decreases, the expected number of observed cases falls below the true expected number, since even a binomial sampling probability of $\rho = 1$ implies that $N \times \rho < 300$.

This simulation suggests a heuristic diagnostic for identifying an appropriate population size that both ensures the approximate validity of posterior estimates, while facilitating faster computation. Such an approach might be appropriate if time is of the essence, as in an outbreak response setting. In principle, separate MCMC chains could be run with a sequence of steadily increasing assumed population sizes. Once the posterior estimates have leveled off, there may be no need to fit the model using the true population size if it is large. Two quantities that appear to be particularly useful in diagnosing the sufficiency of an assumed population size, based on the experiment presented above, are R_0 and the expected number of observed cases in the assumed population, $N \times \rho$.

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 22^{nd} through February 4^{th} were accessed via the `pomp` package in `R` (King et al., 2016), and are displayed in Figure 7.

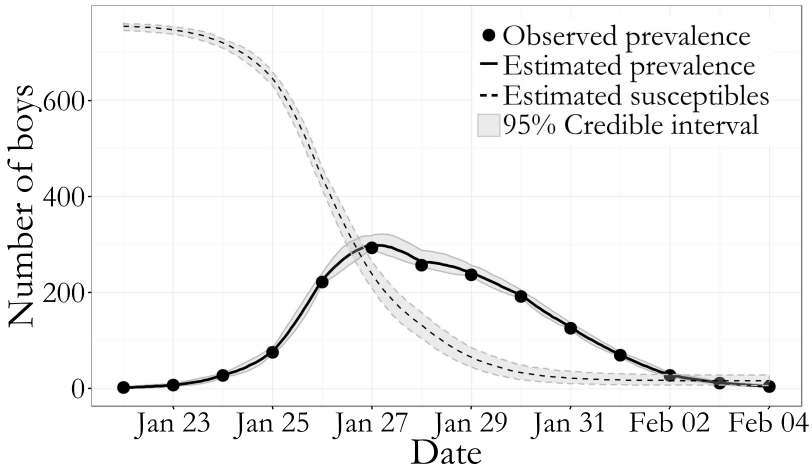

Figure 7: Boarding school data, pointwise posterior median estimates and pointwise 95% credible intervals (grey shaded areas) for the numbers of infected boys (solid line) and susceptible boys (dashed line). Posterior estimates based on a thinned sample, with every 250^{th} configuration retained.

We fit an SIR model to the data, running MCMC five chains in parallel, each for 250,000 iterations.

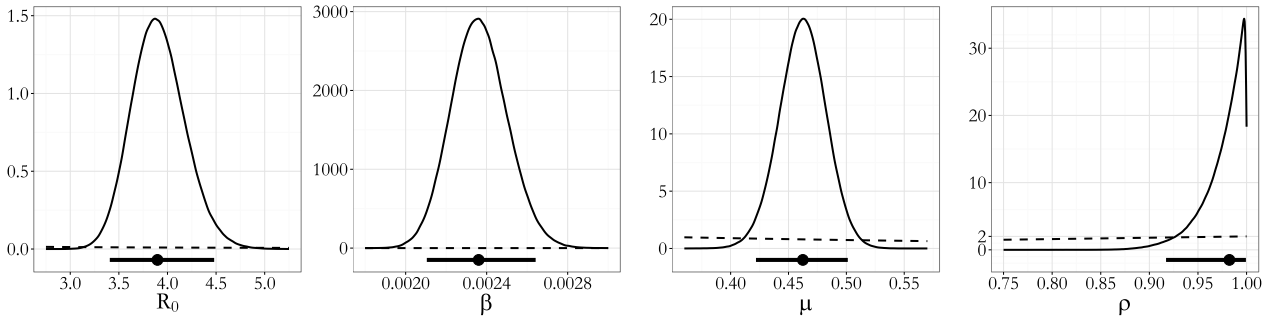

Figure 8: Posterior density estimates for the SIR model parameters fit to the British boarding school data (thin solid lines). The posterior median and 95% Bayesian credible intervals are drawn below the density plots (thick solid lines with circles). The induced prior density for R_0 and the prior densities for the model parameters are shown as dashed lines over the posterior ranges. The effective sample sizes were β: 5,139; μ: 16,130; ρ: 4,733; $p_{S_{t_1}}$: 10,461; $p_{I_{t_1}}$: 908,191; $p_{R_{t_1}}$: 8,620.

We discarded the first 1,000 iterations as burn-in, and sampled ten subject–paths per MCMC iteration. We also ran another set of three chains in parallel for 250,000 iterations each, in which we sampled 75 subject paths per MCMC iteration. The total run time for this second set of chains was over seven times longer than for the set of five chains and yielded virtually identical estimates. Convergence of the chains was assessed visually (see Section S7 of the Supplement), and the retained parameter samples and latent posterior paths following the burn–in period were combined to form the final posterior sample.

We specified a Beta(2,1) prior, which density linearly increases from on its support between 0 to 1, for ρ, since the majority, though perhaps not all, of the infections on any given day would likely have been detected in such a closely monitored environment. We selected diffuse priors for the infectivity and recovery rates, taking $\beta \sim$Gamma(0.001,1) and $\mu \sim$Gamma(1,2), the former being highly diffuse and the latter chosen by assuming that on average individuals developed symptoms and got isolated 2 days after the infection. Since roughly 90% of the students were eventually infected, it is likely that only a few students were either infected or immune prior to the start of the outbreak. Therefore, we set a Dirichlet(900, 3, 9) prior for the initial disease state probabilities.

The posterior median recovery rate corresponds to an average period of 2.17 days (95% BCI: 2.0, 2.38) during which an infectious individual could contact susceptible boys before being confined to the infirmary. Our posterior median estimate of R_0 was 3.90 (95% BCI: 3.41, 4.48). Previous analyses of this dataset, using trajectory matching, estimated the mean infectious duration to be roughly 2.2 days, and R_0 to be roughly 3.7 (Wearing et al., 2005, Keeling and Rohani, 2008). These estimates of R_0 are similar to those from other influenza outbreaks occurring in closed environments (Biggerstaff et al., 2014). Our posterior median estimates of the probabilities that an individual was susceptible, infected, or recovered at the start of the outbreak were, respectively, 0.99 (95% BCI: 0.98, 0.99), 0.003 (95% BCI: 0.001, 0.007), and 0.009 (0.004, 0.017). Finally, our posterior median estimate of ρ was 0.98 (95% BCI: 0.92, 1.00), suggesting that, while almost all of the infections were observed on each day, a handful of cases likely remained undetected. This is consistent with the typical progression of influenza, in which individuals are infectious for about a day before becoming symptomatic (Centers for Disease Control and Prevention).

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 impossible, to carry out using traditional DA methods in the absence of subject–level data. Although we have presented our DA algorithm in the context of fitting an SIR model to prevalence data, the machinery in our algorithm relied neither on the SEM dynamics, nor on the binomial distribution used to model the data. Therefore, our algorithm represents a general solution to the problem at hand, and may be applied, with minor modifications, to a broad class of SEMs. We have demonstrated that updating a small fraction of the subject–paths per MCMC iteration optimizes the MCMC efficiency. Furthermore, moderate misspecification of the assumed population size still yields approximately valid inference. Therefore, our DA algorithm is likely to be of practical use, even in analyzing epidemics in moderately large populations. However, we do not view this algorithm as a solution for analyzing epidemics in very large populations as run length times and MCMC mixing will eventually deteriorate. Still, in many scenarios, this DA algorithm will mitigate the need for extremely computationally expensive simulation–based methods, or for approximate methods. Finally, our DA algorithm is carried out entirely at the subject level, making it possible to also incorporate subject–level covariates and fit models based on subject–level data.

To conclude, we would like to comment on directions for future work that we intend to pursue. First, the DA algorithm in this paper 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, the epidemiological data often consist of incidence counts, which are the number of new cases accumulated in each inter-observation 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 inter-observation interval. Second, although we have shown that it suffices to sample only a small fraction of the subject-level paths, it is likely that our already efficient DA algorithm could be made even more so if the schedule of subject–paths to update could be chosen to maximally perturb the population level path. Finally, an obvious next step in assessing the usefulness of the algorithm is to fit SEMs with more complex dynamics to a variety of datasets.

6 Acknowledgements

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.

References

H. Andersson and T. Britton. Stochastic Epidemic Models and Their Statistical Analysis. Lecture Notes in Statistics. Springer, New York, 2000.

Anon. Influenza in a boarding school. The British Medical Journal, 1:587, 1978.

K. Auranen, E. Arjas, T. Leino, and A.K. Takala. Transmission of pneumococcal carriage in families: a latent Markov process model for binary longitudinal data. Journal of the American Statistical Association, 95:1044–1053, 2000.

N.G. Becker. On a general stochastic epidemic model. Theoretical Population Biology, 11:23–36, 1977.

M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir, and L. Finelli. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC Infectious Diseases, 14:480, 2014.

T. Britton. Stochastic epidemic models: a survey. Mathematical Biosciences, 225:24–35, 2010.

O. Cappé, E. Moulines, and T. Ryden. Inference in Hidden Markov Models. Springer Series in Statistics. Springer, New York, 2006.

S. Cauchemez and N.M. Ferguson. 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, 2008.

S. Cauchemez, F. Carrat, C. Viboud, and A.J. Valleron. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statistics in Medicine, 23: 3469–3487, 2004.

Centers for Disease Control and Prevention. How flu spreads, 2014. URL `http://www.cdc.gov/flu/about/disease/spread.htm`. Accessed on January 3, 2016.

J.R. Davies, A.J. Smith, E.A. Grilli, and T.W. Hoskins. Christ's Hospital 1978–79: An account of two outbreaks of influenza A H1N1. Journal of Infection, 5:151–156, 1982.

V. Dukic, H.F. Lopes, and N.G. Polson. Tracking epidemics with Google flu trends data and a state-space SEIR model. Journal of the American Statistical Association, 107:1410–1426, 2012.

J. Fintzi. ECctmc: Simulation from Endpoint-Conditioned Continuous Time Markov Chains, 2016. URL `https://github.com/fintzij/ECctmc`. R package, version 0.1.0.

G.J. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. Mathematical Medicine and Biology, 15:19–40, 1998.

D.T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics, 22:403–434, 1976.

K. Glass, Y. Xia, and B. Grenfell. Interpreting time-series analyses for continuous-time biological models - measles as a case study. Journal of Theoretical Biology, 223:19–25, 2003.

L. Held and M. Paul. Modeling seasonality in space-time infectious disease surveillance data. Biometrical Journal, 54:824–843, 2012.

L. Held, M. Höhle, and M. Hofmann. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Statistical modelling, 5:187–199, 2005.

A. Hobolth and E.A. Stone. 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, 2009.

M. Hohle and E. Jorgensen. Estimating parameters for stochastic epidemics. Technical Report 102, The Royal Veterinary and Agricultural University, November 2002.

E.L. Ionides, A. Bhadra, Y.Atchadé, A.A. King, et al. Iterated filtering. The Annals of Statistics, 39:1776–1802, 2011.

R Jandarov, M. Haran, O. Bjørnstad, and B. Grenfell. Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. Journal of the Royal Statistical Society: Series C (Applied Statistics), 63:423–444, 2014.

C.P. Jewell, T. Kypraios, P. Neal, and G.O. Roberts. Bayesian analysis for emerging infectious diseases. Bayesian Analysis, 4:465–496, 2009.

M.J. Keeling and P. Rohani. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, Princeton, 2008.

W.O. Kermack and A.G. McKendrick. A contribution to the mathematical theory of epidemics. In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volume 115, pages 700–721. The Royal Society, 1927.

A.A. King, D. Nguyen, and E.L. Ionides. Statistical inference for partially observed markov processes via the R package pomp. Journal of Statistical Software, 69:1–43, 2016. doi: 10.18637/jss.v069.i12.

A.A. Koepke, I.M. Longini Jr, M.E. Halloran, J. Wakefield, and V.N. Minin. Predictive modeling of cholera outbreaks in Bangladesh. Annals of Applied Statistics, In press.

P.E. Lekone and B.F. Finkenstädt. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. Biometrics, 62:1170–1177, 2006.

D. Lindenstrand and A. Svensson. Estimation of the Malthusian parameter in an stochastic epidemic model using martingale methods. Mathematical Biosciences, 246:272–279, 2013.

I.M. Longini Jr. and J.S. Koopman. Household and community transmission parameters from final distributions of infections in households. Biometrics, pages 115–126, 1982.

T. McKinley, A.R. Cook, and R. Deardon. Inference in epidemic models without likelihoods. The International Journal of Biostatistics, 5(1), 2009.

T.J. McKinley, J.V. Ross, R. Deardon, and A.R. Cook. Simulation-based Bayesian inference for epidemic models. Computational Statistics & Data Analysis, 71:434–447, 2014.

P.J. Neal and G.O. Roberts. Statistical inference and model selection for the 1861 Hagelloch measles epidemic. Biostatistics, 5:249–261, 2004.

P.D. O'Neill. Bayesian inference for stochastic multitype epidemics in structured populations using sample data. Biostatistics, 10:779–791, 2009.

P.D. O'Neill. Introduction and snapshot review: relating infectious disease transmission models to data. Statistics in Medicine, 29:2069–2077, 2010.

P.D. ONeill. A tutorial introduction to bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences, 180:103–114, 2002.

P.D. ONeill and G.O. Roberts. Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A (Statistics in Society), 162:121–129, 1999.

M. Plummer, N. Best, K. Cowles, and K. Vines. Coda: Convergence diagnosis and output analysis for MCMC. R News, 6(1):7–11, 2006.

C.M. Pooley, S.C. Bishop, and G. Marion. Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes. Journal of The Royal Society Interface, 12:20150225, 2015.

Z. Qin and C.R. Shelton. Auxiliary Gibbs sampling for inference in piecewise-constant conditional intensity models. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, 2015.

G.O. Roberts and O. Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. Biometrika, 88:603–621, 2001.

S.L. Scott. Bayesian methods for hidden Markov models. Journal of the American Statistical Association, 97(457), 2002.

C.R. Shelton and G. Ciardo. Tutorial on structured continuous-time Markov processes. Journal of Artificial Intelligence Research, 51:725–778, 2014.

A. Sudbury. The proportion of the population never hearing a rumour. Journal of Applied Probability, pages 443–446, 1985.

J.P. Tian and D. Kannan. Lumpability and commutativity of Markov processes. Stochastic Analysis and Applications, 24:685–702, 2006.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M.P.H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface, 6:187–202, 2009.

R. Watson. An application of a martingale central limit theorem to the standard epidemic model. Stochastic Processes and Their Applications, 11:79–89, 1981.

H.J. Wearing, P. Rohani, and M.J. Keeling. Appropriate models for the management of infectious diseases. PLOS Medicine, 2(7), 2005.

S1 SIR Model Construction and Lumpability of CTMCs

In this section, we outline why the SIR model of Section 2.2 is equivalent to the canonical SIR model (Kermack and McKendrick, 1927, Andersson and Britton, 2000) via a property called *lumpability*. The following discussion is not meant to be a comprehensive presentation of the theoretical details behind the connection between the two models. We refer readers seeking a more thorough presentation to (Tian and Kannan, 2006).

Given a Markov process, $\mathbf{X}$ with state space $\mathcal{S} = \{s_1, \ldots, s_P\}$ and initial probability vector π, we define a new process, $\overline{\mathbf{X}}$ on state space $\overline{\mathcal{S}} = \{S_1, \ldots, S_{\mathcal{L}}\}$, a partition of $\mathcal{S}$. The jump chain of this new chain is obtained by taking the sequence of subsets of $\overline{\mathcal{S}}$ that contain the corresponding states of the original jump chain. The initial probability distribution of $\overline{\mathbf{X}}(t)$ is

$$\Pr(\overline{\mathbf{X}}(t_0) = S_i) = \Pr_\pi(\mathbf{X}(t_0) \in S_i)$$

and its transition probabilities are given by

$$\Pr(\overline{\mathbf{X}}(t + \Delta t) = S_j | \overline{\mathbf{X}}(t) = \overline{\mathbf{x}}(t'), t' \leqslant t) = \Pr(\overline{\mathbf{X}}(t + \Delta t) \in S_j | \mathbf{X}(t) = \mathbf{x}(t'), t' \leqslant t),$$

where $\overline{\mathbf{x}}(t')$ and $\mathbf{x}(t')$ denote the paths of the original process and the new process. The new process is called the *lumped process*. We say that the original process is *lumpable* with respect to a partition $\overline{\mathcal{S}}$ of $\mathcal{S}$, and that $\overline{\mathbf{X}}(t)$ is the *lumped Markov process* corresponding to $\mathbf{X}(t)$, if for every choice of π we have that $\overline{\mathbf{X}}(t)$ is Markov and the transition probabilities do not depend on π. A necessary and sufficient condition for a CTMC to be lumpable is that its rate matrix, $\mathbf{\Lambda} = (\lambda_{a,b})$, where $\lambda_{a,b}$ being the rate of transition from s_a to s_b, satisfies

$$\sum_{s_b \in S_B} \lambda_{a,b} = \sum_{s_b \in S_B} \lambda_{c,b}$$

for any pair of sets S_A and S_B and for any pair of states (s_a, s_c) in $S_A \in \overline{\mathcal{S}}$.

In Section 2, we defined the latent process, $\mathbf{X}(\tau) = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$, with state space $\mathcal{S} = \{S, I, R\}^N$. Let $c_u = (x_1, \ldots, x_N)$ denote a configuration of the state labels (e.g. $c_u = (S, I, S, R, I)$), and denote the set of configurations that correspond to a vector of compartment counts by

$$\mathcal{C}_{lmn} = \left\{ c_u : l = \sum_{i=1}^{N} \mathbb{I}\left(x_i = S\right), m = \sum_{i=1}^{N} \mathbb{I}\left(x_i = I\right), n = \sum_{i=1}^{N} \mathbb{I}\left(x_i = R\right),\ l + m + n = N \right\}.$$

The state space of count vectors,

$$\overline{\mathcal{S}} = \left\{ \mathcal{C}_{lmn} : l, m, n \in \{1, \ldots, N\}, l + m + n = N \right\},$$

defines a partition of $\mathcal{S}$ that is obtained by stripping away the subject labels and summing the number of individuals in each disease state.

Given the partition $\overline{\mathcal{S}}$ of $\mathcal{S}$, we may define the CTMC for the canonical SIR model, $\overline{\mathbf{X}} = (S_\tau, I_\tau, R_\tau)$, on the state space of compartment counts. This construction is usually presented for computational

reasons since discarding the subject labels for infection and recovery events substantially reduces the computational overhead. When the sojourn times are exponentially distributed, the transition rates for the time-homogeneous CTMC are

$$\begin{array}{cc} \underline{\text{Transition}} & \underline{\text{Rate}} \\ (S, I, R) \longrightarrow (S-1, I+1, R) & \beta, \\ (S, I, R) \longrightarrow (S, I-1, R+1) & \mu I. \end{array}$$

The state space $\overline{\mathcal{S}}$ partitions the state space $\mathcal{S}$ into groups of configurations for which the triple of compartment counts are the same. The CTMC $\overline{\mathbf{X}}$ trivially satisfies the condition for lumpability, and thus is the lumped Markov chain of $\mathbf{X}$ with respect to this partition.

S2 Forward-Backward Algorithm for Sampling the Disease State at Observation Times

The stochastic forward-backward algorithm (Scott, 2002) enables us to efficiently sample from $\pi\left(\mathbf{X} \mid \mathbf{Y}, \mathbf{X}_{(-j)}, \boldsymbol{\theta}\right)$ by recursively accumulating, in a "forward" pass, information about the probability of various paths through $\mathcal{S}$, conditional on the data, and then recursively sampling a trajectory in a "backwards" pass. Let $\mathbf{Y}_{t_1}^{t_\ell} = (Y_1, \ldots, Y_\ell)$ denote the observations made at times $t_1, \ldots, t_\ell$, and similarly, let $\mathbf{X}_{j,t_{L-\ell+1}}^{t_L} = (\mathbf{X}_j(t_{L-\ell+1}), \ldots, \mathbf{X}_j(t_L))$ denote the state of $\mathbf{X}_j$ at times $t_{L-\ell+1}, \ldots, t_L$. In the forward recursion, we construct a sequence of matrices $\boldsymbol{Q}_j^{(t_2)}, \ldots, \boldsymbol{Q}_j^{(t_L)}$, where $\boldsymbol{Q}_j^{(t_\ell)} = \left(q_{j,r,s}^{(t_\ell)}\right)$, and $q_{j,r,s}^{(t_\ell)} = \Pr\left(\mathbf{X}_j(t_\ell) = s, \mathbf{X}_j(t_{\ell-1}) = r \mid \mathbf{Y}_{t_1}^{t_\ell}, \mathbf{X}_{(-j)}, \boldsymbol{\theta}\right)$. Let $\mathbf{P}_{r,s}^{(j)}(t_{\ell-1}, t_\ell) = \Pr\left(\mathbf{X}_j(t_\ell) = s \mid \mathbf{X}(t_{\ell-1}) = r, \boldsymbol{\theta}; \mathbf{X}_{(-j)}\right)$. If there are changes in the numbers of infected individuals in interval $\mathcal{I}_\ell$, we construct the transition probability matrix for that interval as in (9). Then,

$$q_{j,r,s}^{(t_\ell)} \propto \pi_j^{(t_\ell)}\left(r \mid \mathbf{X}_{(-j)}, \boldsymbol{\theta}\right) \times \mathbf{P}_{r,s}^{(j)}\left(t_{\ell-1}, t_\ell\right) \times f\left(Y_{t_\ell} \mid \mathbf{X}_j(t_\ell), \mathbf{X}_{(-j)}(t_\ell), \rho, \mathbf{p}_{t_1}\right), \tag{13}$$

where $\pi_j^{(t_\ell)}\left(r \mid \mathbf{X}_{(-j)}, \boldsymbol{\theta}, \rho\right) = \sum_r q_{j,r,s}^{(t_j)}$ and with proportionality reconciled via $\sum_r \sum_s q_{j,r,s}^{(t_j)} = 1$.

In the backwards pass, we sample the sequence of states at times $t_1, \ldots, t_L$ from the distribution $\pi\left(\mathbf{X} \mid \mathbf{Y}, \mathbf{X}_{(-j)}, \boldsymbol{\theta}, \rho, \mathbf{p}_{t_1}\right)$. To do this, we first note that

$$\begin{aligned} \pi\left(\mathbf{X} \mid \mathbf{Y}, \mathbf{X}_{(-j)}, \boldsymbol{\theta}, \rho, \mathbf{p}_{t_1}\right) &= \pi\left(\mathbf{X}_j(t_L) \mid \mathbf{Y}_{t_1}^{t_L}, \mathbf{X}_{(-j)}, \boldsymbol{\theta}, \rho, \mathbf{p}_{t_1}\right) \prod_{\ell=1}^{L-1} \pi\left(\mathbf{X}_j(t_{L-\ell}) \mid \mathbf{X}_{j,t_{L-\ell+1}}^{t_L}, \mathbf{X}_{(-j)}, \mathbf{Y}_{t_1}^{t_L}, \boldsymbol{\theta}, \rho, \mathbf{p}_{t_1}\right) \\ &= \pi\left(\mathbf{X}_j(t_L) \mid \mathbf{Y}_{t_1}^{t_L}, \mathbf{X}_{(-j)}, \boldsymbol{\theta}, \rho, \mathbf{p}_{t_1}\right) \prod_{\ell=1}^{L-1} \pi\left(\mathbf{X}_j(t_{L-\ell}) \mid \mathbf{X}_{j,t_{L-\ell+1}}, \mathbf{X}_{(-j)}, \mathbf{Y}_{t_1}^{t_{L-\ell+1}}, \boldsymbol{\theta}, \rho, \mathbf{p}_{t_1}\right), \end{aligned}$$

where the second equality follows from the conditional independence of the HMM. We proceed by first drawing $\mathbf{X}_j(t_L)$ from $\pi_j^{(t_L)}\left(\cdot \mid \mathbf{X}_{(-j)}, \boldsymbol{\theta}, \rho\right)$, and then drawing $\mathbf{X}_j(t_\ell),\ \ell = L-1, \ldots, 1$, each in turn from the categorical distribution with masses proportional to column $\mathbf{x}_j(t_{\ell+1})$ of $\boldsymbol{Q}_j^{(t_{\ell+1})}$.

S3 Simulating Endpoint Conditioned Time–Homogeneous CTMC Paths via Modified Rejection Sampling

In this section, we briefly outline the modified rejection sampling algorithm for simulating a path from an endpoint-conditioned time-homogeneous CTMC. The following discussion is not meant to be comprehensive, and we refer the reader to (Hobolth and Stone, 2009) for a more thorough discussion. We also refer the reader to the `ECctmc R` package for a fast implementation which we relied upon in implementing our data augmentation algorithm (Fintzi, 2016).

Our goal is to simulate a path for a time–homogeneous CTMC, $\mathbf{X}$, in the interval $[0, T]$, conditional on $\mathbf{X}(0) = a$ and $\mathbf{X}(T) = b$. Let $\mathbf{\Lambda}$ be the rate matrix for the process. Let Λ_a denote the a, a diagonal element of $\mathbf{\Lambda}$, and similarly let $\Lambda_{a,b}$ denote the rate given by the a, b element. Let τ denote the hitting time for the first state transition.

The probability that at least one state change occurs in the interval $[0, T]$ given that the chain begins in state a is

$$\Pr(0 \leqslant \tau \leqslant T | \mathbf{X}(0) = a) = 1 - \mathrm{e}^{-T\Lambda_a}. \tag{14}$$

The density of the first transition time given that the chain begins in state a and that at least one transition occurs in the interval $[0, T]$ is

$$f(\tau | 0 \leqslant \tau \leqslant T,\ \mathbf{X}(0) = a) = \frac{\Lambda_a \mathrm{e}^{-\tau\Lambda_a}}{1 - \mathrm{e}^{-T\Lambda_a}}. \tag{15}$$

Thus, the CDF of τ given that at least one transition occurs in $[0, T]$ and that $\mathbf{X}(0) = a$ is

$$F(\tau | 0 \leqslant \tau \leqslant T,\ \mathbf{X}(0) = a) = \int_0^\tau \frac{\Lambda_a \mathrm{e}^{-t\Lambda_a}}{1 - \mathrm{e}^{-T\Lambda_a}} dt = \frac{1 - \mathrm{e}^{-\tau\Lambda_a}}{1 - \mathrm{e}^{-T\Lambda_a}}. \tag{16}$$

We can now sample the first transition time by the inverse-CDF method, sampling $u \sim \mathrm{Unif}(0, 1)$ and applying the inverse-CDF function

$$F^{-1}(u) = \frac{-log\left[1 - u \times \left(1 - \mathrm{e}^{-T\Lambda_a}\right)\right]}{\Lambda_a}. \tag{17}$$

The modified rejection algorithm proposes paths by explicitly sampling the first transition time when it is known that at least one transition occurred (i.e. when $a \neq b$). The remainder of the path is proposed by forward sampling, for instance, via Gillespie's direct algorithm. The proposed path is then accepted if $\mathbf{X}(T) = b$. When it is not known whether a transition occurred (i.e. when $a = b$), a path is proposed via ordinary forward simulation and accepted if the proposed path is valid, i.e. if $\mathbf{X}(T) = b$.

S4 Metropolis-Hastings Ratio Details

Our target distribution is $\pi(\mathbf{X}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\mathbf{X})\pi(\mathbf{X})$. Note that $\mathbf{x}^{\text{new}}$ and $\mathbf{x}^{\text{cur}}$ differ only in the path of the j^{th} subject, so $\Lambda^{(-j)}(\mathbf{x}^{\text{cur}}) = \Lambda^{(-j)}(\mathbf{x}^{\text{new}}) = \boldsymbol{\Lambda}^{(-j)}$. Suppressing the dependence on $\boldsymbol{\theta}$ for clarity, the acceptance ratio is

$$a_{\mathbf{x}^{\text{cur}} \longrightarrow \mathbf{x}^{\text{new}}} = \min\left\{ \frac{\pi(\mathbf{x}^{\text{new}}|\mathbf{y})}{\pi(\mathbf{x}^{\text{cur}}|\mathbf{y})} \frac{q(\mathbf{x}^{\text{cur}}|\mathbf{x}^{\text{new}})}{q(\mathbf{x}^{\text{new}}|\mathbf{x}^{\text{cur}})},\ 1 \right\}$$

Now,

$$\begin{aligned} \pi(\mathbf{x}^{\text{new}}|\mathbf{y}) &\propto \Pr(\mathbf{y}|\mathbf{x}^{\text{new}})\pi(\mathbf{x}^{\text{new}}), \\ \pi(\mathbf{x}^{\text{cur}}|\mathbf{y}) &\propto \Pr(\mathbf{y}|\mathbf{x}^{\text{cur}})\pi(\mathbf{x}^{\text{cur}}), \end{aligned}$$

where $\Pr(\mathbf{y}|\mathbf{x}^{\text{new}})$ and $\Pr(\mathbf{y}|\mathbf{x}^{\text{cur}})$ are binomial probabilities for the measurement process, and $\pi(\mathbf{x}^{\text{new}})$ and $\pi(\mathbf{x}^{\text{cur}})$ are the time–homogenous CTMC densities of the current and the proposed population–level paths that appear in Equation (4). Let $\pi(\mathbf{x}_j^{\text{new}}|\boldsymbol{\Lambda}^{(-j)};\mathcal{I})$ and $\pi(\mathbf{x}_j^{\text{cur}}|\boldsymbol{\Lambda}^{(-j)};\mathcal{I})$ denote the time–inhomogeneous subject–level CTMC proposal densities given by (8). Then,

$$\begin{aligned} q(\mathbf{x}^{\text{new}}|\mathbf{x}^{\text{cur}}) &= \Pr(\mathbf{x}^{\text{new}}|\mathbf{y};\boldsymbol{\Lambda}^{(-j)}(\mathbf{x}^{\text{cur}}),\mathcal{I}) \\ &= \frac{\pi(\mathbf{x}^{\text{new}},\mathbf{y};\boldsymbol{\Lambda}^{(-j)}(\mathbf{x}^{\text{cur}}),\mathcal{I})}{\Pr(\mathbf{y};\boldsymbol{\Lambda}^{(-j)},\mathcal{I})} \\ &= \frac{\Pr(\mathbf{y}|\mathbf{x}^{\text{new}})\pi(\mathbf{x}_j^{\text{new}}|\boldsymbol{\Lambda}^{(-j)};\mathcal{I})}{\Pr(\mathbf{y};\boldsymbol{\Lambda}^{(-j)}(\mathbf{x}^{\text{new}}),\mathcal{I})} \end{aligned}$$

and similarly,

$$q(\mathbf{x}^{\text{cur}}|\mathbf{x}^{\text{new}}) = \frac{\Pr(\mathbf{y}|\mathbf{x}^{\text{cur}})\pi(\mathbf{x}_j^{\text{cur}}|\boldsymbol{\Lambda}^{(-j)};\mathcal{I})}{\Pr(\mathbf{y};\boldsymbol{\Lambda}^{(-j)}(\mathbf{x}^{\text{cur}}),\mathcal{I})}.$$

Therefore,

$$\begin{aligned} \frac{\pi(\mathbf{x}^{\text{new}}|\mathbf{y})}{\pi(\mathbf{x}^{\text{cur}}|\mathbf{y})} \frac{q(\mathbf{x}^{\text{cur}}|\mathbf{x}^{\text{new}})}{q(\mathbf{x}^{\text{new}}|\mathbf{x}^{\text{cur}})} &= \frac{\Pr(\mathbf{y}|\mathbf{x}^{\text{new}})\pi(\mathbf{x}^{\text{new}})}{\Pr(\mathbf{y}|\mathbf{x}^{\text{cur}})\pi(\mathbf{x}^{\text{cur}})} \frac{\Pr(\mathbf{y}|\mathbf{x}^{\text{cur}})\pi(\mathbf{x}_j^{\text{cur}};\boldsymbol{\Lambda}^{(-j)})}{\Pr(\mathbf{y}|\mathbf{x}^{\text{new}})\pi(\mathbf{x}_j^{\text{new}};\boldsymbol{\Lambda}^{(-j)})} \\ &= \frac{\pi(\mathbf{x}^{\text{new}})}{\pi(\mathbf{x}^{\text{cur}})} \frac{\pi(\mathbf{x}_j^{\text{cur}}|\boldsymbol{\Lambda}^{(-j)};\mathcal{I})}{\pi(\mathbf{x}_j^{\text{new}}|\boldsymbol{\Lambda}^{(-j)};\mathcal{I})}. \end{aligned}$$

Hence,

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

S5 Detailed MCMC Results for Simulated Data

Presented below are traceplots of the log-likelihood and the posterior parameter samples, along with posterior density plots and pairwise posterior density plots of model parameters, under all three prior regimes in the first simulation. Effective sample sizes are also presented below.

Prior Regime	β	μ	ρ	$\Pr(X_j(t_1) = S)$	$\Pr(X_j(t_1) = I)$
1	1521	177	148	372	4403
2	1385	965	1129	345	11350
3	1192	241	193	375	5492

Table S1: Effective sample sizes under each of the three prior regimes. Regime 1 consisted of diffuse priors for the rate parameters and the binomial sampling probability. Regime 2 assumed more informative priors for the rates. Regime 3 assumed informative priors for all model parameters. The same Dirichlet(9, 0.2, 0.5) distribution was used for $\mathbf{p}_{t_1}$ under all three prior regimes.

S6 Selecting the Number of Subject–Paths to Resample per Set of Parameter Updates

S6.1 Simulation details - Low immunity scenario

We simulated an epidemic in a population of 500 individuals under the following conditions: $R_0 = 2.5$, $\mu = 0.5$, $\rho = 0.5$, $\mathbf{p}_{t_1} = (0.94, 0.01, 0.05)$. For each run of 100,000 iterations, we initialized the MCMC using the same initial configuration of subject-level paths and parameter values. All that was varied from one MCMC run to the next was the number of subject-level paths per set of parameter updates. We resampled the following sequence of numbers of subjects to resample: 1, 2, 3, 4, 5, 10, 25, 50, 100, ..., 550, 600. We used informative priors for all model parameters (presented in the table below). Effective sample sizes and the Raftery-Lewis convergence diagnostic were calculated using the `coda` package (Plummer et al., 2006).

Simulation details and results - moderate immunity setting

We repeated the simulation detailed above with the only difference being that the proportion of already immune individuals was 35% of the total population. Accordingly, we also used a $Dirichlet(65, 1, 30)$ prior for the initial state probability parameters.

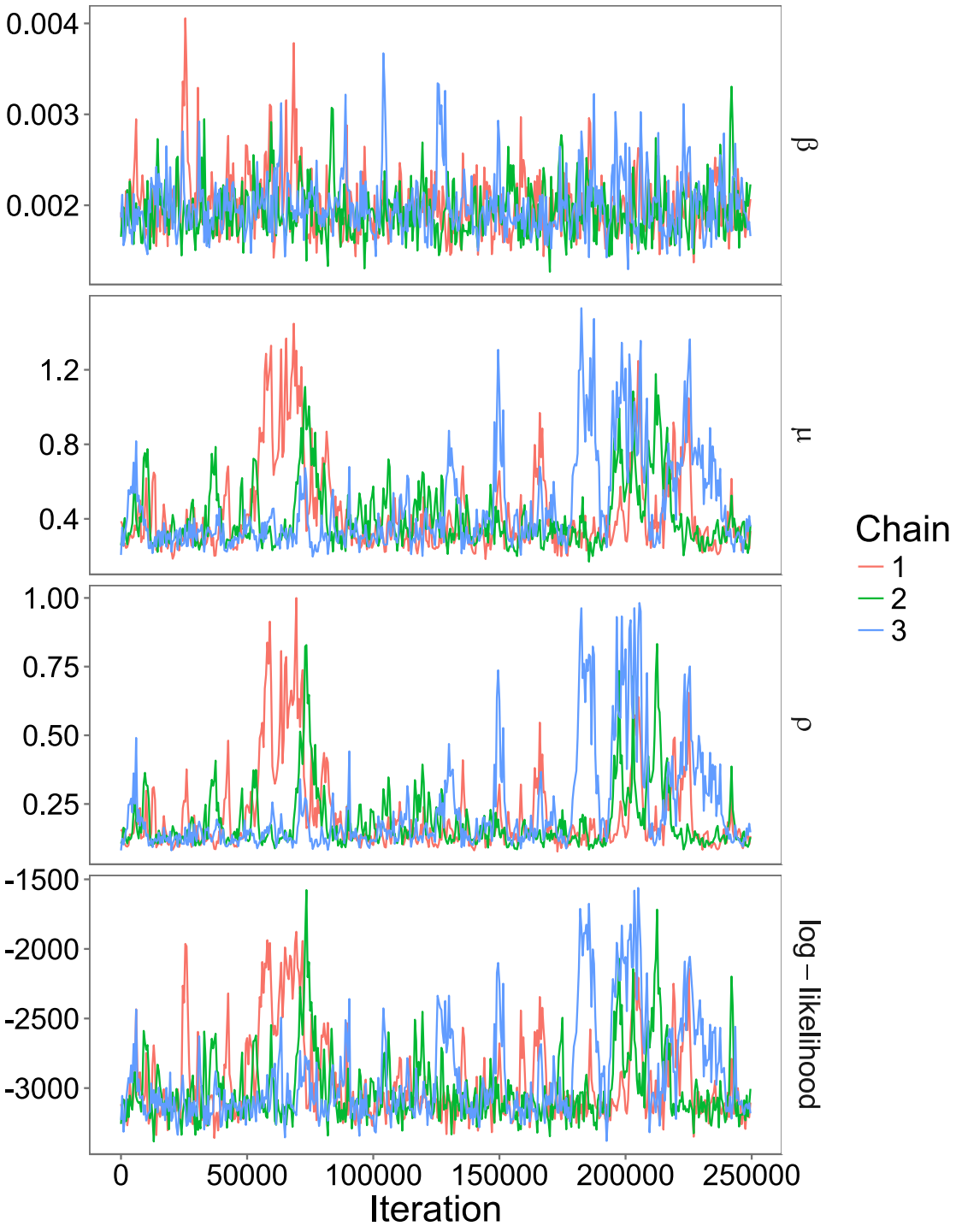

Figure S1: Log-likelihood and parameter traceplots under diffuse priors for the rate parameters and the binomial sampling probability (Regime 1), thinned to show every 500^{th} sample.

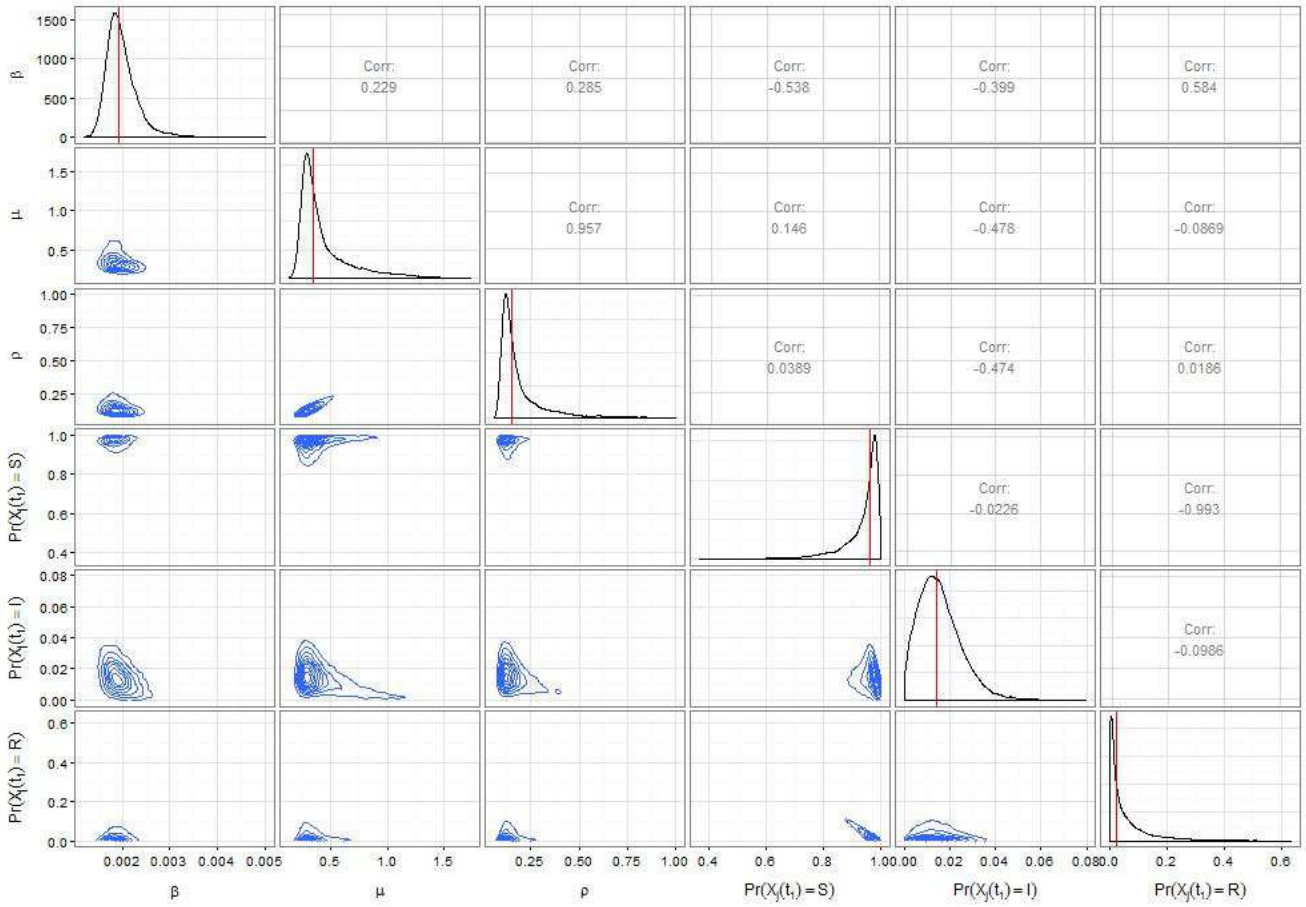

Figure S2: Posterior densities and contour plots of pairwise posterior samples of SIR model parameters under diffuse priors for the rate parameters and for the binomial sampling probability (Regime 1). Vertical red lines in posterior density plots indicate posterior median estimates.

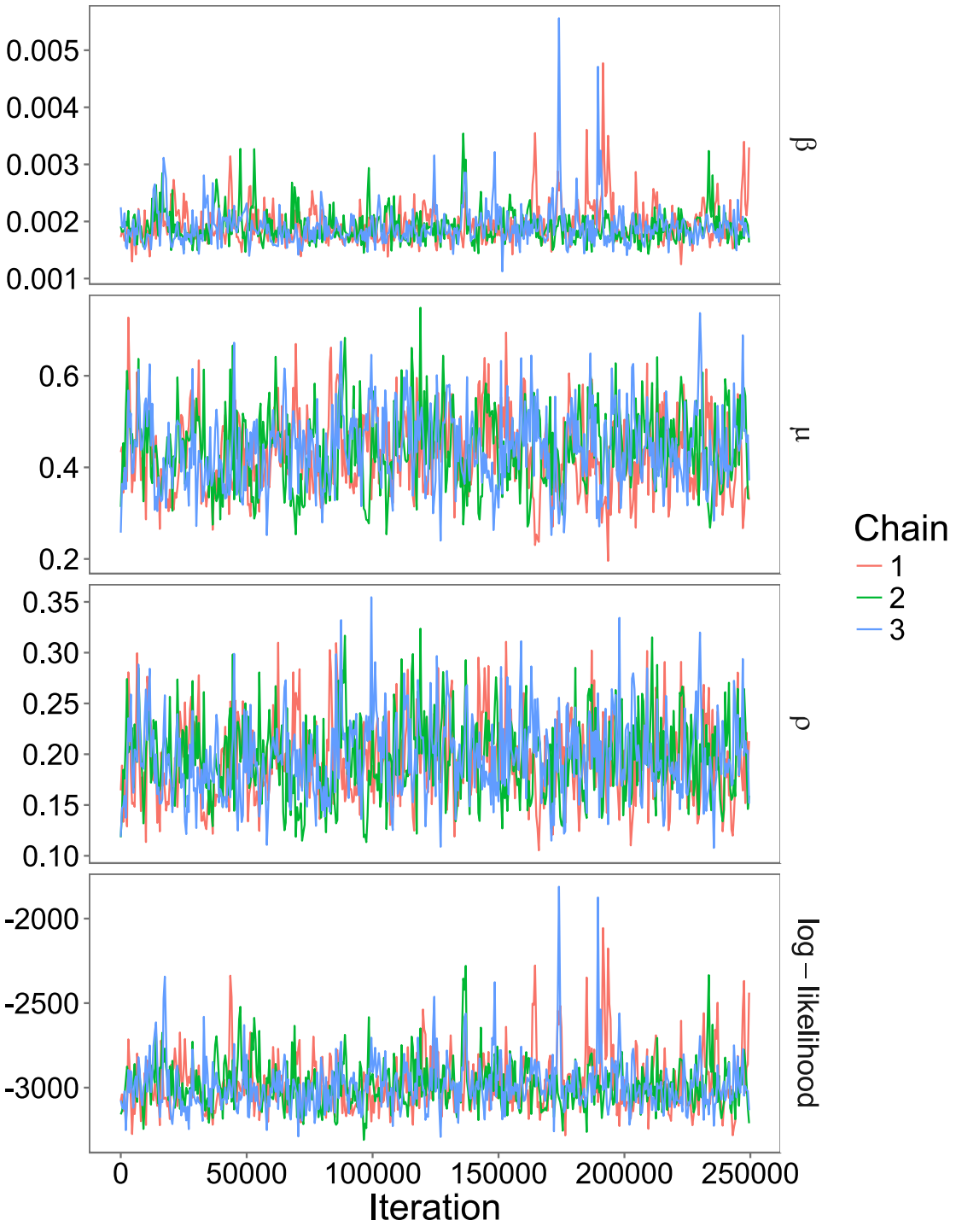

Figure S3: Log-likelihood and parameter traceplots under informative priors for the rate parameters and a diffuse prior for the binomial sampling probability (Regime 2), thinned to show every 500^{th} sample.

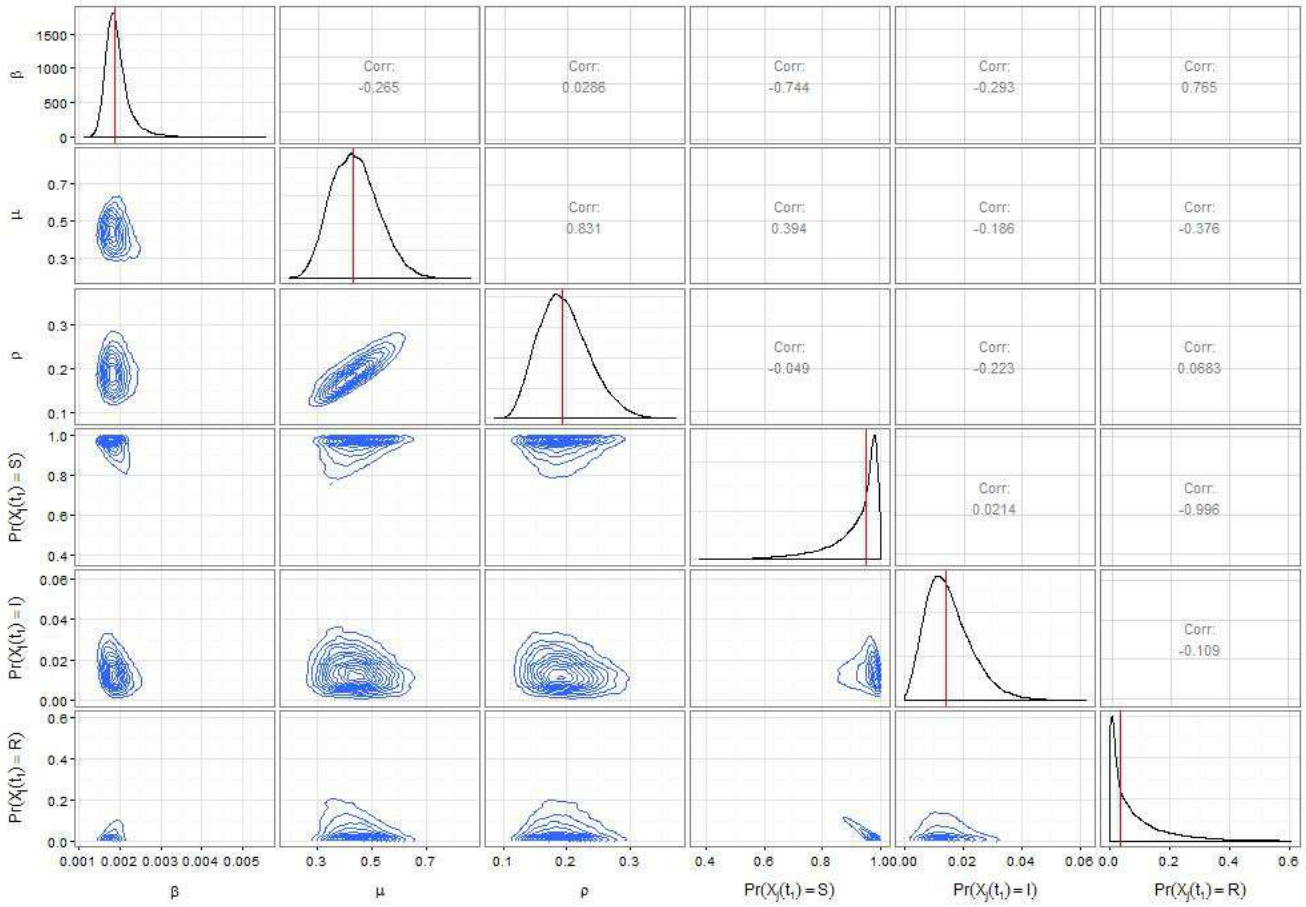

Figure S4: Posterior densities and contour plots of pairwise posterior samples of SIR model parameters under informative priors for rate parameters and a diffuse prior for the binomial sampling probability (Regime 2). Vertical red lines in posterior density plots indicate posterior median estimates.

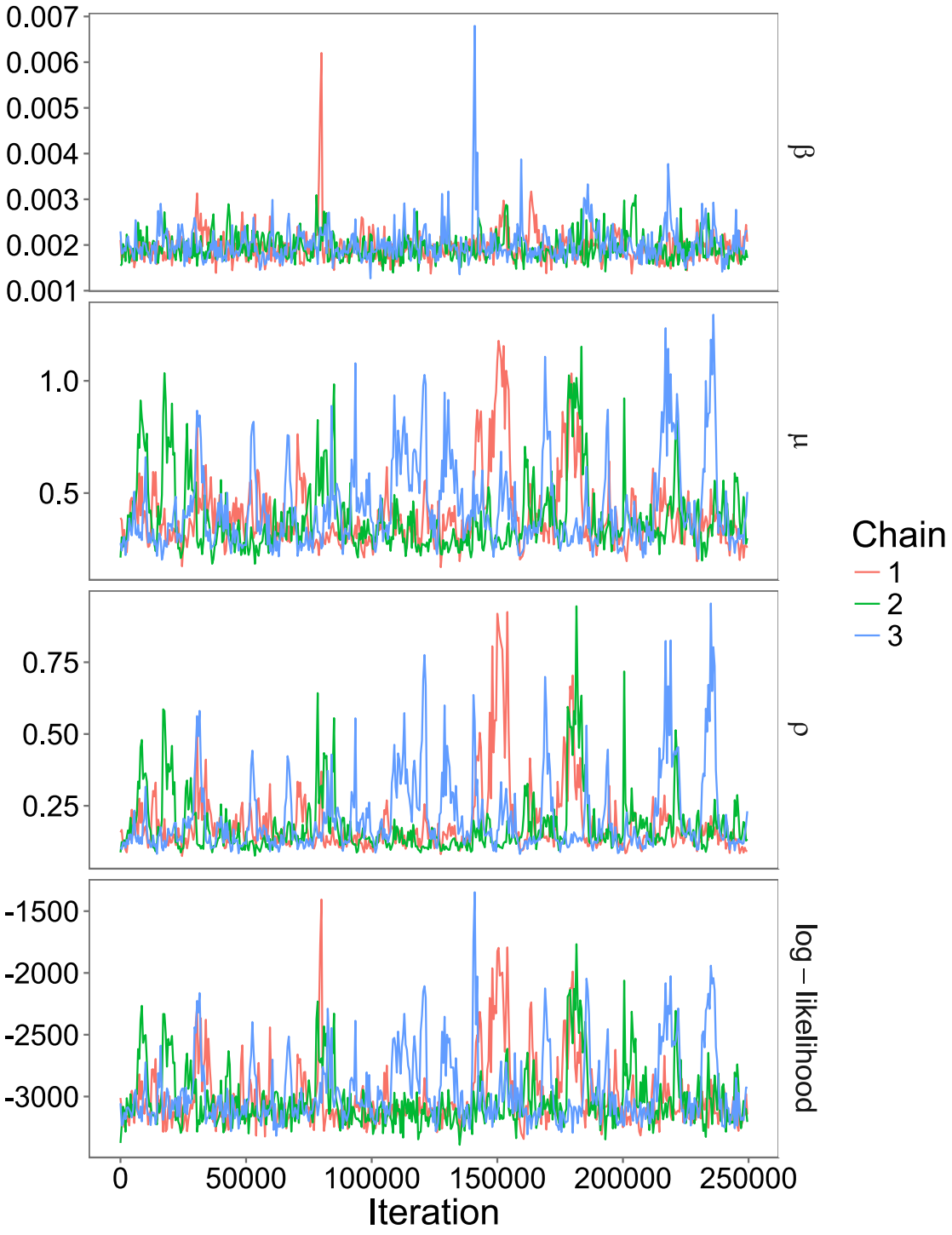

Figure S5: Log-likelihood and parameter traceplots under informative priors for the rate parameters and the binomial sampling probability (Regime 3), thinned to show every 500^{th} sample.

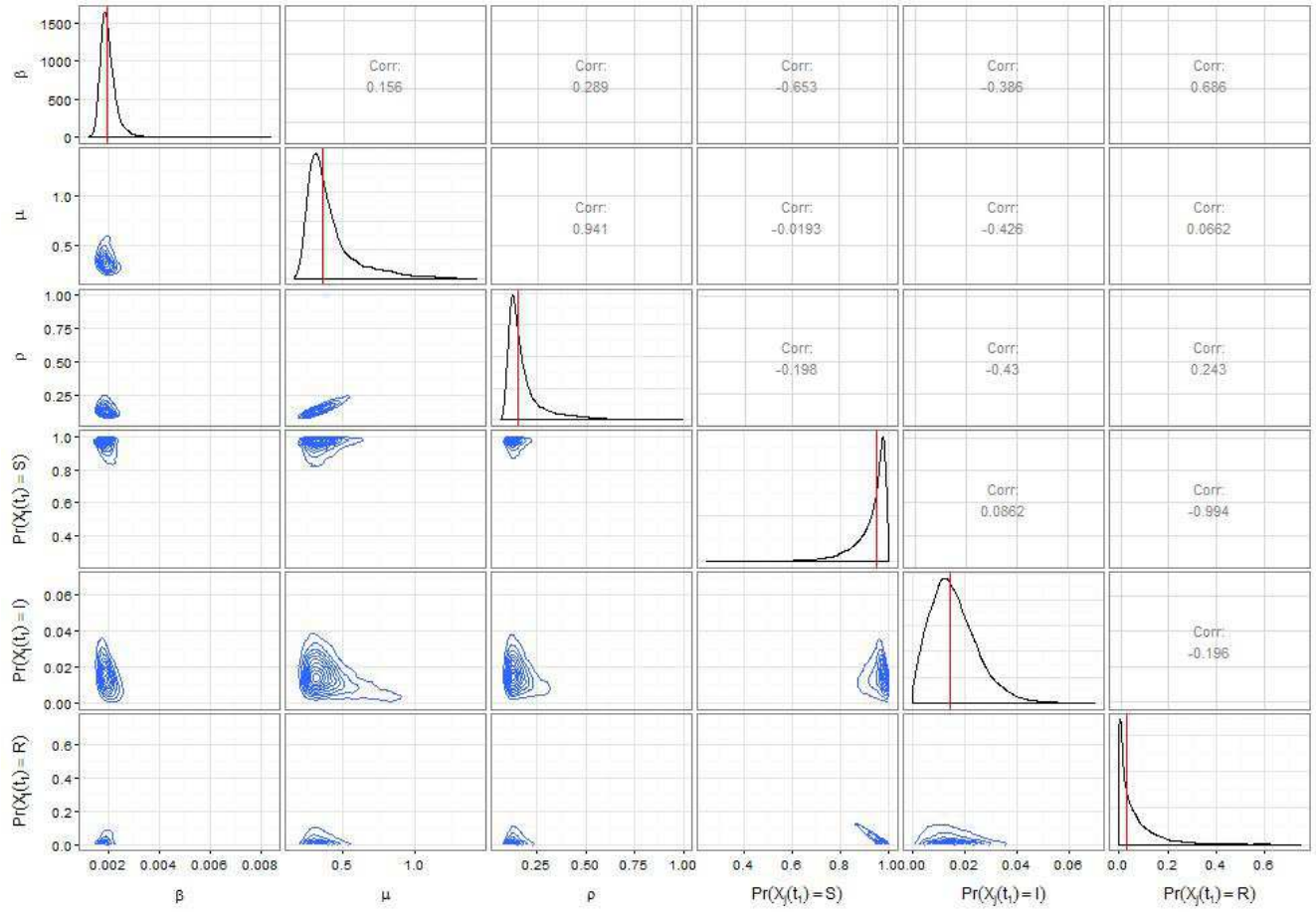

Figure S6: Posterior densities and contour plots of pairwise posterior samples of SIR model parameters under informative priors for the rate parameters and for the binomial sampling probability (Regime 3). Vertical red lines in posterior density plots indicate posterior median estimates.

Parameter	Prior Distribution
R_0	$\frac{1}{2}\text{Beta}'(0.0024,\ 0.96)$
β	$Gamma(0.0024, 1)$
μ	$Gamma(0.96, 2)$
ρ	$Beta(3.75, 4.25)$
$p_{S_{t_1}}$ $p_{I_{t_1}}$ $p_{R_{t_1}}$	$Dirichlet(95, 1, 5)$

Table S2: Prior distributions used for model parameters in determining the optimal number of subject paths per parameter update. The prior for R_0 is the implied prior distribution induced by the prior choices for β and μ.

S7 MCMC output and convergence diagnostics for British boarding school example

Parameter	Prior Distribution	Posterior Median (95% CrI)
R_0	$\frac{1}{2}\text{Beta}'(0.001,\ 1)$	3.89 (3.41, 4.48)
β	$Gamma(0.001, 1)$	0.0024 (0.0021, 0.0026)
μ	$Gamma(1, 2)$	0.46 (0.42, 0.50)
ρ	$Beta(1, 2)$	0.98 (0.92, 0.99)
$p_{S_{t_1}}$	$Dirichlet(900, 3, 9)$	0.99 (0.98, 0.99)
$p_{I_{t_1}}$		0.003 (0.001, 0.007)
$p_{R_{t_1}}$		0.009 (0.004, 0.017)

Table S3: Prior distributions and posterior estimates for boarding school SIR model parameters from three chains with 75 subject path updates per parameter update. The prior for R_0 is the implied prior induced by the priors for β and μ. Effective sample size were β: 20,958; μ: 37,266; ρ: 9,336; $p_{S_{t_1}}$: 49,605; $p_{I_{t_1}}$: 715,284; $p_{R_{t_1}}$: 41,806.

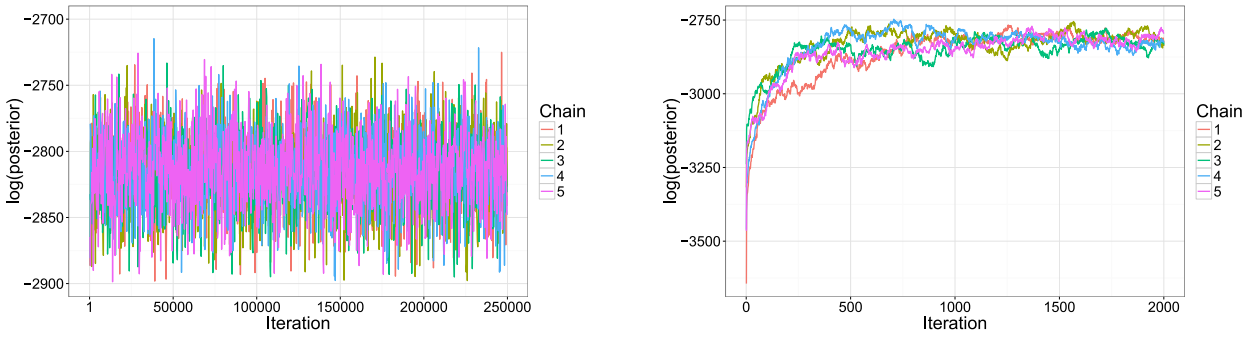

Figure S7: Log-posterior traceplots for five chains with 10 subjects subject path updates per parameter set update. (a) Log-posterior traceplots, values thinned by plotting every 250^{th} value. (b) First 2,000 log-posterior values.

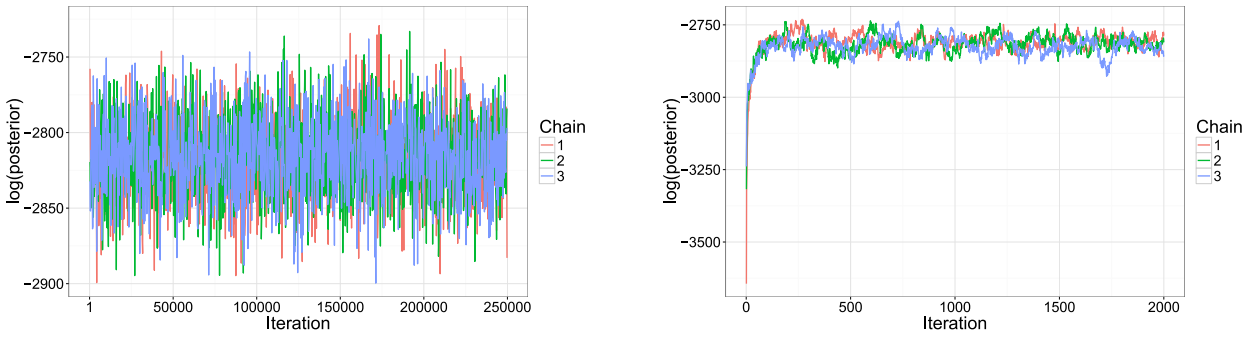

Figure S8: Log-posterior traceplots for five chains with 75 subjects subject path updates per parameter set update. (a) Log-posterior traceplots, values thinned by plotting every 250^{th} value. (b) First 2,000 log-posterior values.

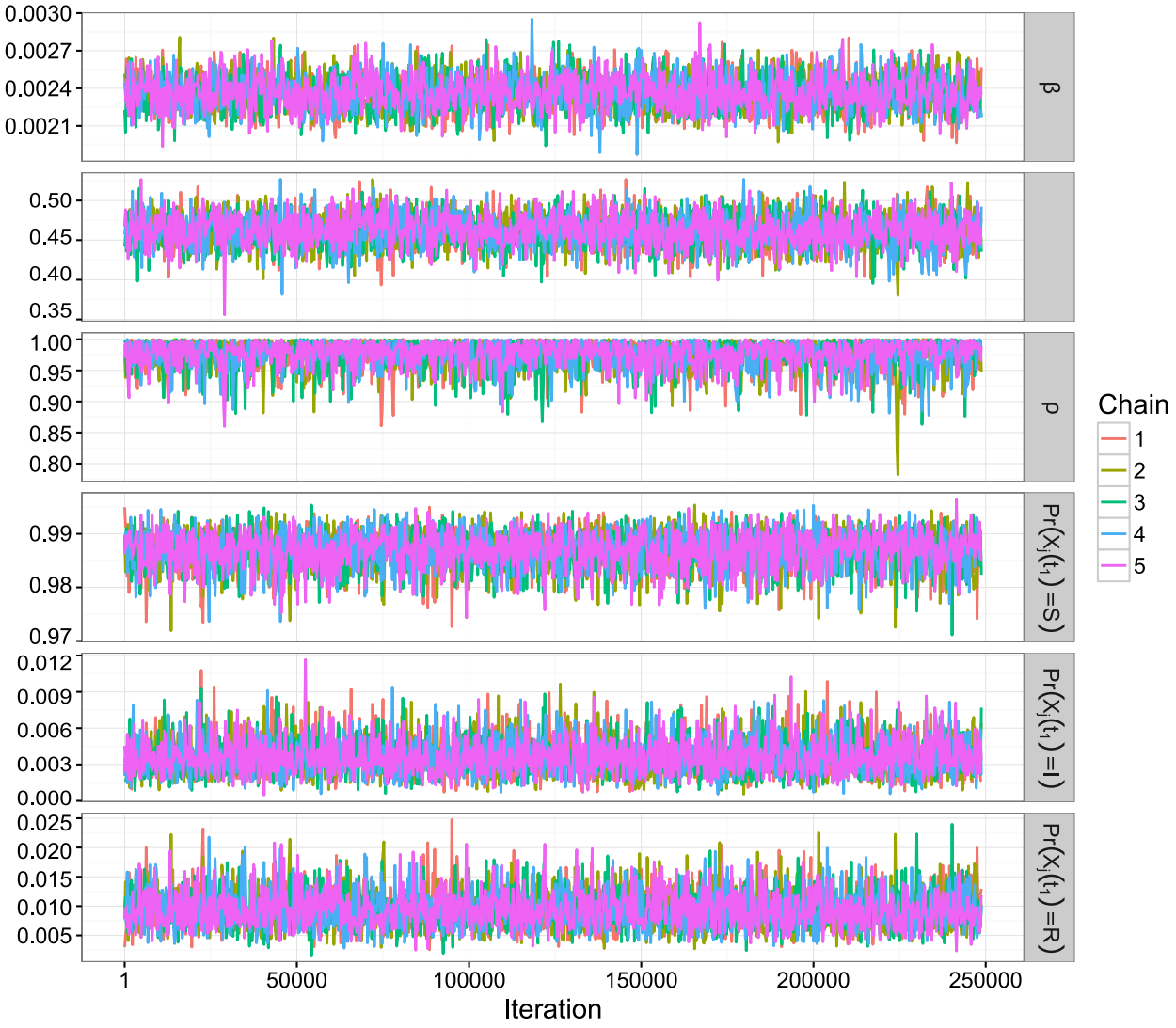

Figure S9: Traceplots of SIR model parameters for boarding school data. Estimates based on three chains sampling 10 subject paths per set of model parameter updates.

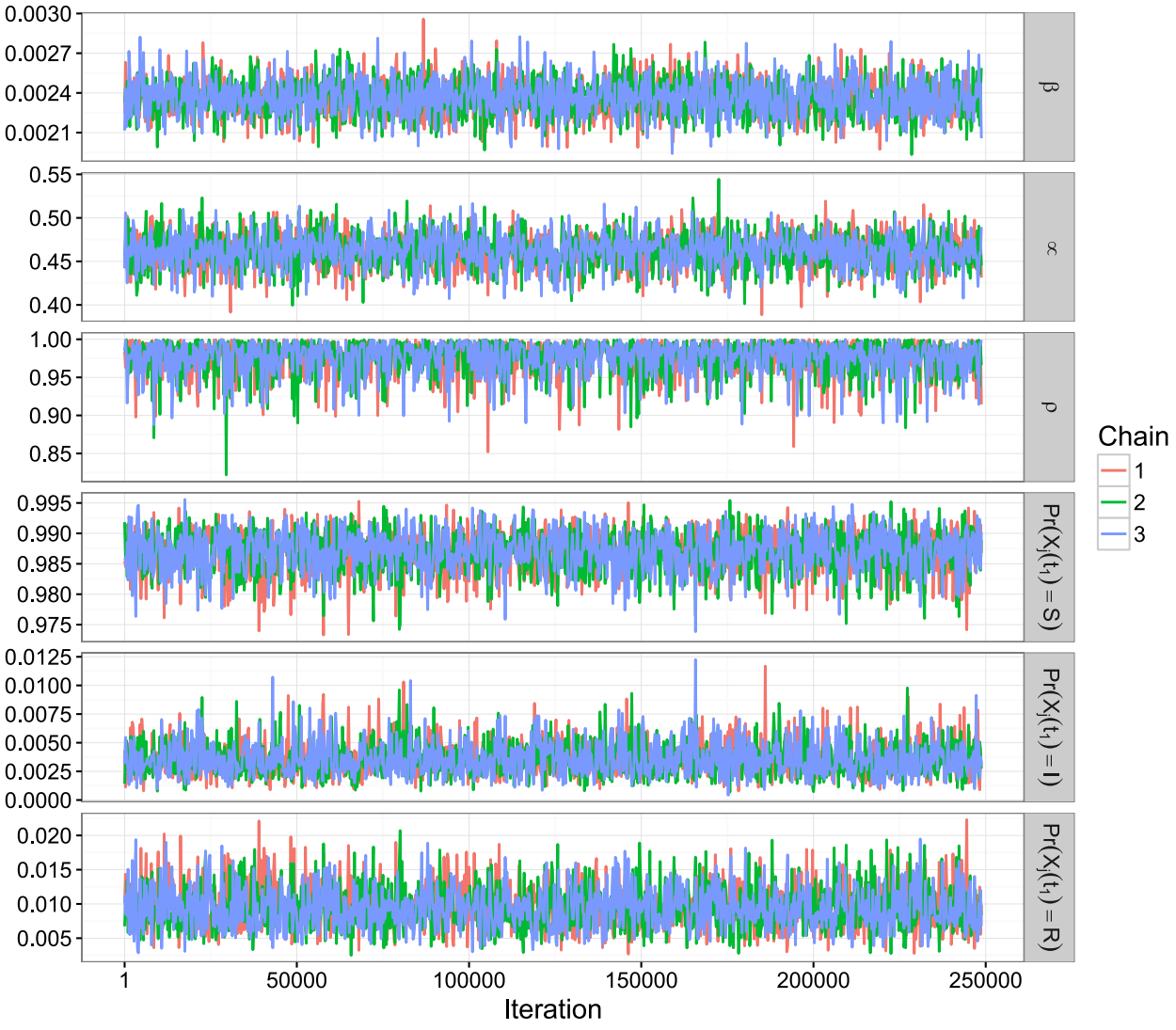

Figure S10: Traceplots of SIR model parameters for boarding school data. Estimates based on three chains sampling 75 subject paths per set of model parameter updates.

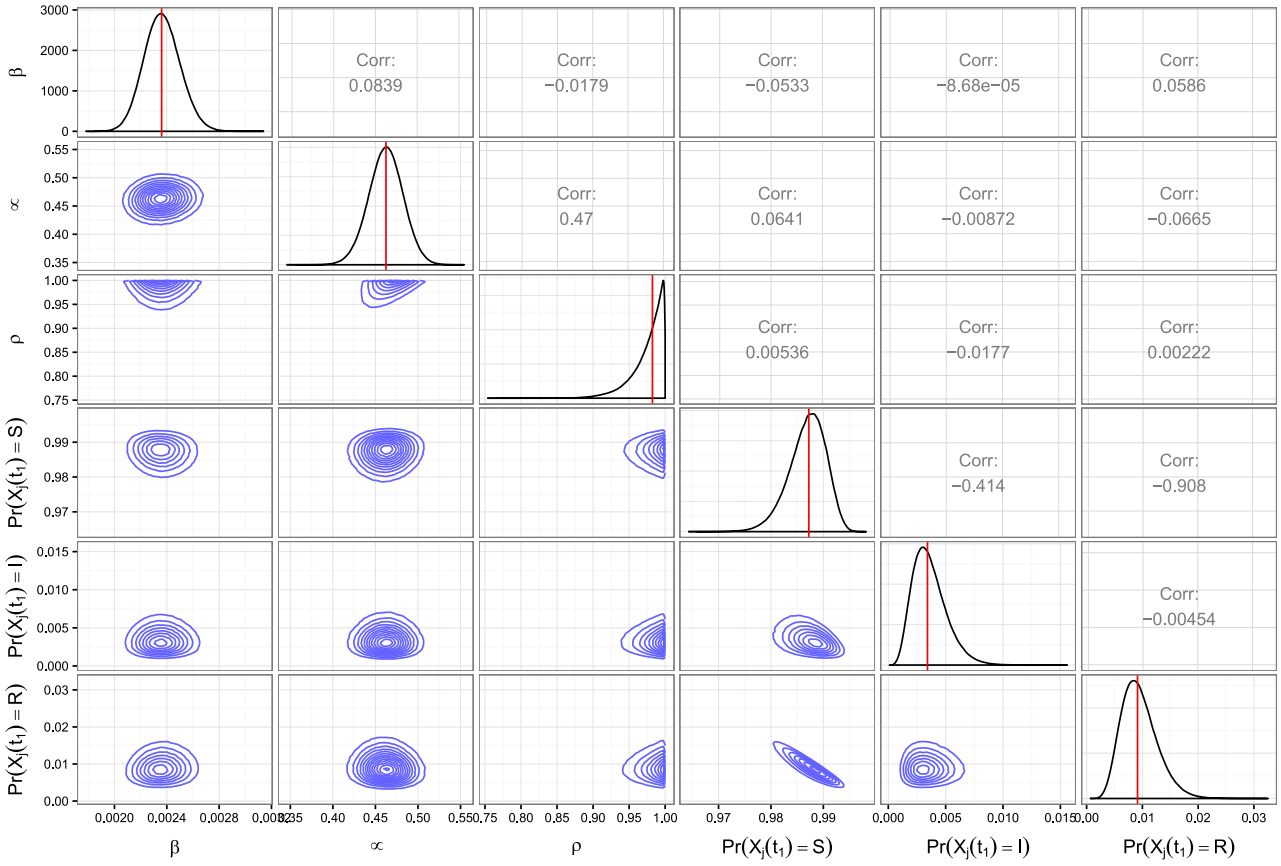

Figure S11: Posterior densities contour plots of pairwise posterior samples of SIR model parameters for boarding school data. Vertical red lines in posterior density plots indicate posterior median estimates. Estimates based on three chains sampling 10 subject paths per set of model parameter updates.

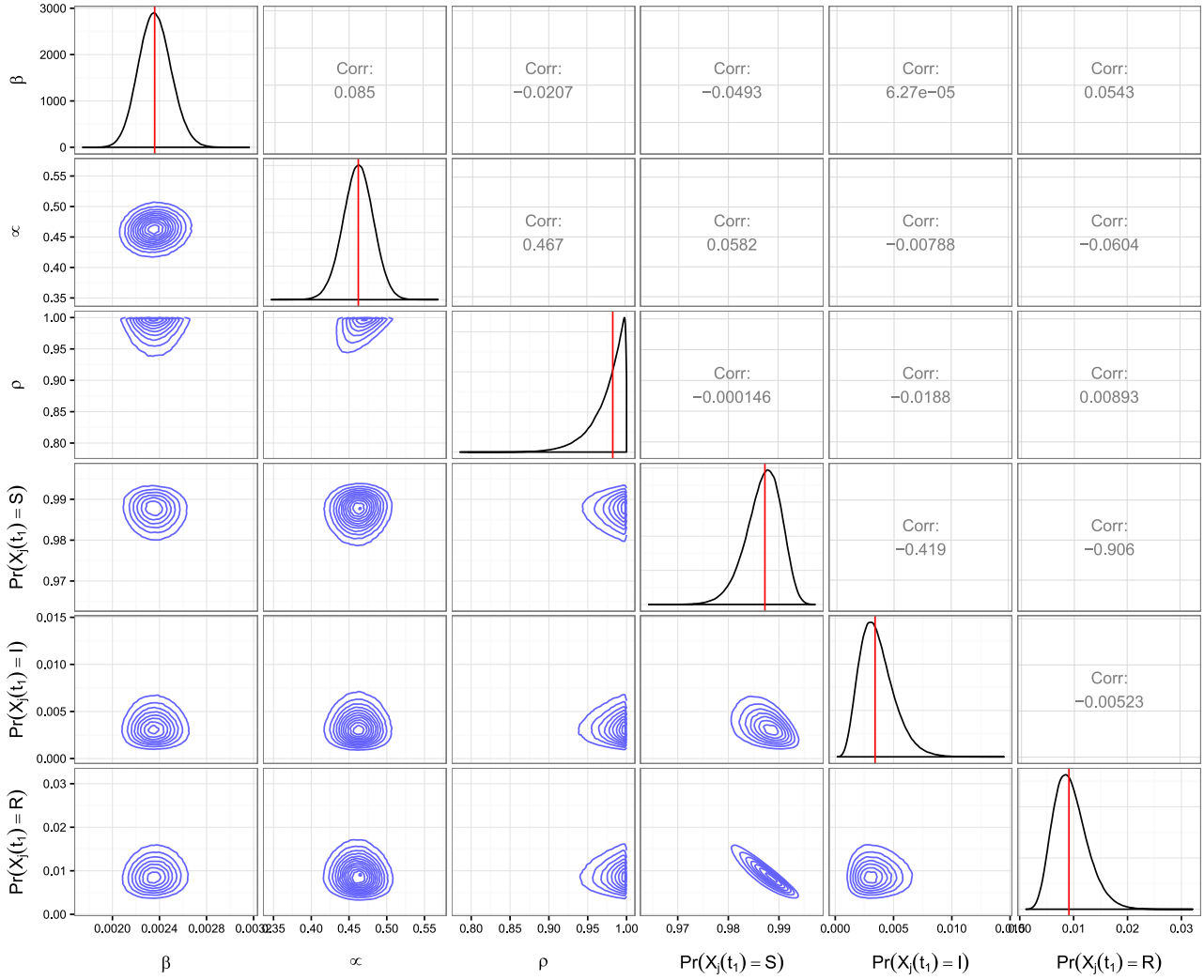

Figure S12: Posterior densities contour plots of pairwise posterior samples of SIR model parameters for boarding school data. Vertical red lines in posterior density plots indicate posterior median estimates. Estimates based on three chains sampling 75 subject paths per set of model parameter updates.