Efficient Data Augmentation for Fitting Stochastic Epidemic Models to Prevalence Data

Jonathan Fintzi[1], Xiang Cui[2], Jon Wakefield[1,2], and Vladimir N. 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 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–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 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.

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 the conclusions we draw about its dynamics (Keeling and Rohani, 2008, Allen, 2008). Moreover, many questions — e.g., 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 a 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, e.g., 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 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 within the SEM, e.g., 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 models, and the resulting estimates are specific to the SEM dynamics. Approximation methods replace the epidemic model with a simpler model whose likelihood is more 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 and to construct a transition model for the population flow between model compartments at successive times (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 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 events span several orders of magnitude (Glass et al., 2003, Shelton and Ciardo, 2014). 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, Toni et al., 2009), pseudo–marginal methods (McKinley et al., 2014), and sequential Monte Carlo (or particle filter) methods (Toni et al., 2009, Andrieu et al., 2010, Ionides et al., 2011, Dukic et al., 2012, Koepke et al., 2016). Within this class of methods, the particle marginal Metropolis–Hastings algorithm of Andrieu et al. (2010) stands out in being a general method for Bayesian inference and is used as a benchmark method in this paper. 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 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 this approach is in 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, 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), Qin and Shelton (2015), and Shestopaloff and Neal (2016) have presented methods that do not rely on subject–level data. However, their algorithms forgo the flexibility of agent–based DA, and in the case of the latter two papers 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 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 in the absence of any subject–level information. 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 outline in Section S6 the minimal adaptations required for fitting Susceptible–Exposed–Infected–Recovered (SEIR) and Susceptible–Infected–Recovered–Susceptible (SIRS) models, which we describe in Sections 3.1, 3.2, and 4.

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). For simplicity, we assume a closed, homogeneously mixing population in which each individual becomes infectious immediately upon becoming infected. 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 constant 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 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

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

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, 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 $\boldsymbol{\tau} = \{\tau_0, \ldots, \tau_{K+1}\}$, 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 $\mathbb{I}\left(\tau_k \triangleq I\right)$ and $\mathbb{I}\left(\tau_k \triangleq R\right)$ 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

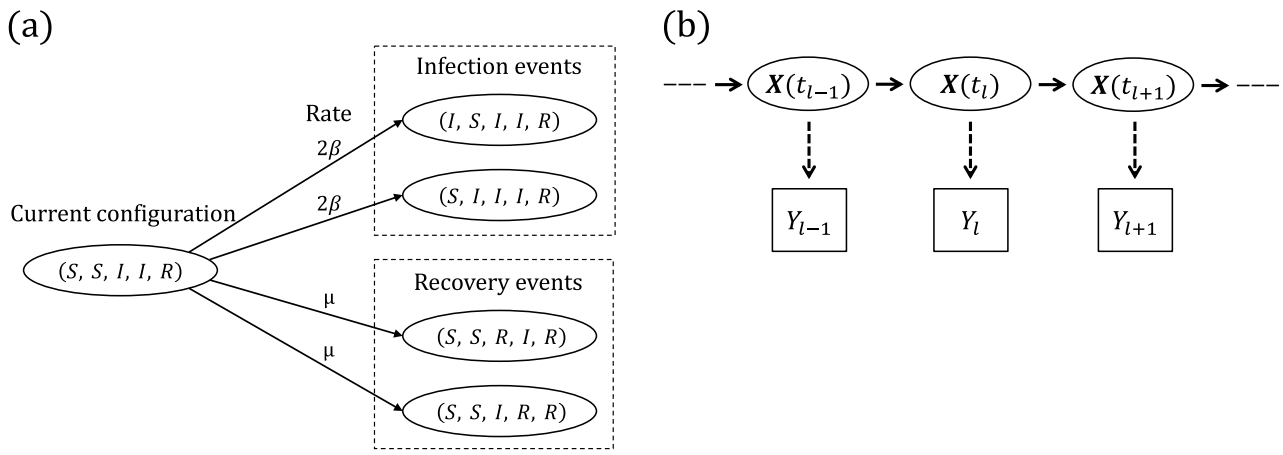


Figure 1: (a) SIR dynamics in a population of five subjects. The number of infecteds can increase from two to three via an infection of the first or second subject, reaching each of those configurations at rate 2β. The number of recovered individuals can increase from one to two via a recovery of the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

parameters. The complete data likelihood is

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

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 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 in the posterior $\pi(\boldsymbol{\theta}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \int L(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})\pi(\mathbf{X}|\boldsymbol{\theta})\pi(\theta)\mathrm{d}\pi(\mathbf{X})$ is analytically and numerically 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 data augmentation 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 in (4). The joint posterior distribution is

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

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 the path of a single subject $\mathbf{X}_j$, conditionally on the data, using a time–inhomogeneous 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 $\boldsymbol{\tau}^{(-j)} = \{t_1, t_L\} \cup \{\boldsymbol{\tau} \backslash \boldsymbol{\tau}^{(j)}\} = \left\{\tau_0^{(-j)}, \tau_1^{(-j)}, \ldots, \tau_M^{(-j)}, \tau_{M+1}^{(-j)}\right\}$, where $t_1 \equiv \tau_0^{(-j)}$ and $t_L \equiv \tau_{M+1}^{(-j)}$, to be the set of $M \leqslant K$ (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 $[t_1, t_L]$, i.e. $\mathcal{I}_1 = \left[\tau_0^{(-j)}, \tau_1^{(-j)}\right)$, $\mathcal{I}_2 = \left[\tau_1^{(-j)}, \tau_2^{(-j)}\right), \ldots,$ $\mathcal{I}_{M+1} = \left[\tau_M^{(-j)}, \tau_{M+1}^{(-j)}\right)$. 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. Let $\boldsymbol{\Lambda}^{(-j)} = \left\{\boldsymbol{\Lambda}_1^{(-j)}(\boldsymbol{\theta}), \ldots, \boldsymbol{\Lambda}_{M+1}^{(-j)}(\boldsymbol{\theta})\right\}$ be the sequence of rate matrices corresponding to each interval in $\mathcal{I}$, where for $m = 1, \ldots, M+1$. The rate matrix for subject j is

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

We can construct the transition probability matrix for interval m as

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

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, which may be carried out

efficiently by computing the decomposition analytically. We may further lessen the total computational burden by caching the eigen decompositions to avoid duplicate computations. One additional point to note is the eigen–values of any SIR rate matrix are always real. However, this is not generally true, e.g., it is possible for the rate matrix of an SIRS model to have complex eigenvalues. In this case, we obtain a real valued transition probability matrix by first applying a rotation to each rate matrix with complex eigenvalues in order to obtain its real canonical form (Hirsch et al., 2013). This is discussed in Section S2.

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

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

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 sub–intervals in $\mathcal{I}_\ell$, within which the subject–level CTMC is time–homogeneous. Thus, the transition probability matrix over an inter–observation interval, $\mathcal{I}_\ell = [t_{\ell-1}, t_\ell]$, partitioned by S transition events that define inter–event intervals with endpoints given by times $t_{\ell-1} = \tau_{\ell,0}^{(-j)} < \tau_{\ell,1}^{(-j)} < \cdots < \tau_{\ell,S-1}^{(-j)} < \tau_{\ell,S}^{(-j)} \equiv t_\ell$, 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–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.

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

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

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

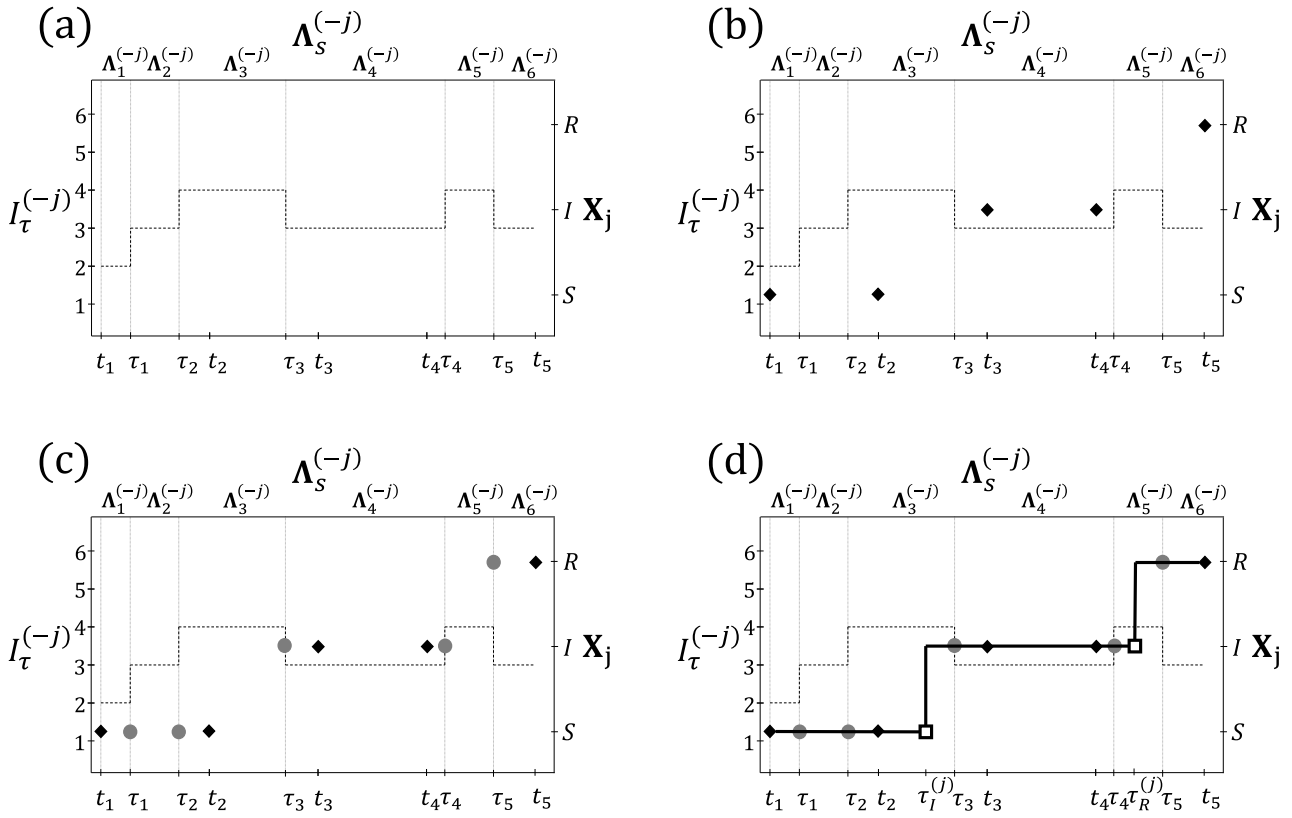


Figure 2: Procedure for constructing a subject–path proposal with SIR dynamics. (a) The dashed line depicts the number of infected individuals, excluding $\mathbf{X}_j$, the subject whose path is being sampled. The observation times, $t_1, \ldots, t_5$, and times at which other subjects change disease states, $\tau_1, \ldots, \tau_5$, are shown on the bottom axis. Rate matrices of the time–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) *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–homogeneous CTMC distributions, conditional on the sequence of disease states sampled in the HMM and discrete time skeleton steps.

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 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 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 $t_{\ell-1} = \tau_{\ell,0}^{(-j)} < \tau_{\ell,1}^{(-j)} < \cdots < \tau_{\ell,n-1}^{(-j)} < \tau_{\ell,n}^{(-j)} = t_\ell$, 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

$$
\begin{aligned}
&\Pr\left(\mathbf{X}_j(\tau_{l,i}^{(-j)}) = x_{l,i} | \mathbf{X}_j(\tau_{i-1}^{(-j)}) = x_{i-1}, \mathbf{X}_j(\tau_n^{(-j)}) = x_n\right) \\
&\qquad = \frac{\Pr\left(\mathbf{X}_j(\tau_{l,i}^{(-j)}) = x_{l,i}, \mathbf{X}_j(\tau_{l,i-1}^{(-j)}) = x_{l,i-1}, \mathbf{X}_j(\tau_n^{(-j)}) = x_n\right)}{\Pr\left(\mathbf{X}_j(\tau_{l,i-1}^{(-j)}) = x_{l,i-1}, \mathbf{X}_j(\tau_n^{(-j)}) = x_n\right)} \\
&\qquad = \frac{\Pr\left(\mathbf{X}_j(\tau_{l,i}^{(-j)}) = x_{l,i} | \mathbf{X}_j(\tau_{l,i-1}^{(-j)}) = x_{l,i-1}\right) \Pr\left(\mathbf{X}_j(\tau_n^{(-j)}) = x_n | \mathbf{X}_j(\tau_{l,i}^{(-j)}) = x_{l,i}\right)}{\Pr\left(\mathbf{X}_j(\tau_n^{(-j)}) = x_n | \mathbf{X}_j(\tau_{l,i-1}^{(-j)}) = x_{l,i-1}\right)} \\
&\qquad = \frac{\left[\mathbf{P}^{(j)}(\tau_{l,i-1}^{(-j)}, \tau_{l,i}^{(-j)})\right]_{x_{l,i-1}, x_{l,i}} \left[\prod_{k=i}^{n-1} \mathbf{P}^{(j)}(\tau_k^{(-j)}, \tau_{k+1}^{(-j)})\right]_{x_{l,i}, x_n}}{\left[\prod_{k=i-1}^{n-1} \mathbf{P}^{(j)}(\tau_k^{(-j)}, \tau_{k+1}^{(-j)})\right]_{x_{l,i-1}, x_n}}.
\end{aligned} \tag{9}
$$

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). 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.

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

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

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 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 non–zero 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. 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 Section S8).

2.5 Implementation

We provide the `R` and `C++` code base for this paper, along with examples and the code for reproducing the results we present 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 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 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 Section S9). Priors for the rate parameters and binomial sampling probability were scaled 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 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" 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 visa–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 inter–event 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 Section S7 for further discussion on this point). Simulation details, along with minor adaptations to our algorithm for fitting the SEIR and SIRS models, are presented in Section S6.

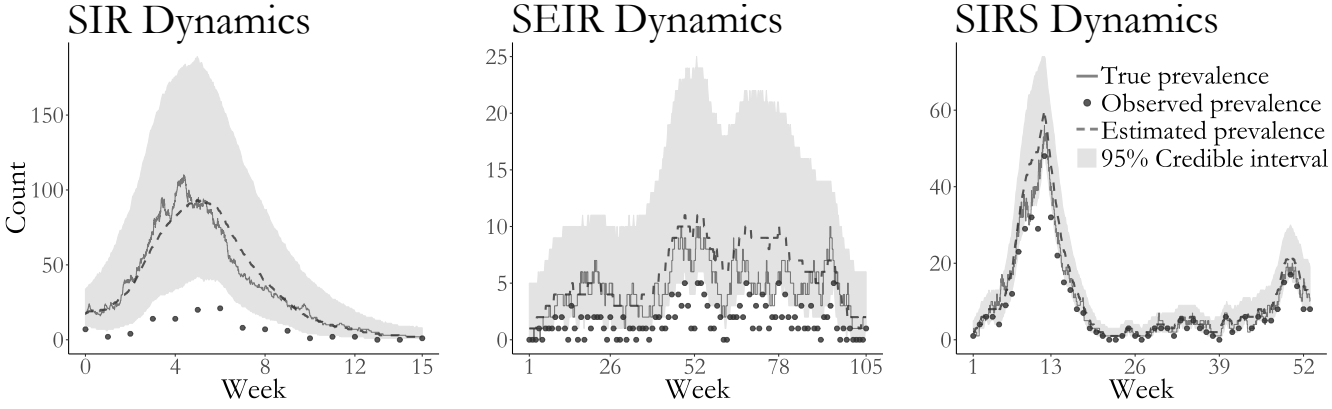


Figure 3: Estimated latent posterior distributions of disease prevalence in outbreaks simulated under SIR (left), SEIR (middle), and SIRS (right) dynamics. Depicted are the true unobserved prevalence (solid line), observed data (dots), pointwise posterior median prevalence (dashed line), and pointwise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

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; Figure S12 presents estimated latent posterior distributions and true epidemic paths for all model compartments). The acceptance rates for subject–path proposals were roughly 92% for the SIR model, 91% for the SEIR model, and 77% for the SIRS model. Our posterior estimates of the model parameters also closely match estimates obtained using the particle marginal Metropolis–Hastings (PMMH) algorithm of Andrieu et al. (2010), implemented using the `pomp` package in `R` (King et al., 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 $350\times$ greater for the SIR model, $4.4\times$ greater for the SEIR model, and $13\times$ 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 $10.5\times$ greater for PMMH in fitting the SIR model, $2\times$ for the SEIR model, and $0.7\times$ 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 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 imposible 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 interpretted counterfactually — e.g. "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 MCMC converges to some reasonable estimate of the model dynamics, 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 struggle in many instances, and it is here that we highlight a critical advantage of our data augmentation 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

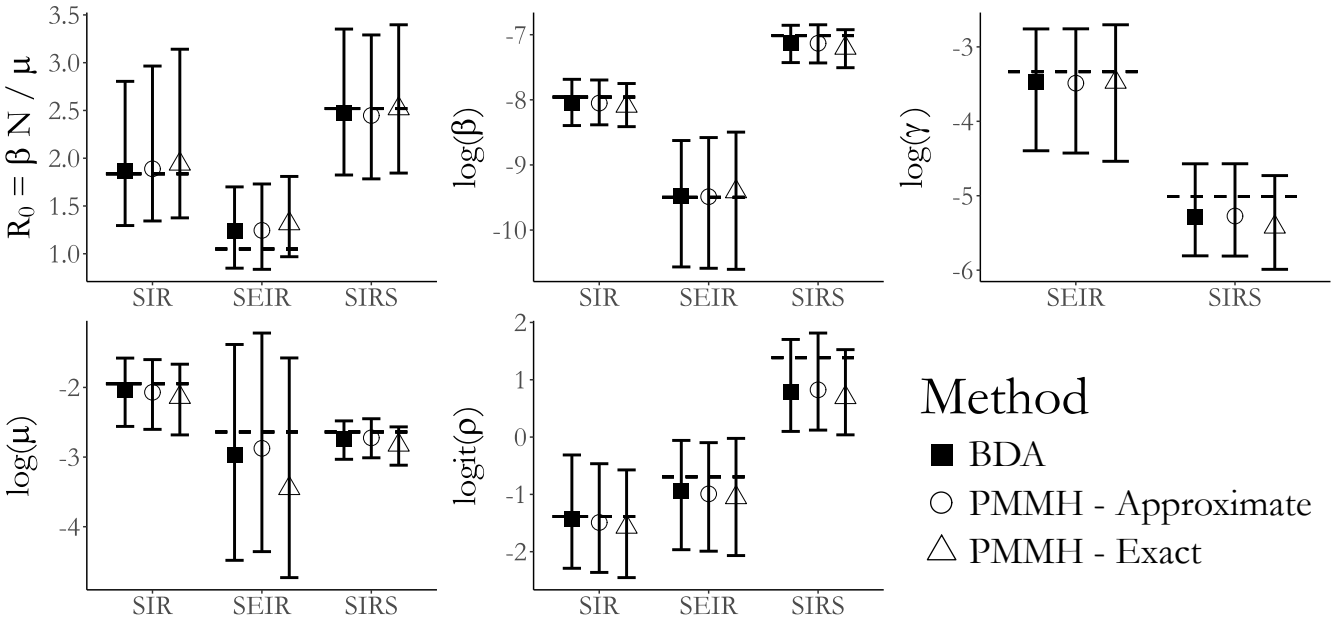
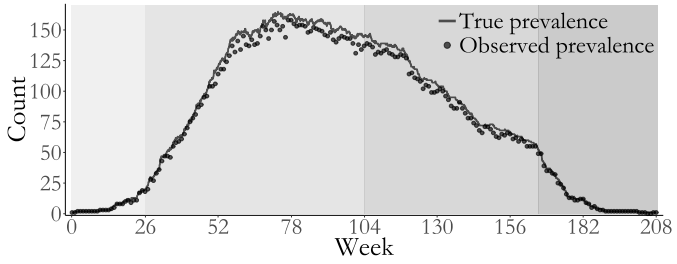


Figure 4: Posterior medians and 95% credible intervals of parameters in the SIR, SEIR, and SIRS models fit with Bayesian data augmentation (BDA) and particle marginal Metropolis–Hastings (PMMH) with particle paths simulated approximately (using τ–leaping) and exactly (using Gillespie's direct algorithm). 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.

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 four 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 Section S10).



	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 & Table 1: 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, $1/\mu$.

We fit SIR and SEIR models to the data using our DA algorithm, and using PMMH with 2,500

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 (Table S11). The MCMC chains for models fit with PMMH suffered from severe particle degeneracy and did not converge (see 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 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 outsize the 95% credible interval at times, we are still able to reconstruct a reasonabe 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 (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.

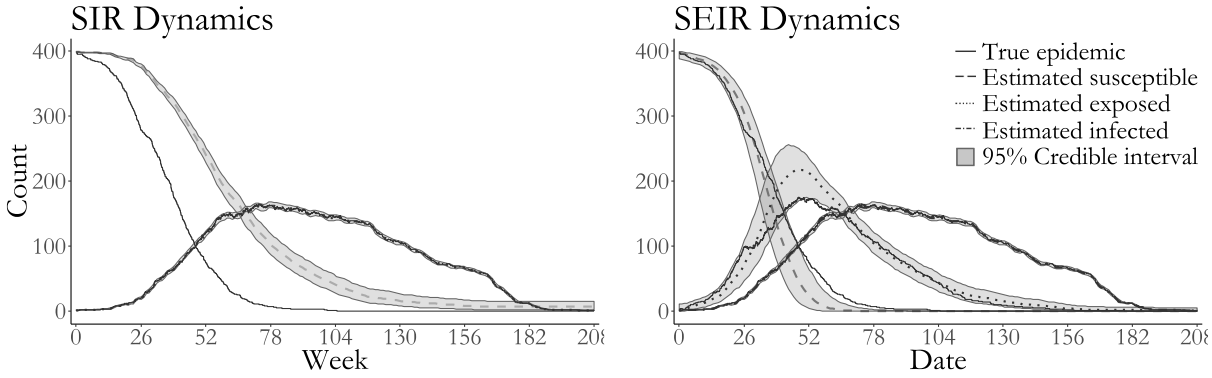


Figure 6: True epidemic path (solid lines), pointwise posterior median estimate of the numbers of susceptibles (dashed line), exposed (dotted line), and infected individuals (dash–dotted line) and pointwise 95% credible intervals (shaded regions) under SIR and SEIR dynamics.

3.3 Inference under population size misspecification

Model misspecficiation often extends not only to the SEM dynamics, but also to the assumed population size. This is most often the case in settings where subject–level data is unvailable, e.g.

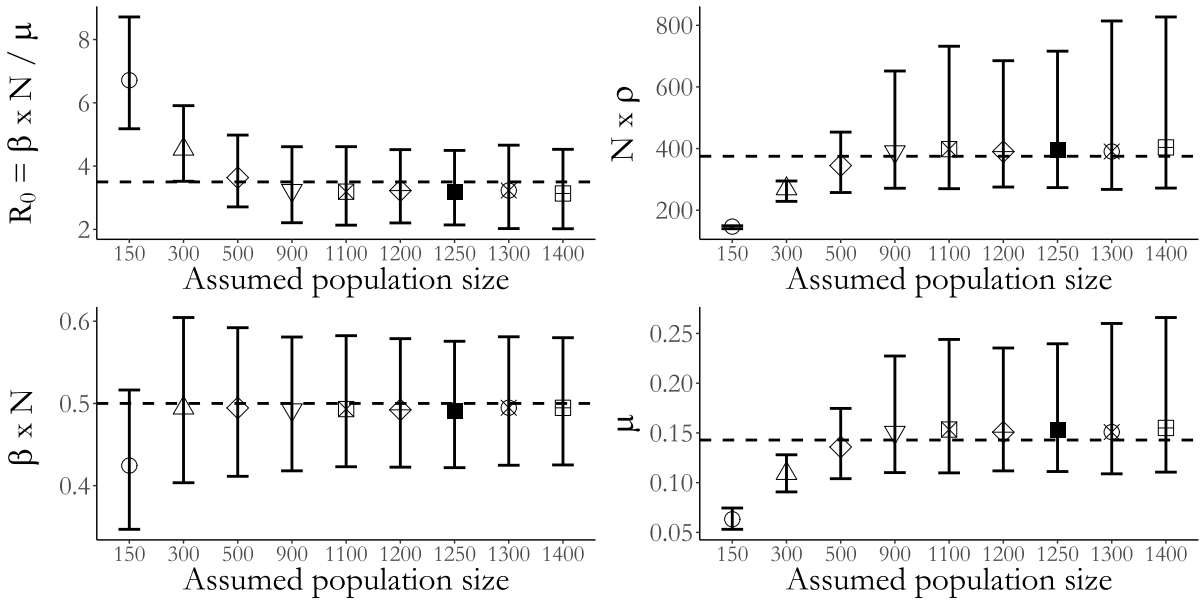


Figure 7: Posterior medians and 95% credible intervals for the basic reproductive number, R_0, infectivity rate, recovery rate, and binomial sampling probability scaled by the assumed population size. The dashed lines indicate the true values in the population of size 1,250. The population size, N, indicates the assumed population size used in fitting the model.

surveillance settings, and may result in biased estimates of the SEM dynamics. This bias is the result of a missmatch between the intensive dynamics of the epidemic process, which are a function of the fractions of individuals in the population who 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 counts reflects a high prevalence/low detection rate setting, or visa–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 $N = 1,250$ 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 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. The constrained ranges of model dynamics also manifest in a narrowing of the widths of the pointwise credible intervals for disease prevalence (Figure 8). Under severe misspecification of the population size (N = 150), the latent posterior distribution has 95% of its mass within only

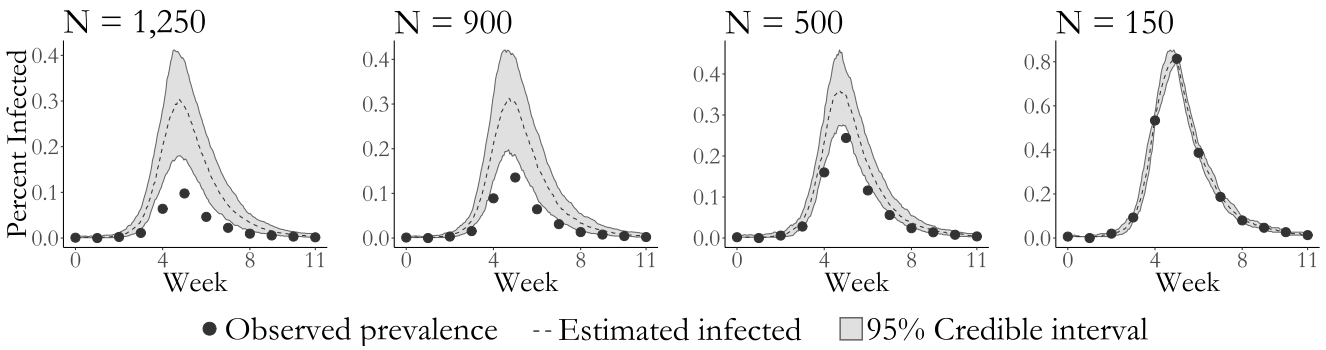


Figure 8: Estimated latent posterior distributions of disease prevalence under SIR dyamics. The true population size is 1,250. Depicted are the observed prevalence (dots), pointwise posterior median prevalence (dashed line), and pointwise 95% credible intervals (shaded region) all scaled by the assumed population size. Latent posterior estimates are based on a thinned sample, with every 250^{th} sample retained.

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 in order 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× 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 is available, we must be concerned with 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 $R_0 = \beta \times 763/\mu = 1.8375$ and the mean infectious period was $1/\mu = 7$ days. We fit four 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. Complete simulation details and convergence diagnostics are supplied in

Section S12.

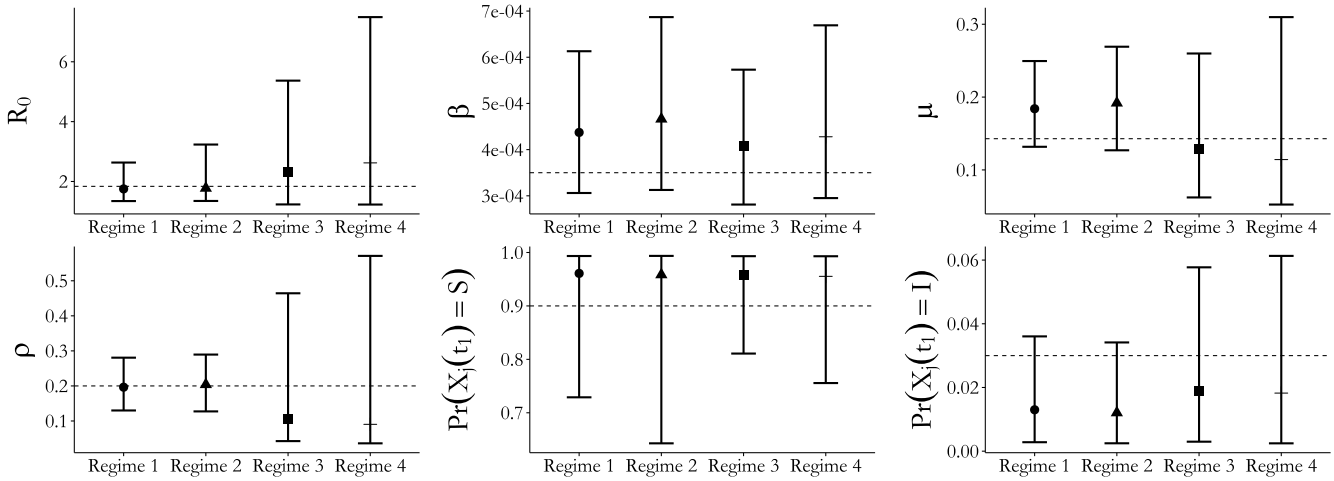


Figure 9: Posterior median estimates and 95% credible intervals for all SIR model parameters under four different prior regimes (Table S14). Regimes 1 and 3 set informative priors for the per–contact infectivity and recovery rates. Regimes 1 and 2 set informative priors for the binomial sampling probability.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

The true values for all model parameters fell within the 95% credible intervals under all four prior regimes. Informative priors tended to result in narrower credible intervals for the parameters (Figure 9) as well as for the latent process (Figure 10), though the effects of the prior for the detection probability were particularly pronounced. 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 mix somewhat better than chains with diffuse priors for the sampling probabilty (see traceplots in Section S12).

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 11.

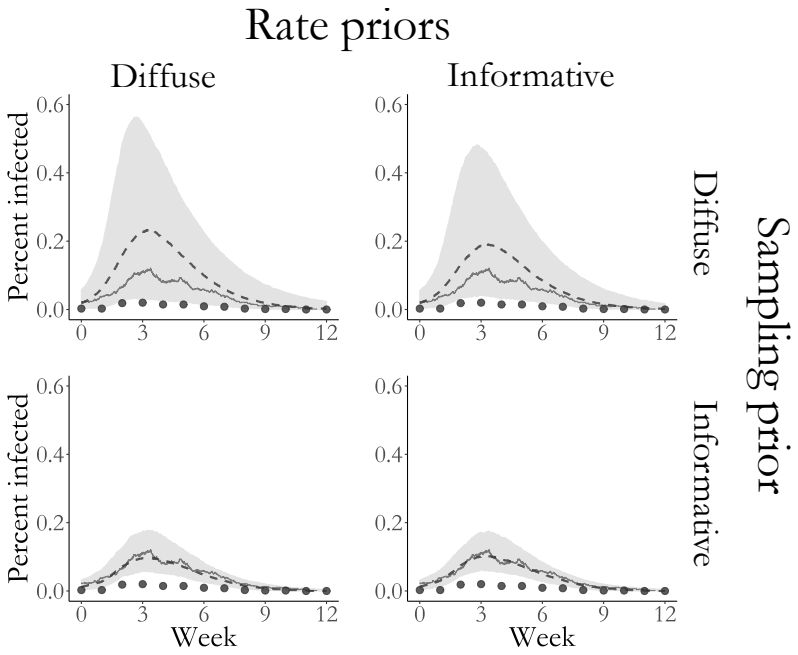


Figure 10: Estimated latent posterior distributions of disease prevalence in outbreaks simulated under four prior regimes for SIR model rate parameters and the binomial sampling probability. Depicted are the true unobserved prevalence (solid line), observed data (dots), pointwise posterior median prevalence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

We used our DA algorithm and PMMH to fit SIR and SEIR models with a binomial emission distribution to the data (see 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 in order 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 likely went unnoticed. The posterior median recovery rate under SIR dynamics corresponds to an average period of 2.16 days (95% 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 one to four days before symptoms manifest, and are typically infectious for

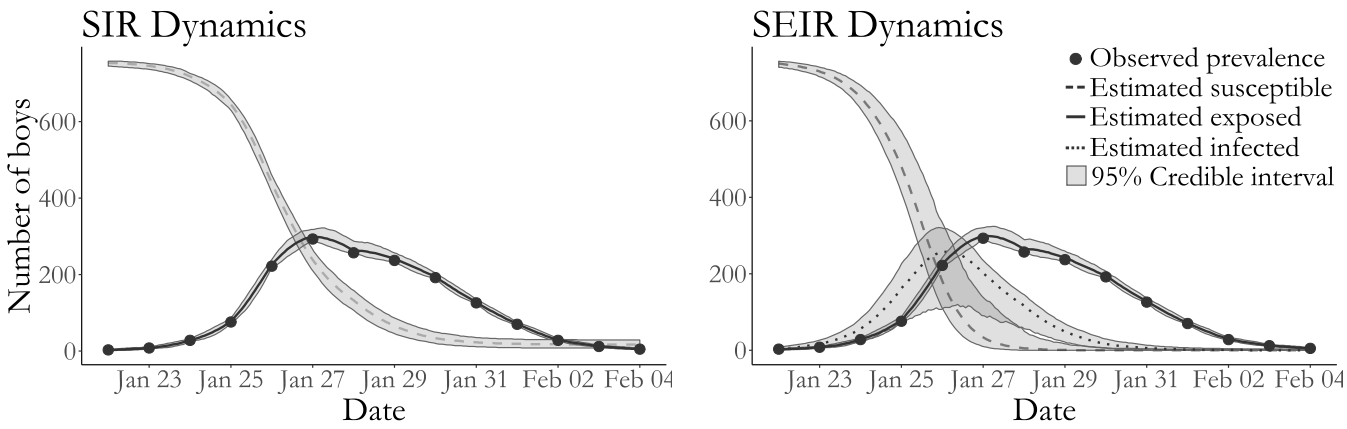


Figure 11: Boarding school data, pointwise posterior median estimates and pointwise 95% credible intervals (grey shaded areas) under SIR and SEIR dynamics of the numbers of susceptible boys (dashed line), exposed boys (dotted line), and infected boys (solid line). Posterior estimates based on a thinned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

one day before, and up to a week after, symptom onset (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 et al., 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 impossible, to carry out using traditional agent–based DA methods in the absence of subject–level data. Although we outlined the BDA 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 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

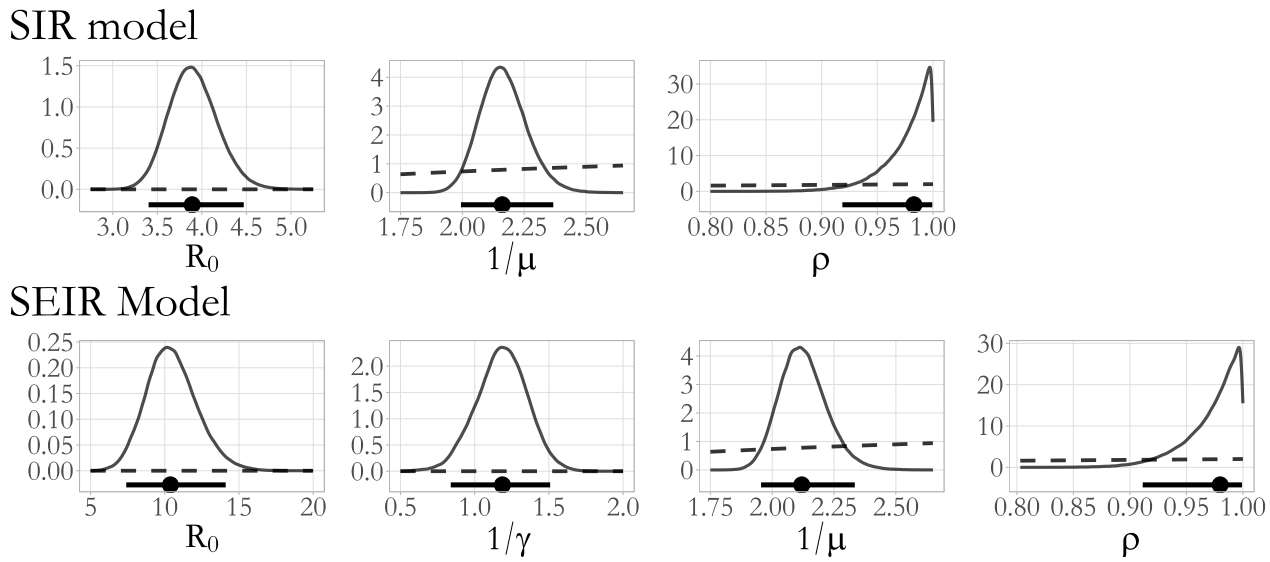


Figure 12: Posterior density estimates for $R_0 = \beta N/\mu$, the mean latent and infectious periods, $1/\gamma$ and $1/mu$, and the binomial sampling probability, ρ, from SIR and SEIR model parameters fit to the British boarding school data (solid lines). The posterior median and 95% Bayesian credible 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.

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 re-sampling 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 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, outbreak 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. 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 the application of our algorithm to more complex models fit to epidemics in large, structured populations.

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. We would also like to 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 paper.

References

A.J.S. Allen. An introduction to stochastic epidemic models. In *Mathematical Epidemiology*, pages 81–130. Springer, New York, 2008.

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

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:269–342, 2010.

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.

C Bretó and E.L. Ionides. Compound Markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Processes and their Applications*, 121:2571–2591, 2011.

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.

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 0.2.2.

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.

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.

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

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. Höhle and E. Jørgensen. 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.

A.A. Koepke, I.M. Longini Jr, M.E. Halloran, J. Wakefield, and V.N. Minin. Predictive modeling of cholera outbreaks in Bangladesh. *The Annals of Applied Statistics*, 10:575–595, 2016.

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 Å. 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*, 38:115–126, 1982.

T. McKinley, A.R. Cook, and R. Deardon. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5:1–40, 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.

C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45:3–49, 2003.

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

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. O'Neill 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.

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: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97:337–351, 2002.

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

A.Y. Shestopaloff and R.M. Neal. Sampling latent states for high-dimensional non-linear state space models with the embedded HMM method. *arXiv preprint arXiv:1602.06030v2*, 2016.

A. Sudbury. The proportion of the population never hearing a rumour. *Journal of Applied Probability*, 22:443–446, 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:e174, 2005.

D.J. Wilkinson. *Stochastic Modelling for Systems Biology*. CRC Press, Boca Raton, 2011.

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, depicted in Figure S1. This construction is usually

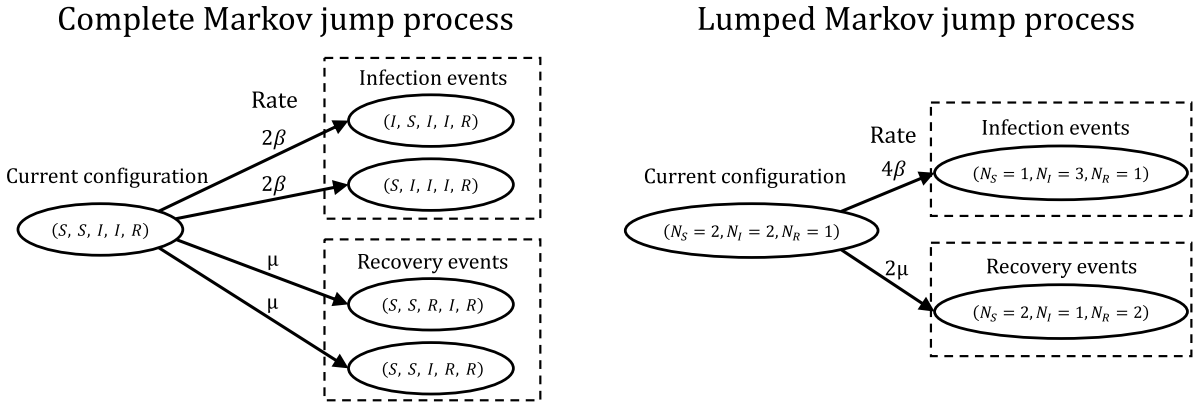


Figure S1: Complete and lumped representations of SIR dynamics in a population of five individuals. The per–contact infectivity rate, β, and the recovery rate, μ, parameterize exponential waiting time distributions between transition events. The complete Markov jump process evolves on the state space of subject state labels, $\mathcal{S} = \{S, I, R\}^N$, with dynamics determined by the subject–level transition rates. Each susceptible may contact two infected individuals, while each infected individual recovers independently. The lumped process evolves on the state space of compartment counts, $\widetilde{\mathcal{S}} = \{N_S, N_I, N_R : \ N_S + N_I + N_R = N\}$, with dynamics determined by lumped transition rates. The waiting time distributions between transitions are derived by noting that if $\tau_1 \sim \exp(\lambda_1)$ and $\tau_2 \sim \exp(\lambda_2)$, then $\tau_{\min} = \min(\tau_1, \tau_2) \sim \exp(\lambda_1 + \lambda_2)$.

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

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

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 Computing the matrix exponential

The transition probability matrix (TPM), $\mathbf{P}(t)$, for a time–homogeneous CTMC over an interval of length t, solves the matrix differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{P}(t) = \mathbf{\Lambda}\mathbf{P}(t), \quad \text{s.t. } \mathbf{P}(0) = \mathbf{I},$$

where $\mathbf{\Lambda}$ is the transition rate matrix for the CTMC and $\mathbf{I}$ is an identity matrix of the same size as $\mathbf{\Lambda}$ (Wilkinson, 2011). Therefore, $\mathbf{P}$ is computed using the matrix exponential solution of the above differential equation, $\mathbf{P} = \exp(t\mathbf{\Lambda})$. This is the most intensive step in our algorithm. However, we may lessen the computational burden to a large extent by leveraging the fact that we are computing the matrix exponential for the same rate matrix for possibly many values of t. Therefore, computing the matrix exponential using the eigen decomposition of $\mathbf{\Lambda}$ and caching the resulting eigenvalues and eigenvectors will be relatively efficient (Moler and Van Loan, 2003). We outline this computation in the following two cases: when the eigenvalues of $\mathbf{\Lambda}$ are all real (e.g. as with the SIR and SEIR models), and when $\mathbf{\Lambda}$ has complex eigenvalues (e.g. as is possibly the case with the SIRS model).

S2.1 Case 1: $\mathbf{\Lambda}$ has real eigenvalues

Suppose that $\mathbf{\Lambda}_{n\times n} = \mathbf{U}\mathbf{V}\mathbf{U}^{-1}$, where $\mathbf{V}$ is a diagonal matrix of eigenvalues, $v_1, \ldots, v_n$, and $\mathbf{U}$ is the matrix whose columns are the corresponding right eigenvectors. Then,

$$\mathrm{e}^{t\mathbf{V}} = \mathrm{diag}(\mathrm{e}^{v_1 t}, \ldots, \mathrm{e}^{v_n t}).$$

That $\mathbf{U}$ is nonsingular yields

$$\mathrm{e}^{t\mathbf{\Lambda}} = \mathbf{U}\mathrm{e}^{t\mathbf{V}}\mathbf{U}^{-1}.$$

S2.2 Case 2: $\mathbf{\Lambda}$ has complex eigenvalues

In the event that $\mathbf{\Lambda}$ has complex eigenvalues, we may obtain a real-valued TPM by transforming $\mathbf{\Lambda}$ into its real canonical form (Hirsch et al., 2013). Suppose that $\mathbf{\Lambda}$ has r real eigenvalues, $v_1, \ldots, v_r$, with corresponding real eigenvectors, $\mathbf{u}_1, \ldots, \mathbf{u}_r$, and $n - r$ pairs of complex conjugate eigenvalues. Let $(\mathbf{u}_j|\mathbf{w}_j)$ denote the real and imaginary parts of the eigenvector corresponding to the j^{th} eigenvalue, $\alpha_j + i\beta_j$, for $j = r+1, \ldots, n$, and define the matrix $\mathbf{T} = (\mathbf{u}_1|\ldots|\mathbf{u}_r|\mathbf{u}_{r+1}|\mathbf{w}_{r+1}|\ldots|\mathbf{u}_n|\mathbf{w}_n)$. The real canonical form for a rate matrix with complex eigenvalues can now be written as $\mathbf{V} = \mathbf{T}^{-1}\mathbf{\Lambda}\mathbf{T}$, where $\mathbf{V} = \mathrm{diag}(v_1, \ldots, v_r, \mathbf{B}_{r+1}, \ldots, \mathbf{B}_n)$, and each $\mathbf{B}_j,\ j = r+1, \ldots, n$ is given by

$$\mathbf{B}_j = \begin{pmatrix} \alpha_j & \beta_j \\ -\beta_j & \alpha_j \end{pmatrix},$$

which implies that

$$\mathrm{e}^{t\mathbf{B}_j} = \mathrm{e}^{\alpha_j t} \begin{pmatrix} \cos(\beta_j t) & \sin(\beta_j t) \\ -\sin(\beta_j t) & \cos(\alpha_j) \end{pmatrix},$$

and hence $\mathrm{e}^{t\mathbf{V}} = \mathrm{diag}(\mathrm{e}^{v_1 t}, \ldots, \mathrm{e}^{v_r t}, \mathrm{e}^{t\mathbf{B}_{r+1}}, \ldots, \mathrm{e}^{t\mathbf{B}_n})$. Therefore, we can compute the matrix exponential of $t\mathbf{\Lambda}$ as $\mathrm{e}^{t\mathbf{\Lambda}} = \mathbf{T}\mathrm{e}^{t\mathbf{V}}\mathbf{T}^{-1}$.

S3 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 (2.3). Then,

$$q_{j,r,s}^{(t_\ell)} \propto \pi_j^{(t_\ell)}\left(r \mid \mathbf{X}_{(-j)}, \boldsymbol{\theta}\right) \times \mathbf{P}_{r,s}^{(j)}\left(t_{\ell-1}, t_\ell\right) \times f\left(Y_{t_\ell} \mid \mathbf{X}_j(t_\ell), \mathbf{X}_{(-j)}(t_\ell), \rho, \mathbf{p}_{t_1}\right), \tag{11}$$

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 $\boldsymbol{Q}_j^{(t_{\ell+1})}$.

S4 Simulating endpoint Conditioned Time–Homogeneous CTMC Paths

In this section, we briefly summarize the modified rejection sampling and uniformization algorithms 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 the excellent paper by Hobolth and Stone (2009) for a more thorough discussion. We also refer the reader to the `ECctmc R` package for a fast implementation of these algorithms 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. We also denote by $\mathbf{P}(T)$ the transition probability matrix for the CTMC over $[0, T]$, and $P_{ab}(T)$ the probability of beginning in state a and ending in state b.

S4.1 Modified rejection sampling

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 $\mathbf{X}(T) = b$.

We sample the first transition time via the inverse–CDF method, sampling $u \sim \text{Unif}(0, 1)$ and applying the inverse-CDF function

$$F^{-1}(u) = \frac{-log\left[1 - u \times \left(1 - \mathrm{e}^{-T\Lambda_a}\right)\right]}{\Lambda_a}. \tag{12}$$

We found that the modified rejection algorithm worked well in fitting the SIR and SIRS models. In the examples we studied in which these models were fit, subject–paths over intervals where the endpoints required multiple jumps ($S \to R$, or $I \to S$) were almost never considered. Therefore, usually only a single transition time was required to be sampled in a given interval, and accomplishing this using the inverse–CDF method was quite fast.

S4.2 Uniformization

The uniformization algorithm samples the path for a time–homogeneous CTMC conditional on the state at the interval endpoints by coupling the original process to a Markov chain determined by an auxilliary Poisson point process. State transitions, including virtual transition where the state

does not change, occur at points of this auxilliary process, and the sequence of state labels is drawn from the corresponding Markov chain.

We construct the transition rate matrix of the auxilliary Markov chain, $\mathbf{Y}$, as $R = I + \frac{1}{\mu}\mathbf{\Lambda}$, where $\mu = \max_a \mathbf{\Lambda}_a$. Then number of state transitions, N, conditional on $\mathbf{X}(0) = a,\ \mathbf{X}(T) = b$, can be shown to be

$$P(N = n|\mathbf{X}(0) = a, \mathbf{X}(T) = b) = \mathrm{e}^{-\mu T}\frac{(\mu T)^n}{n!}R^n_{ab/}P_{ab}(T). \tag{13}$$

The algorithm proceeds by first sampling the number of state transitions from this distribution. If there are no transitions, or if there is one transition and the states at the endpoints are the same, the algorithm terminates. Otherwise, we drawn n independent uniform values in $[0, T]$ and sort them to obtain the times of state transitions. The state labels at the sorted sequence of times, τ_i, $i = 1, \ldots, n-1$, is then drawn from the discrete distribution with masses given by

$$P(\mathbf{X}(\tau_i)|\mathbf{X}(\tau_{i-1}, \mathbf{X}(T) = b) = \frac{R_{x_{i-1},x_i}(R^{n-i})_{x_i b}}{(R^{n-i+1})_{x_{i-1}b})}. \tag{14}$$

We found that uniformization was preferable to modified rejection sampling when fitting the SEIR model. In this case, modified rejection sampling tended to get hung up when sampling paths in intervals where the endpoints suggested that at least two state transitions occured (which though it seldom occured, significantly slowed down the MCMC). We also note that the transition probability, $P_{ab}(T)$, is computed and cached in carrying out the HMM step of our algorithm. Therefore, there are no additional eigen–decompositions or matrix exponentiations required in using the uniformization algorithm to sample the exact times of state transition.

S5 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 (7). 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,

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

S6 Fitting SEIR and SIRS models via Bayesian data augmentation

S6.1 SEIR model formulation

The Susceptible–Exposed–Infectious–Recovered (SEIR) model adds an additional latent state to the SIR model in which subjects who are exposed to an infected individual incubate before becoming infectious. As with the SIR model, recovery is assumed to confer lifelong immunity. The structure of this model does not affect any of the machinery involved in the subject–path proposal mechanism, but rather merely redefines the population–level time–homogeneous CMTC for the epidemic process, and the subject–level time–inhomogeneous CTMC used in the subject–path proposals.

Under this model, we suppose that 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 exposed, infectious, and recover. The state space of this process is $\mathcal{S} = \{S, E, I, R\}^N$, the Cartesian product of N state labels taking values in $\{S, E, I, R\}$. The state space of a single subject, $\mathbf{X}_j$, is $\mathcal{S}_j = \{S, E, I, R\}$, and a realized subject–path is of the form $\mathbf{x}_j(\tau) = \left(S,\ \tau < \tau_{\mathrm{E}}^{(j)}; E,\ \tau_{\mathrm{E}}^{(j)} \leqslant \tau < \tau_{\mathrm{I}}^{(j)}; I,\ \tau_{\mathrm{I}}^{(j)} \leqslant \tau < \tau_{\mathrm{R}}^{(j)}; R\ ,\ \tau_{\mathrm{R}}^{(j)} \leqslant \tau\right)$ where $\tau_{\mathrm{E}}^{(j)}$, $\tau_{\mathrm{I}}^{(j)}$, and $\tau_{\mathrm{R}}^{(j)}$ are the times at which subject J becomes exposed, infectious, and recovers. As with the SIR model, some or all of these events may not transpire in the observation period $[t_1, t_L]$, or at all. We let β be the per–contact infectivity rate, γ be the rate at which an exposed individual becomes infectious, and μ be the rate at which an infectious individual recovers. Furthermore, we write the vector of disease state probabilities as $\mathbf{p}_{t_1} = (p_S, p_E, p_I, p_R)$. The latent epidemic process evolves according to a time–homogeneous CTMC, with transition rate from configuration $\mathbf{x}$ to $\mathbf{x}'$ that differ only in the state of one subject j is given by $\mathbf{\Lambda} = \beta I$ if $\mathbf{X}_j = S$ and $\mathbf{X}'_j = E$, γ if $\mathbf{X}_j = E$ and $\mathbf{X}'_j = I$, and μ if $\mathbf{X}_j = I$ and $\mathbf{X}'_j = R$. Finally, the time–inhomogeneous CTMC rate matrices used in the subject–path proposal distribution have the form

$$\mathbf{\Lambda}_m^{(-j)}(\boldsymbol{\theta}) = \begin{array}{c} \\ S \\ E \\ I \\ R \end{array} \begin{array}{c} \begin{array}{cccc} S & E & I & R \end{array} \\ \begin{pmatrix} -\beta I_{\tau_m}^{(-j)} & \beta I_{\tau_m}^{(-j)} & 0 & 0 \\ 0 & -\gamma & \gamma & 0 \\ 0 & 0 & -\mu & \mu \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{15}$$

As is the case with the SIR model, the eigen–values of the CTMC rate matrices for the SEIR model are always real valued. The only computational modification, relative to the SIR model, that we suggest is that times of state transition in inter–event intervals be sampled conditional on the state at the endpoints via uniformization (see Section S4 of the supplement).

S6.2 SIRS model formulation

The Susceptible–Infected–Recovered–Susceptible (SIRS) model modifies the SIR model to allow for loss of immunity. Again, fitting this model using our Bayesian data augmentation algorithm

does not affect any of the machinery involved in the subject–path proposal mechanism, although the recurrent nature of the disease dynamics increase the computational burden of the algorithm since the disease state at the interval endpoints does absolve us of sampling the path within each inter–event interval where the states at the endpoints are the same.

Under the SIRS model, we suppose that 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 exposed, infectious, 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
$\mathbf{x}_j(\tau) = \Big(S,\ \tau < \tau^{(j)}_{\mathrm{I}_1}; I,\ \tau^{(j)}_{\mathrm{I}_1} \leqslant \tau < \tau^{(j)}_{\mathrm{R}_1}; R\ ,\ \tau^{(j)}_{\mathrm{R}_1} \leqslant \tau < \tau^{(j)}_{\mathrm{L}_1}; S\ ,\ \tau^{(j)}_{\mathrm{L}_1} \leqslant \tau < \tau^{(j)}_{\mathrm{I}_2}; \ldots\Big)$,
where $\tau^{(j)}_{\mathrm{I}_k}$, $\tau^{(j)}_{\mathrm{R}_k}$, and $\tau^{(j)}_{\mathrm{L}_k}$ are times at which subject J becomes infected, recovers, and loses immunity, and are ennumerated by the subscript k as the process may revisit each state multiple time. As with the SIR and SEIR models, some or all of these events may not transpire in the observation period $[t_1, t_L]$, or at all. We let β be the per–contact infectivity rate, μ be the rate at which an infectious individual recovers, and γ be the rate at which immunity is lost. Furthermore, we write the vector of disease state probabilities as $\mathbf{p}_{t_1} = (p_S, p_I, p_R)$. The latent epidemic process evolves according to a time–homogeneous CTMC, with transition rate from configuration $\mathbf{x}$ to $\mathbf{x}'$ that differ only in the state of one subject j is given by $\mathbf{\Lambda} = \beta I$ if $\mathbf{X}_j = S$ and $\mathbf{X}'_j = E$, μ if $\mathbf{X}_j = I$ and $\mathbf{X}'_j = R$, and γ if $\mathbf{X}_j = R$ and $\mathbf{X}'_j = S$. Finally, the time–inhomogeneous CTMC rate matrices used in the subject–path proposal distribution have the form

$$\mathbf{\Lambda}^{(-j)}_m(\boldsymbol{\theta}) = \begin{array}{c} \\ S \\ I \\ R \end{array} \begin{array}{c} \begin{array}{ccc} S & I & R \end{array} \\ \begin{pmatrix} -\beta I^{(-j)}_{\tau_m} & \beta I^{(-j)}_{\tau_m} & 0 \\ 0 & -\mu & \mu \\ \gamma & 0 & -\gamma \end{pmatrix} \end{array}. \tag{16}$$

Unlike the SIR and SEIR models, eigenvalues of each CTMC rate matrix may be complex. In order to obtain a real valued transition probability matrix over an interval for which eigen–values of the rate matrix are complex, we must rotate that rate matrix to obtain its real canonical form. This is further discussed in Section S2 of the Supplement.

S7 Selecting the Number of Subject–Paths to Update per MCMC Iteration

There is no need to re–sample the path of every subject within each MCMC 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, efficiency of the implementation, and the degree of model misspecification could affect the optimal number subject–path updates per MCMC iteration. It is clearly impossible to disentangle the effects of all of the possible factors that could affect the optimal number of subject–path updates per iteration. In the main paper, we set the number of subject–paths per iteration on the basis of log–posterior effective sample size (ESS) per CPU time in an initial run of 5,000–10,000 iterations (depending on the simulation).

S8 Prior and Full–Conditional Distributions of SEM parameters

Parameter	Conjugate Prior Dist.	Prior Hyperparameters	Full Conditional Hyperparameters
R_0	Beta$'$	$a_\beta,\ a_\mu,\ 1,\ \frac{b_\mu N}{b_\beta}$	—
β	Gamma	$a_\beta,\ b_\beta$	$a_\beta + \sum_{j=1}^{M} \mathbb{I}\left(\tau_j \triangleq I\right),\ b_\beta + \sum_{j=1}^{M} S_{\tau_{j-1}} I_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
μ	Gamma	$a_\mu,\ b_\mu$	$a_\mu + \sum_{j=1}^{M} \mathbb{I}\left(\tau_j \triangleq R\right),\ b_\mu + \sum_{j=1}^{M} I_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
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 S1: Prior and full conditional distributions for SIR model parameters. β is the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability, and $\mathbf{p}_{t_1}$ is the vector of initial state probabilities. Gamma priors are parameterized with rates, so a Gamma(a, b) distribution has mean a/b. The Beta prime prior for $R_0 = \beta N/\mu$ is the implied prior induced by the prior distributions for β and μ. The indicators $\mathbb{I}\left(\tau_j \triangleq I\right)$ and $\mathbb{I}\left(\tau_j \triangleq R\right)$ equal 1 if τ_j corresponds to a time when an individual becomes infected or recovers.

Parameter	Conjugate Prior Dist.	Prior Hyperparameters	Full Conditional Hyperparameters
R_0	Beta$'$	$a_\beta,\ a_\mu,\ 1,\ \frac{b_\mu N}{b_\beta}$	—
β	Gamma	$a_\beta,\ b_\beta$	$a_\beta + \sum_{j=1}^{M} \mathbb{I}\left(\tau_j \triangleq E\right),\ b_\beta + \sum_{j=1}^{M} S_{\tau_{j-1}} I_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
γ	Gamma	$a_\gamma,\ b_\gamma$	$a_\gamma + \sum_{j=1}^{M} \mathbb{I}\left(\tau_j \triangleq I\right),\ b_\gamma + \sum_{j=1}^{M} E_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
μ	Gamma	$a_\mu,\ b_\mu$	$a_\mu + \sum_{j=1}^{M} \mathbb{I}\left(\tau_j \triangleq R\right),\ b_\mu + \sum_{j=1}^{M} I_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
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 S2: Prior and full conditional distributions for SEIR model parameters. β is the per–contact infectivity rate, γ is the rate at which an exposed individual becomes infectious, μ is the recovery rate, ρ is the binomial sampling probability, and $\mathbf{p}_{t_1}$ is the vector of initial state probabilities. Gamma priors are parameterized with rates, so a Gamma(a, b) distribution has mean a/b. The Beta prime prior for $R_0 = \beta N/\mu$ is the implied prior induced by the prior distributions for β and μ. The indicators $\mathbb{I}\left(\tau_j \triangleq E\right)$, $\mathbb{I}\left(\tau_j \triangleq I\right)$ and $\mathbb{I}\left(\tau_j \triangleq R\right)$ equal 1 if τ_j corresponds to a time when an individual becomes exposed, becomes infectious, or recovers.

Parameter	Conjugate Prior Dist.	Prior Hyperparameters	Full Conditional Hyperparameters
R_0	Beta$'$	$a_\beta,\ a_\mu,\ 1,\ \frac{b_\mu N}{b_\beta}$	—
β	Gamma	$a_\beta,\ b_\beta$	$a_\beta + \sum_{j=1}^{M} \mathbb{I}\left(\tau_j \triangleq E\right),\ b_\beta + \sum_{j=1}^{M} S_{\tau_{j-1}} I_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
μ	Gamma	$a_\mu,\ b_\mu$	$a_\mu + \sum_{j=1}^{M} \mathbb{I}\left(\tau_j \triangleq R\right),\ b_\mu + \sum_{j=1}^{M} I_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
γ	Gamma	$a_\gamma,\ b_\gamma$	$a_\gamma + \sum_{j=1}^{M} \mathbb{I}\left(\tau_j \triangleq L\right),\ b_\gamma + \sum_{j=1}^{M} R_{\tau_{j-1}} (\tau_j - \tau_{j-1})$
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 S3: Prior and full conditional distributions for SIRS model parameters. β is the per–contact infectivity rate, μ is the recovery rate, γ is the rate at which a recovered individual loses immunity, ρ is the binomial sampling probability, and $\mathbf{p}_{t_1}$ is the vector of initial state probabilities. Gamma priors are parameterized with rates, so a Gamma(a, b) distribution has mean a/b. The Beta prime prior for $R_0 = \beta N/\mu$ is the implied prior induced by the prior distributions for β and μ. The indicators $\mathbb{I}\left(\tau_j \triangleq I\right)$, $\mathbb{I}\left(\tau_j \triangleq R\right)$, and $\mathbb{I}\left(\tau_j \triangleq L\right)$ equal 1 if τ_j corresponds to a time when an individual becomes infected, recovers, or loses immunity.

S9 Simulation 1 — Inference Under Various Epidemic Dynamics — Setup and Additional Results

S9.1 Simulation details for the SIR model

We simulated an epidemic in a population of 750 individuals, 90% of whom were initially susceptible and 3% of whom were initially infected. Prevalence was observed with detection probability $\rho = 0.2$ at weekly intervals over a four month period which captured both the exponential growth and decline of the epidemic. The mean infectious period was $1/\mu = 7$ days and the per-contact infectivity rate was 0.00035, which combined to give a basic reproductive number was $R_0 = \beta N/\mu \approx 1.8$.

We ran three chains for 100,000 iterations each, sampling the paths for 75 subjects, chosen uniformly at random, per MCMC iteration. We discarded the first 10 iterations from each chain as burn-in. Priors for the rate parameters (summarized in Table S4) were scaled so that the prior mass spanned a reasonable range of values, but were otherwise mild. Similarly, the prior for the binomial sampling probability reflected a general prior belief that fewer than 40% of cases were detected. The prior for the initial distribution parameters was informative, and was chosen as such because of the paucity of data available for estimation of the initial distribution parameters.

We also fit the SIR model to the data using PMMH. We ran two sets of three MCMC chains with the PMMH algorithm for 50,000 iterations each with 100 particles per chain, and discarded the first 100 iterations as burn-in. The first set of chains simulated particle paths approximately using τ–

Param.	True Value	Prior distribution
R_0	1.8	Beta′(0.3, 1, 1, 6)
β	0.00035	Gamma(0.3, 1000)
μ	0.14	Gamma(1, 8)
$\mathbf{p}_{t_1}$	(0.9, 0.03, 0.07)	Dirichlet(90, 2, 5)
ρ	0.2	Beta(2, 7)

Table S4: Prior distributions for SIR model and measurement process parameters. The prior for R_0 is the induced prior implied by β and μ. The per–contact infectivity rate is β, the recovery rate is μ, the binomial sampling probability is ρ, and the initial state probabilities are $\mathbf{p}_{t_1}$.

leaping with a time step of two hours, while the second chain simulated paths exactly via Gillespie's direct algorithm. Parameters were updated using random walk Metropolis–Hastings (RWMH) with a proposal covariance matrix estimated from an initial run of 5,000 iterations using an adaptive RWMH algorithm with a target acceptance rate of 23.4%. We updated parameters on transformed scales in order to remove restrictions on the parameter space, applying a log transformation to β and μ, a logit transformation to ρ, and a generalized logit transformation to $\mathbf{p}_{t_1}$.

S9.2 Additional results and MCMC diagnostics for the SIR model

Method	Chain	Hours	ESS	ESS per CPU time
BDA	1	9.9	87.7	8.8
BDA	2	8.7	67.9	7.8
BDA	3	8.5	63.8	7.5
PMMH–A	1	0.6	1847.4	2871.6
PMMH–A	2	0.6	1942.2	2995.7
PMMH–A	3	0.7	1876.6	2615.9
PMMH–E	1	26.1	1568.3	60.1
PMMH–E	2	20.4	2123.7	104.0
PMMH–E	3	20.5	1849.4	90.2

Table S5: Log–posterior run times, effective sample sizes (ESSs), and effective sample sizes per CPU time measure in hours (ESS.per.CPU.time). BDA indicates our Bayesian data augmentation algorithm, PMMH–A indicates PMMH with paths simulated approximately via τ–leaping algorithm, and PMMH–E indicates PMMH with paths simulated exactly using Gillespie's direct algorithm. The BDA chains were run for 100,000 iterations each, while the PMMH chains were run for 50,000 iterations following a tuning run of 5,000 iterations.

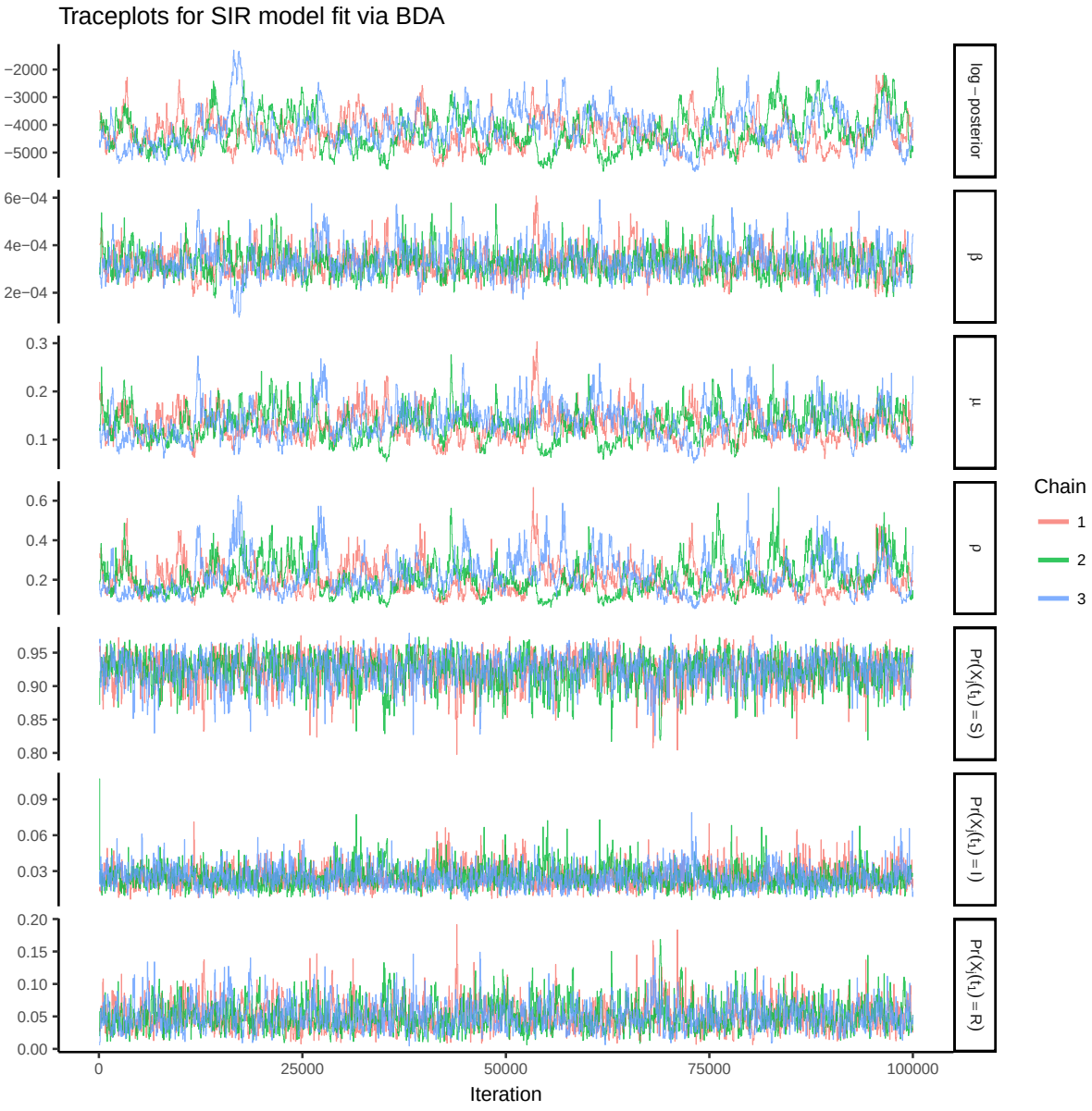


Figure S2: Traceplots of the log–posterior and model parameters for the SIR model fit using Bayesian data augmentation following an initial burn–in of 10 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

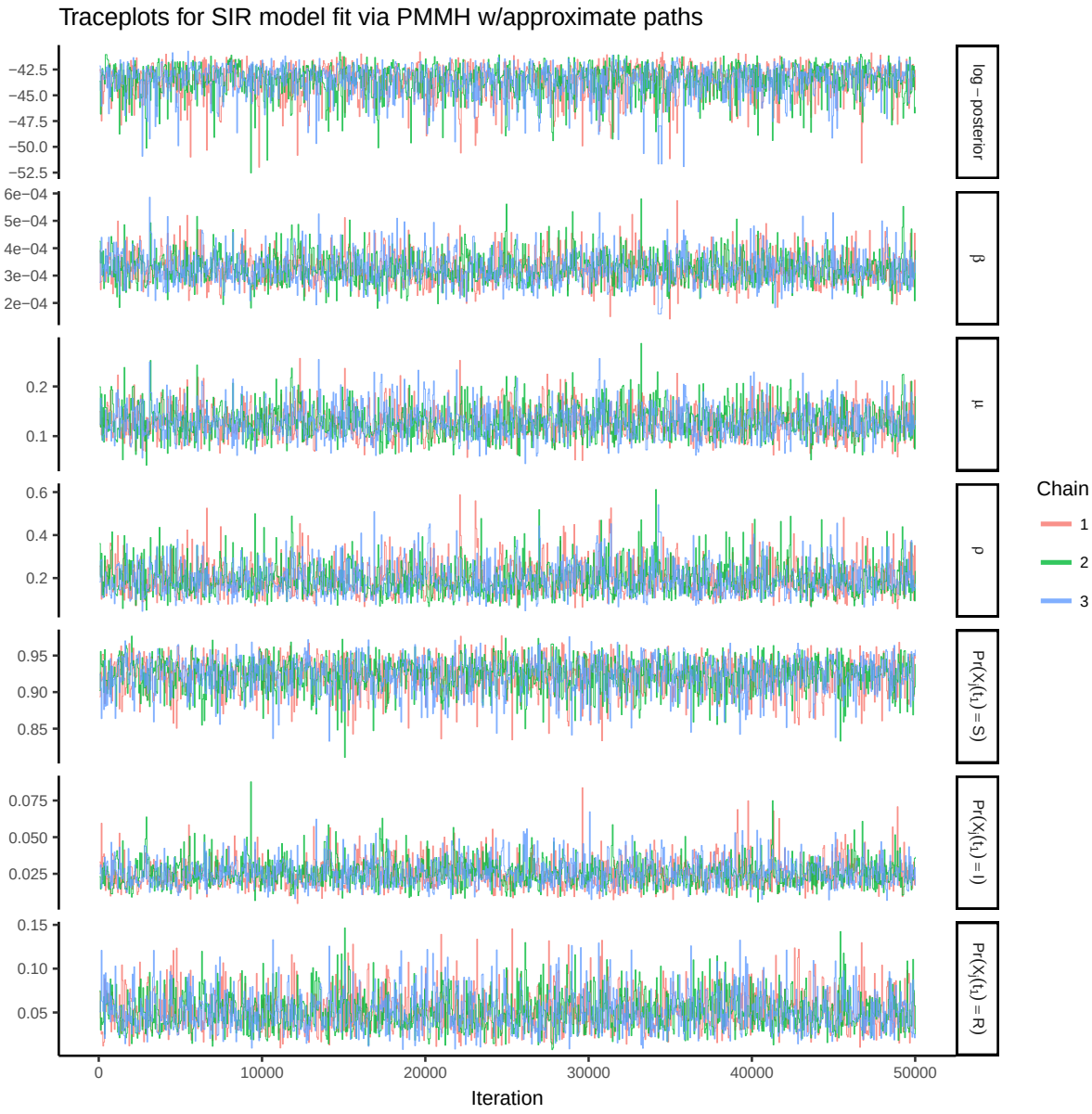


Figure S3: Traceplots of the log–posterior and model parameters for the SIR model fit using PMMH with 100 particles and a time step of 8 hours, following a tuning run of 5,000 iterations used to estimate the covariance matrix for the RWMH and an initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

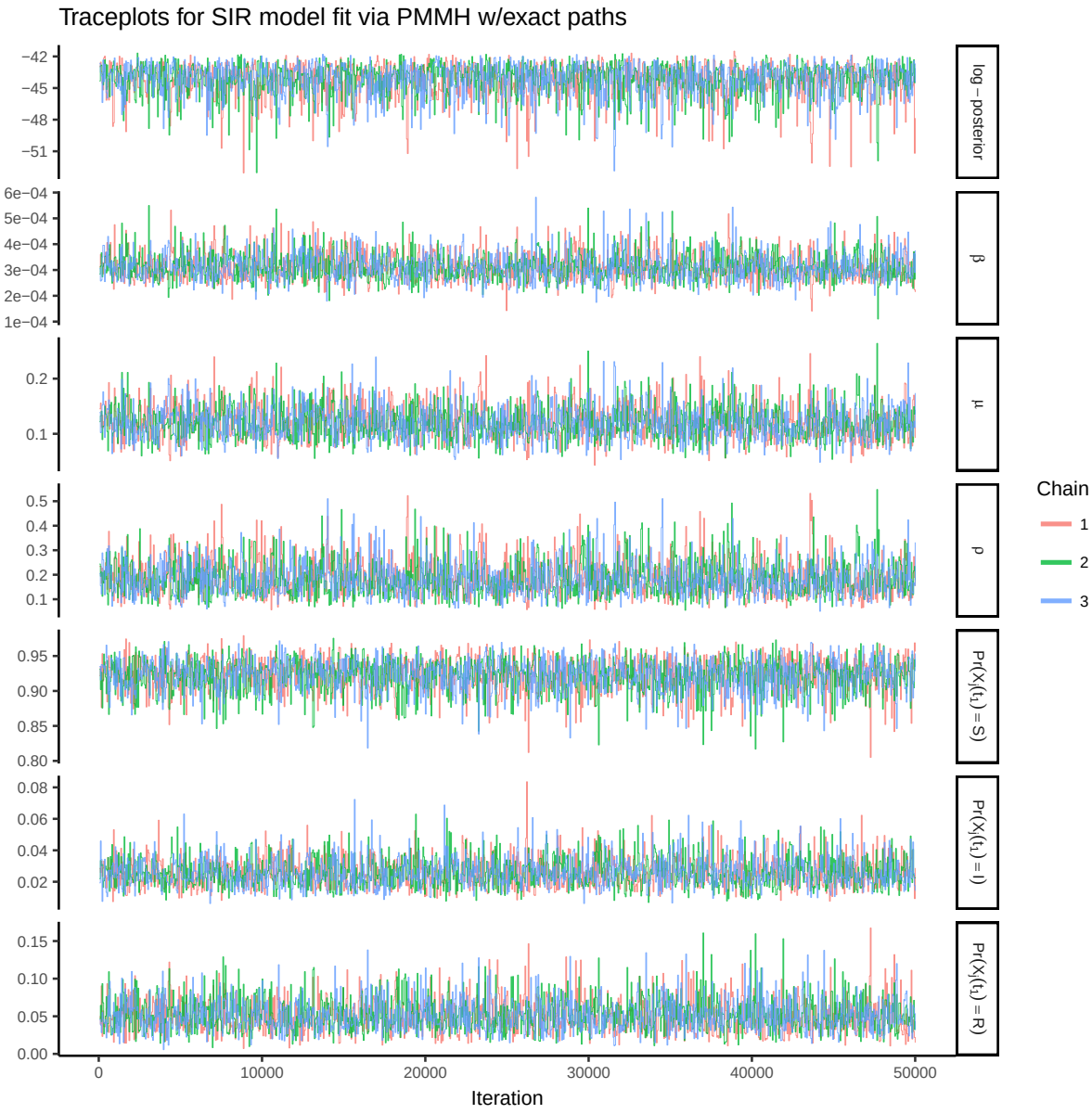


Figure S4: Traceplots of the log–posterior and model parameters for the SIR model fit using PMMH with 100 particles, following a tuning run of 5,000 iterations used to estimate the covariance matrix for the RWMH and an initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

S9.3 Simulation details for the SEIR model

We simulated an outbreak under near-endemic SEIR dynamics, with $R_0 = \beta N/\mu = 1.05$, in a population of 500 individuals. The outbreak was initiated by a single infected individual in an otherwise susceptible population, 121 of whom eventually became infected. The mean sojourn time in the exposed state was $1/\gamma = 14$ days, while the mean infectious period duration was $1/\mu = 28$ days. Prevalence was observed at weekly intervals, with detection probability $\rho = 0.3$, over a two year period.

We ran three chains for 100,000 iterations each, sampling the paths for 100 subjects, chosen uniformly at random, per MCMC iteration. We discarded the first 10 iterations from each chain as burn-in. Priors for the rate parameters (summarized in Table S6) were scaled so that the prior mass spanned a reasonable range of values, but were otherwise mild. The prior for the binomial sampling probability was chosen so that 80% of the mass was between roughly 15 and 55 percent. The prior for the initial distribution parameters was informative.

Param.	True Value	Prior distribution
R_0	1.05	Beta′(1, 3.2, 1, 5)
β	0.000075	Gamma(1, 10000)
γ	0.071	Gamma(1, 11)
μ	0.036	Gamma(3.2, 100)
$\mathbf{p}_{t_1}$	(0.998, 0.006, 0.002, 0, 0)	Dirichlet(100, 0.1, 0.4, 0.01)
ρ	0.3	Beta(3.5, 6.5)

Table S6: Prior distributions for SEIR model and measurement process parameters. The prior for R_0 is the induced prior implied by β and μ. The per–contact infectivity rate is β, the rate at which an exposed individual becomes infectious is γ, the recovery rate is μ, the binomial sampling probability is ρ, and the initial state probabilities are $\mathbf{p}_{t_1}$.

We also fit the SEIR model to the data using PMMH. We ran two sets of three MCMC chains with the PMMH algorithm for 50,000 iterations each with 200 particles per chain, and discarded the first 100 iterations as burn-in. The first set of chains simulated particle paths approximately using τ–leaping with a time step of 8 hours, while the second chain simulated paths exactly via Gillespie's direct algorithm. Parameters were updated using random walk Metropolis–Hastings (RWMH) with a proposal covariance matrix estimated from an initial run of 5,000 iterations using an adaptive RWMH algorithm with a target acceptance rate of 23.4%. We updated parameters on transformed scales in order to remove restrictions on the parameter space, applying a log transformation to β, γ, and μ, a logit transformation to ρ, and a generalized logit transformation to $\mathbf{p}_{t_1}$.

S9.4 Additional results and MCMC diagnostics for the SEIR model

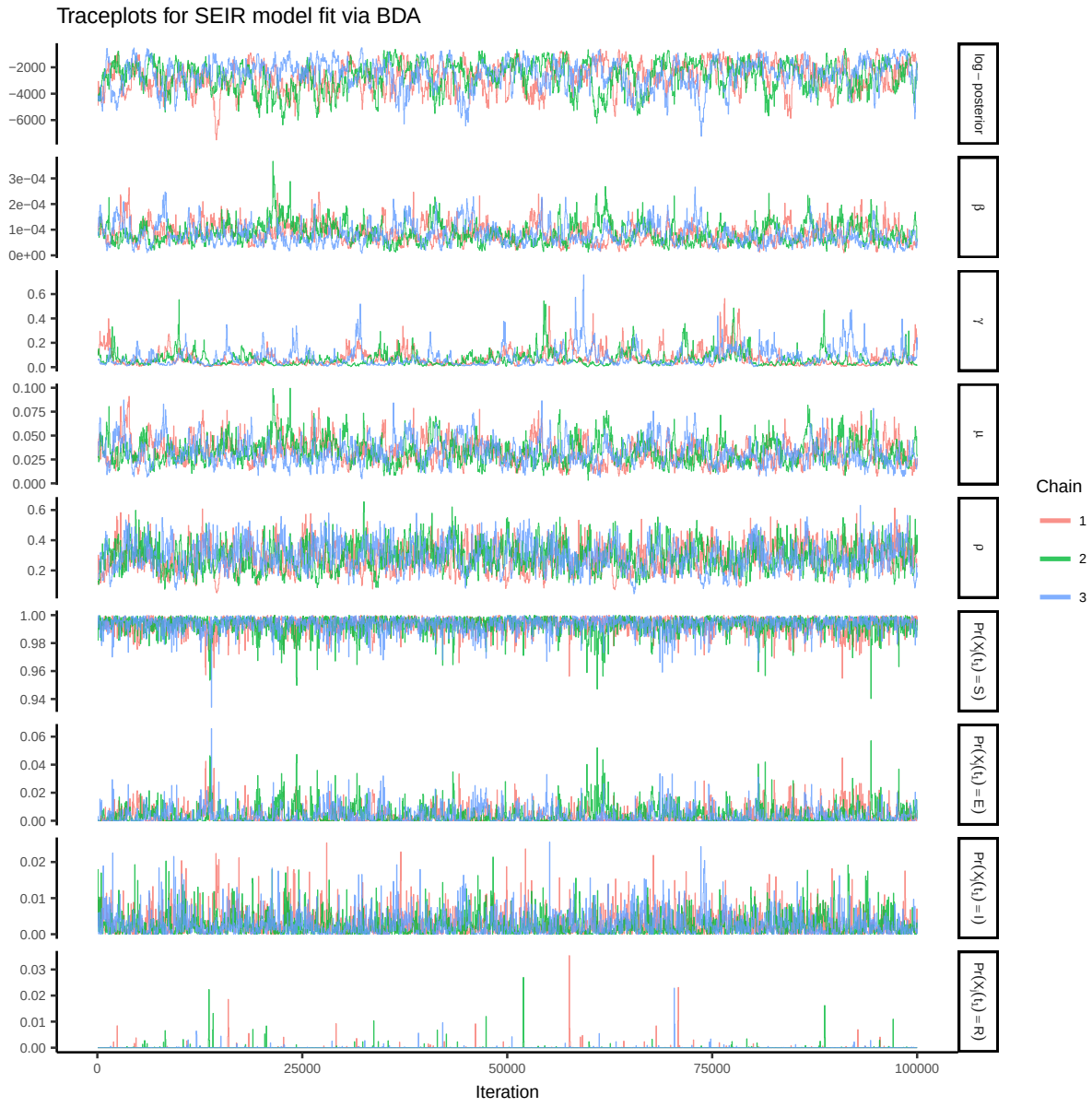


Figure S5: Traceplots of the log–posterior and model parameters for the SIR model fit using Bayesian data augmentation following an initial burn–in of 10 iterations. β denotes the per–contact infectivity rate, γ is the rate at which exposed individuals become infectious, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

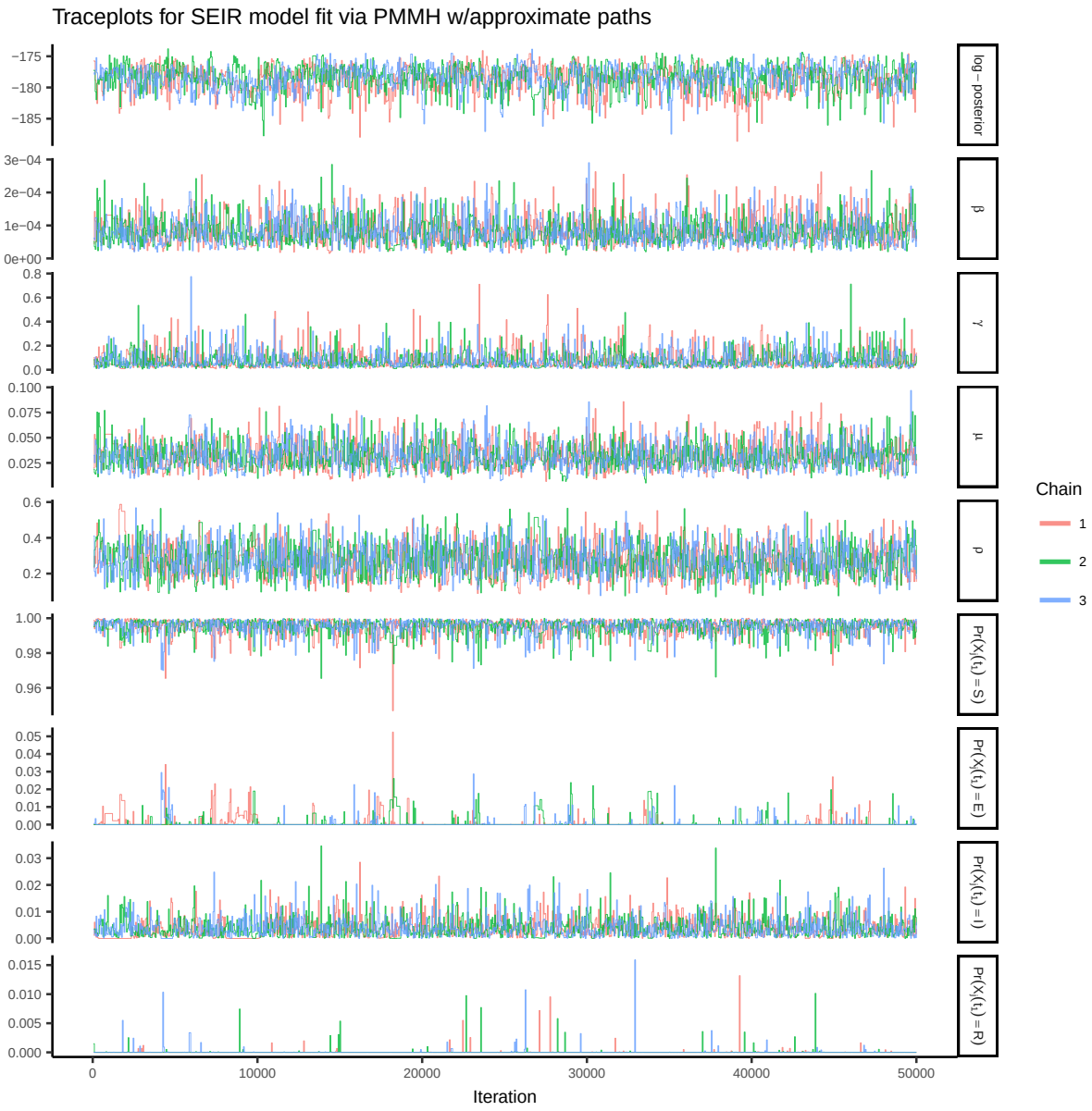


Figure S6: Traceplots of the log–posterior and model parameters for the SIR model fit using PMMH with 200 particles and a time step of 8 hours, following a tuning run of 5,000 iterations used to estimate the covariance matrix for the RWMH and an initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, γ is the rate at which exposed individuals become infectious, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

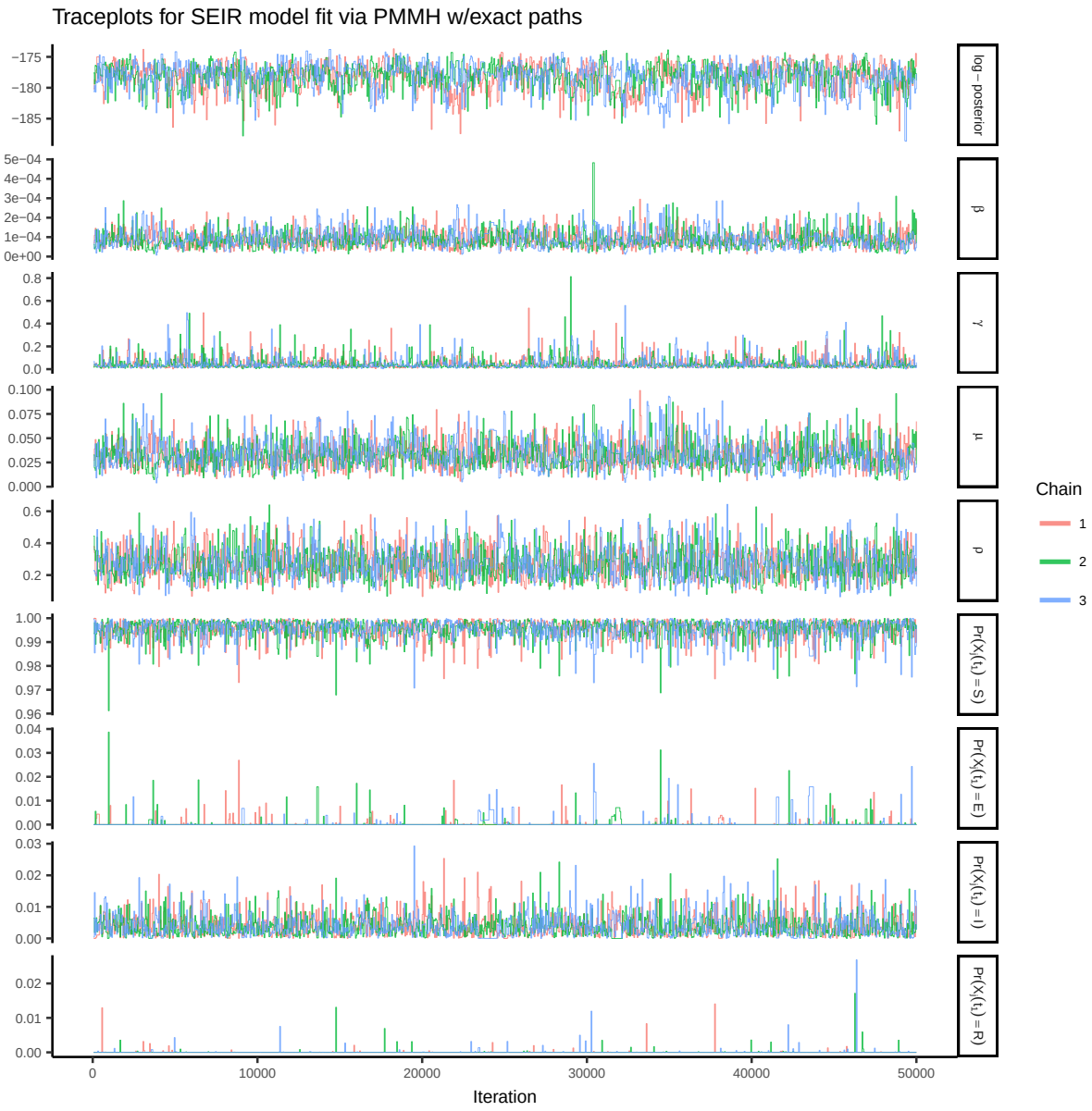


Figure S7: Traceplots of the log–posterior and model parameters for the SIR model fit using PMMH with 200 particles, following a tuning run of 5,000 iterations used to estimate the covariance matrix for the RWMH and an initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, γ is the rate at which exposed individuals become infectious, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

Model	Method	Chain	Time	ESS	ESS per CPU time
SEIR	BDA	1	9.2	149.9	16.2
SEIR	BDA	2	9.2	146.0	15.9
SEIR	BDA	3	9.0	143.9	16.0
SEIR	PMMH - A	1	8.1	483.6	59.5
SEIR	PMMH - A	2	8.3	684.8	82.2
SEIR	PMMH - A	3	8.4	570.5	67.9
SEIR	PMMH - E	1	15.8	411.9	26.1
SEIR	PMMH - E	2	15.9	589.8	37.1
SEIR	PMMH - E	3	14.1	466.3	33.1

Table S7: Log–posterior run times, effective sample sizes (ESSs), and effective sample sizes per CPU time measure in hours (ESS.per.CPU.time). BDA indicates our Bayesian data augmentation algorithm, PMMH–A indicates PMMH with paths simulated approximately via τ–leaping algorithm, and PMMH–E indicates PMMH with paths simulated exactly using Gillespie's direct algorithm. The BDA chains were run for 100,000 iterations each, while the PMMH chains were run for 50,000 iterations following a tuning run of 5,000 iterations.

S9.5 Simulation details for the SIRS model

The final outbreak was simulated under SIRS dynamics in a population of 200 individuals, in which $R_0 = \beta N/\mu = 2.52$, the mean infectious period was $1/\mu = 14$ days, and the mean time until loss of immunity was $1/\gamma = 150$ days. One percent of of the population was initially at the time of the first observation and the rest of the individuals were susceptible. Prevalence was observed weekly, with detection probability $\rho = 0.95$, over a one year period that spanned the initial wave of the epidemic as well as most of the second wave of the epidemic.

We ran three chains for 100,000 iterations each, sampling the paths for 100 subjects, chosen uniformly at random, per MCMC iteration. We discarded the first 2,000 iterations from each chain as burn-in. Priors for the rate parameters (summarized in Table S8) were scaled so that the prior mass spanned a reasonable range of values, but were otherwise mild. Similarly, the prior for the binomial sampling probability reflected a general prior belief that more than 60% of cases were detected, but was not otherwise particularly informative. The prior for the initial distribution parameters was informative.

We also fit the SIRS model to the data using PMMH. We ran three MCMC chains with the PMMH algorithm for 50,000 iterations each with 500 particles per chain, and discarded the first 100 iterations as burn-in. We also ran a set of chains with 200 particles but mixing was poor and not all of the chains converged. We attempted to exactly simulate particle paths but ultimately failed due to degeneracies in the algorithm. The time step for the τ–leaping algorithm was 8 hours. Parameters were updated using random walk Metropolis–Hastings (RWMH) with a proposal covariance matrix estimated from an initial run of 5,000 iterations using an adaptive RWMH algorithm with a target acceptance rate of 23.4%. We updated parameters on transformed scales in order to remove restrictions on the parameter space, applying a log transformation to β, μ, and γ, a logit transformation to ρ, and a generalized logit transformation to $\mathbf{p}_{t_1}$.

Param.	True Value	Prior distribution
R_0	2.52	Beta$'$(0.1, 1.5, 1, 28)
β	0.1	Gamma$(0.1, 100)$
μ	0.036	Gamma(1.8, 14)
γ	0.071	Gamma$(0.0625, 10)$
$\mathbf{p}_{t_1}$	(0.99, 0.01, 0)	Dirichlet(90, 1.5, 0.01)
ρ	0.95	Beta(5, 1)

Table S8: Prior distributions for SIRS model and measurement process parameters. The prior for R_0 is the induced prior implied by β and μ. The per–contact infectivity rate is β, the recovery rate is μ, the rate at which immunity is lost is γ, the binomial sampling probability is ρ, and the initial state probabilities are $\mathbf{p}_{t_1}$.

S9.6 Additional results and MCMC diagnostics for the SIRS model

Model	Method	Chain	Time	ESS	ESS per CPU time
SIRS	BDA	1	14.2	167.7	11.8
SIRS	BDA	2	10.9	194.8	17.8
SIRS	BDA	3	10.8	243.0	22.6
SIRS	PMMH - A	1	3.1	670.8	214.1
SIRS	PMMH - A	2	3.0	799.5	267.3
SIRS	PMMH - A	3	3.5	766.2	217.1
SIRS	PMMH - E	1	50.2	570.9	11.4
SIRS	PMMH - E	2	48.6	667.6	13.7
SIRS	PMMH - E	3	48.8	592.6	12.1

Table S9: Log–posterior run times, effective sample sizes (ESSs), and effective sample sizes per CPU time measure in hours (ESS.per.CPU.time). BDA indicates our Bayesian data augmentation algorithm, PMMH–A indicates PMMH with paths simulated approximately via τ–leaping algorithm, and PMMH–E indicates PMMH with paths simulated exactly using Gillespie's direct algorithm. The BDA chains were run for 100,000 iterations each, while the PMMH chains were run for 50,000 iterations following a tuning run of 5,000 iterations.

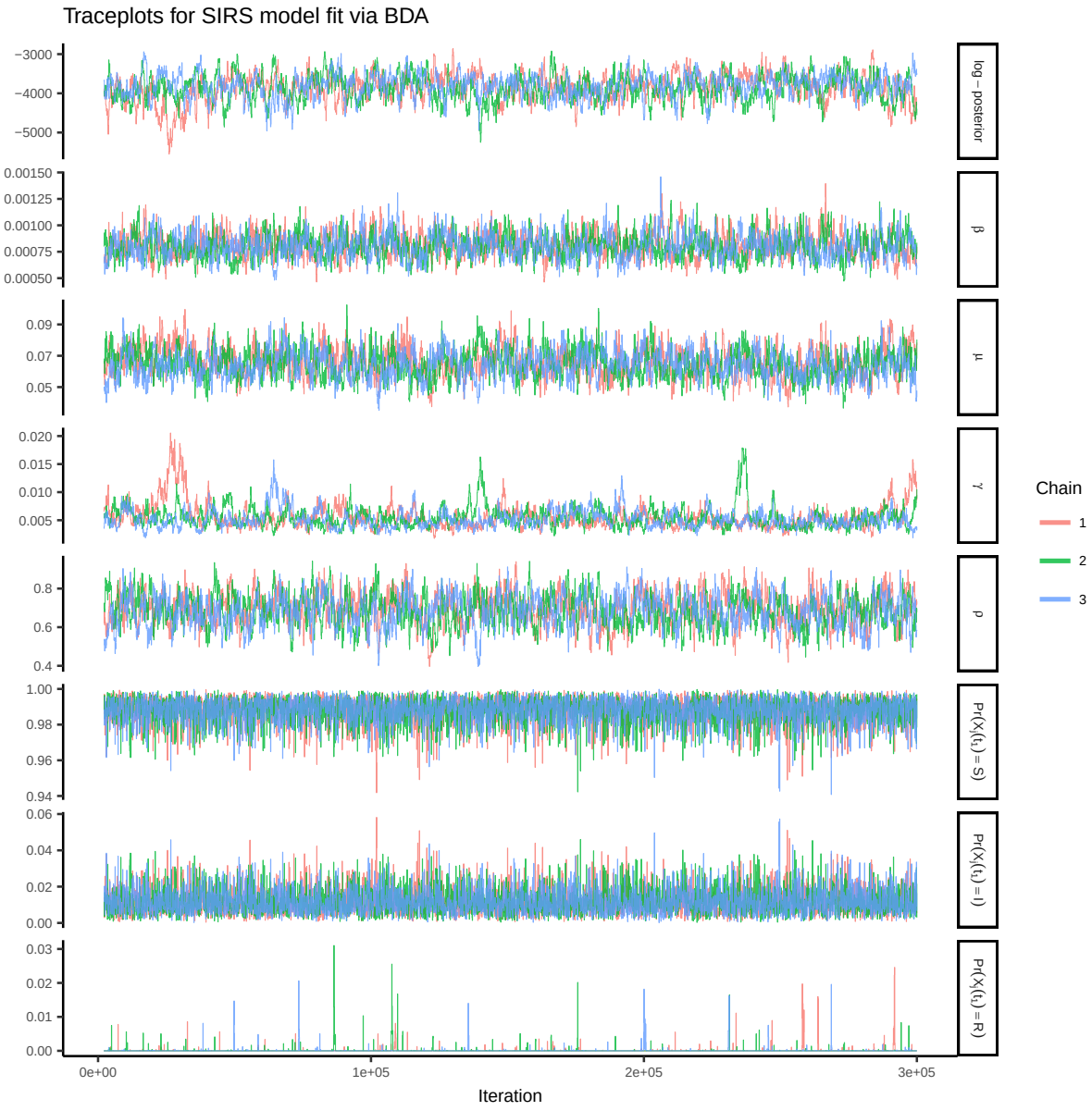


Figure S8: Traceplots of the log–posterior and model parameters for the SIR model fit using Bayesian data augmentation following an initial burn–in of 2,000 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, *gamma* is the rate at which immunity is lost, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

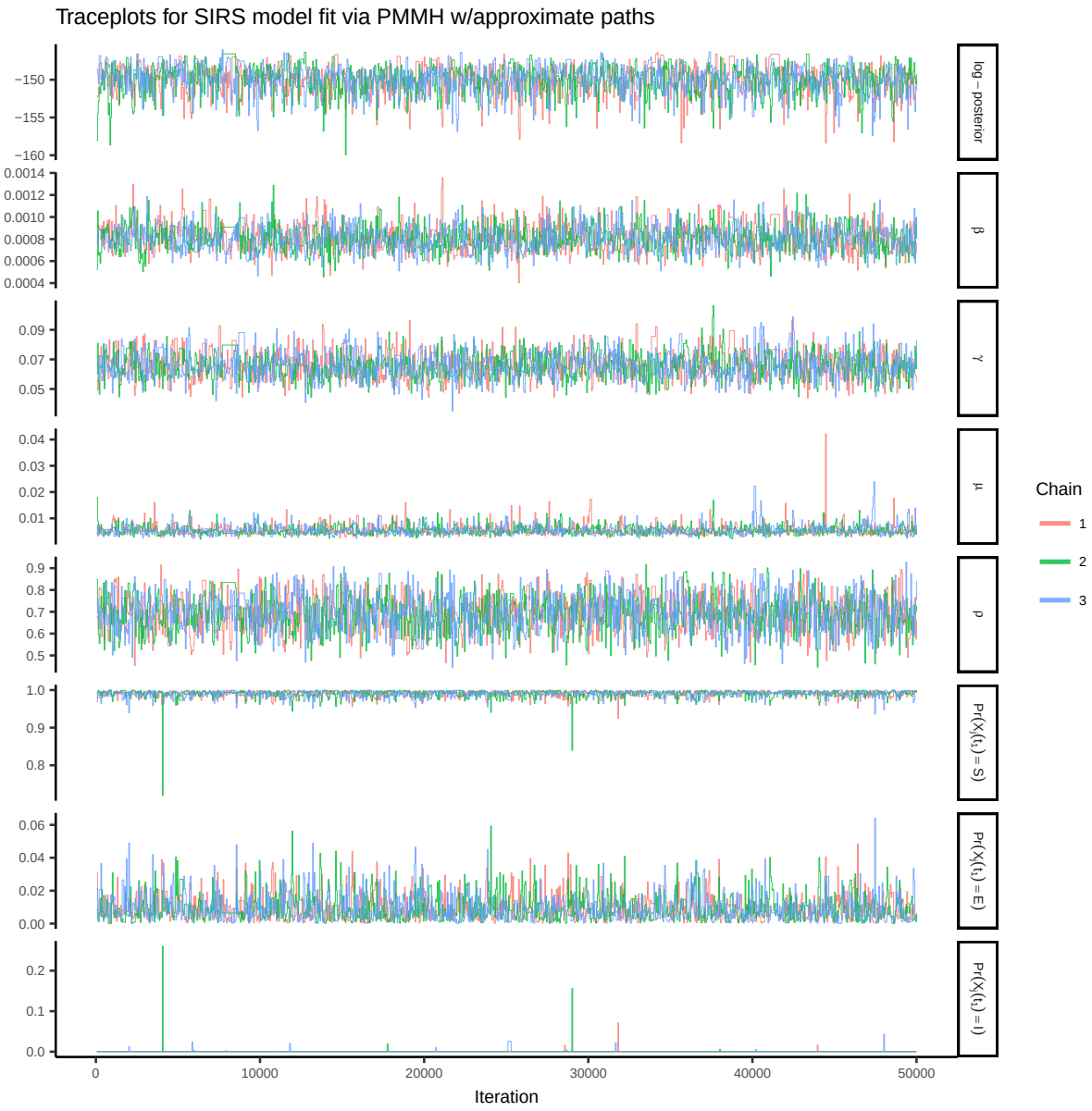


Figure S9: Traceplots of the log–posterior and model parameters for the SIR model fit using PMMH with 500 particles per chain and a time–step of 8 hours in the approximate τ–leaping algorithm, following a tuning run of 5,000 iterations to estimate the RWMH covariance matrix and in initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, *gamma* is the rate at which immunity is lost, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

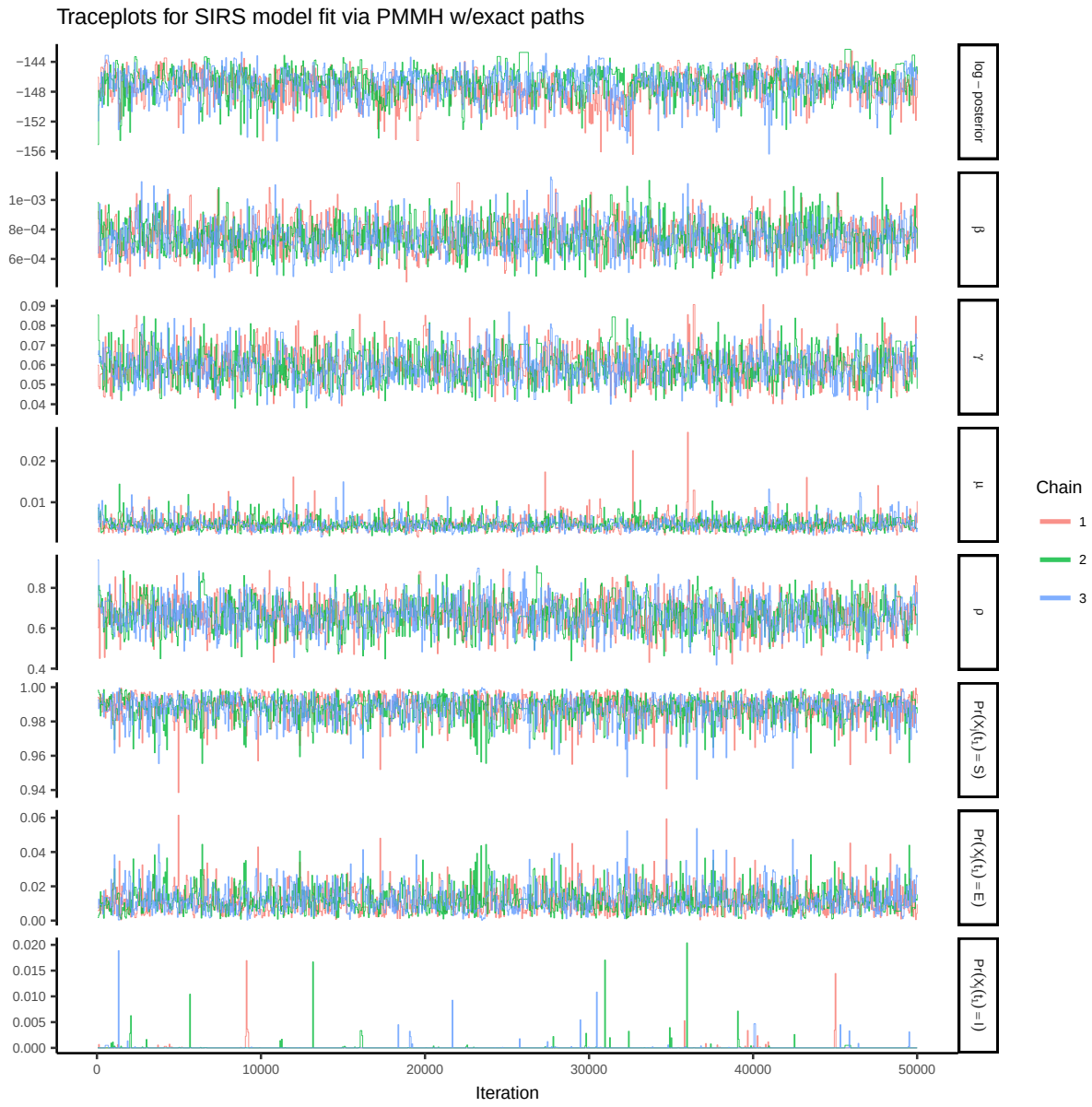


Figure S10: Traceplots of the log–posterior and model parameters for the SIR model fit using PMMH with 500 particles per chain and particle paths simulated exactly via Gillespie's direct algorithm, following a tuning run of 5,000 iterations to estimate the RWMH covariance matrix and in initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, *gamma* is the rate at which immunity is lost, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

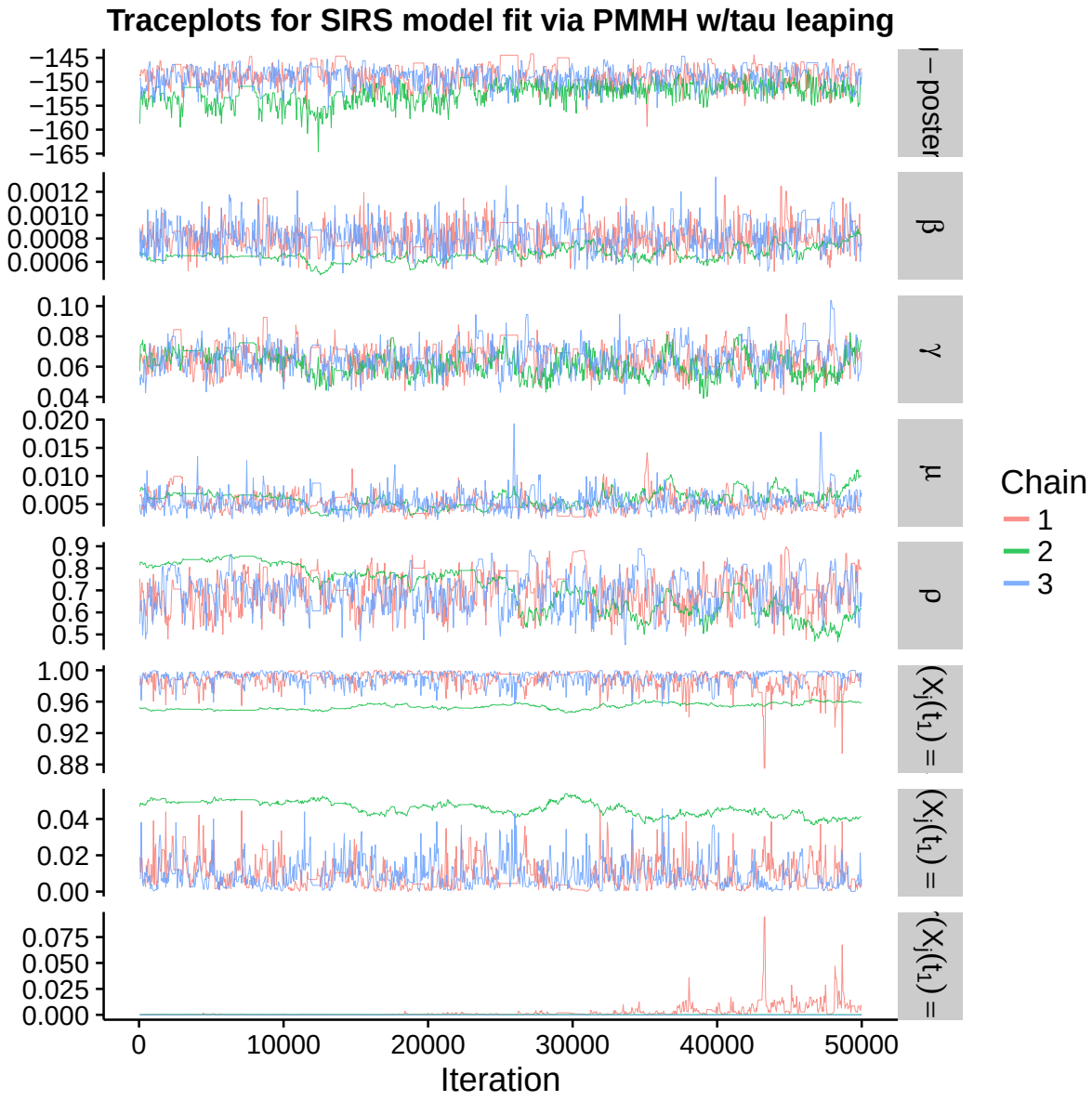


Figure S11: Traceplots of the log–posterior and model parameters for the SIR model fit using PMMH with 200 particles per chain and a time–step of 8 hours in the approximate τ–leaping algorithm, following a tuning run of 5,000 iterations to estimate the RWMH covariance matrix and in initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, *gamma* is the rate at which immunity is lost, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

S9.7 Estimated latent posterior distributions for all models

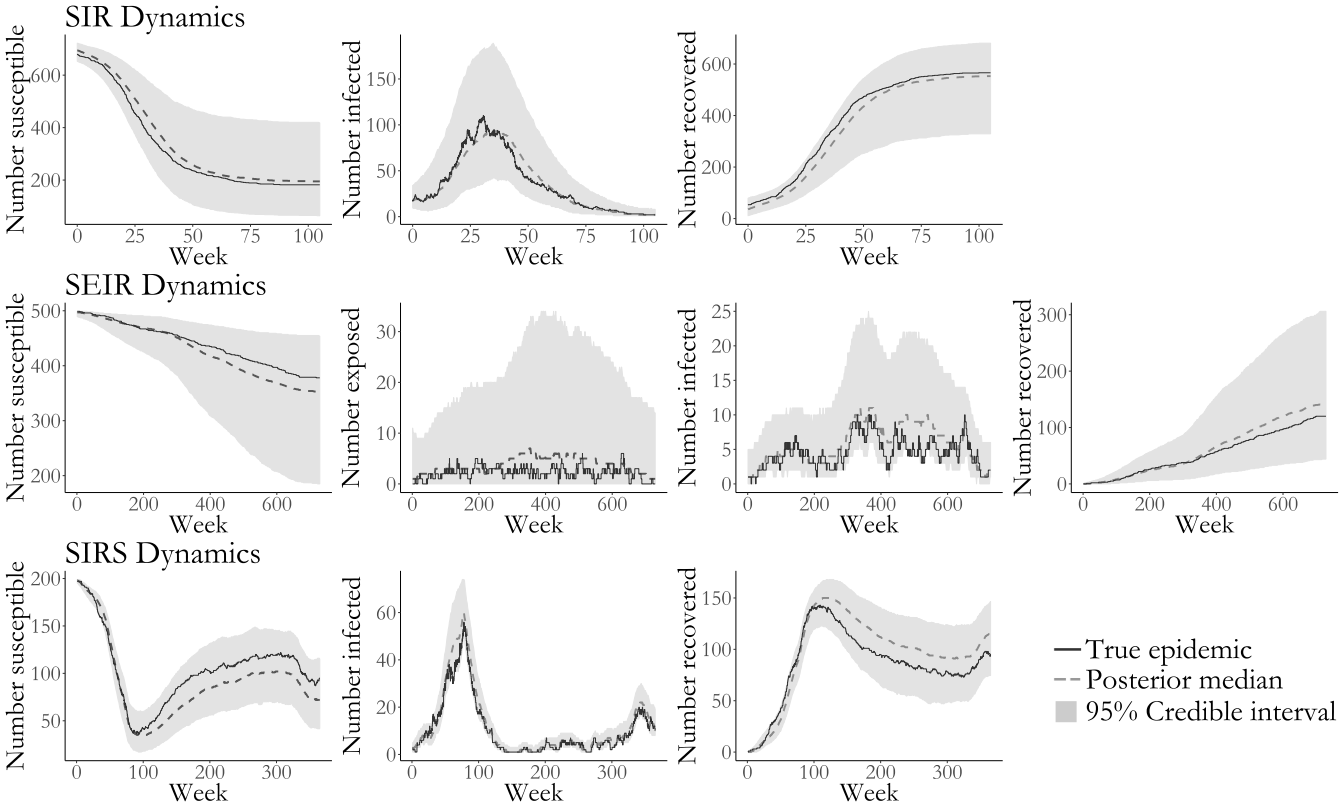


Figure S12: Pointwise posterior medians (dashed lines) and pointwise 95% credible intervals for the numbers of individuals in each disease state for the SIR, SEIR, and SIRS models. True compartment counts are shown as solid lines. Estimates are based on a thinned sample, retaining the collection of disease histories at the end of every 250^{th} MCMC iteration.

S10 Simulation 2 — Inference under Model Misspecification — Setup and Additional Results

S10.1 Simulation setup

We simulated an epidemic in a population of size N=400 with time–varying dynamics using Gillespie's direct algorithm over a four year period. Weekly prevalence counts were binomially distributed with detection probability $\rho = 0.95$. The epidemic dynamics varied over four epochs, based on the parameters given in Table S10. We fit SIR and SEIR models to the data, running three MCMC chains per model, discarding the first 100 iterations as burn–in, and sampling the paths of 150 subjects, chosen uniformly at random, per MCMC iteration. After discarding the burn–in, the resulting samples were combined to form the final sample. We also attempted to fit the models using PMMH. We ran three chains per model, each using 2,500 particles, the paths for which were simulated approximately via τ–leaping with a one day time step. The PMMH chains were plagued by severe particle degeneracy and did not converge.

Epoch 1: Weeks 0 – 26

Param.	True value
β	0.00025
γ	1/210
μ	1/150
ρ	0.95
$\mathbf{X}(t_0)$	$S_0 = 397, E_0 = 2, I_0 = 1, R_0 = 0$

Epoch 2: Weeks 26–105

β	0.0001
γ	1/210
μ	1/330
ρ	0.95
$\mathbf{X}(t_{26})$	$S_0 = 279, E_0 = 98, I_0 = 20, R_0 = 3$

Epoch 3: Weeks 105–167

β	0.00035
γ	1/90
μ	1/300
ρ	0.95
$\mathbf{X}(t_{105})$	$S_0 = 1, E_0 = 43, I_0 = 145, R_0 = 211$

Epoch 4: Weeks 167 – 209

β	0.0001
γ	1/180
μ	1/70
ρ	0.95
$\mathbf{X}(t_{167})$	$S_0 = 0, E_0 = 1, I_0 = 52, R_0 = 347$

Table S10: Parameter values governing the time–varying SEIR dynamics and binomial emissions process. The epidemic was simulated using Gillespie's direct algorithm and the process was restarted with the new parameter values at the beginning of each epoch.

S10.2 Additional results

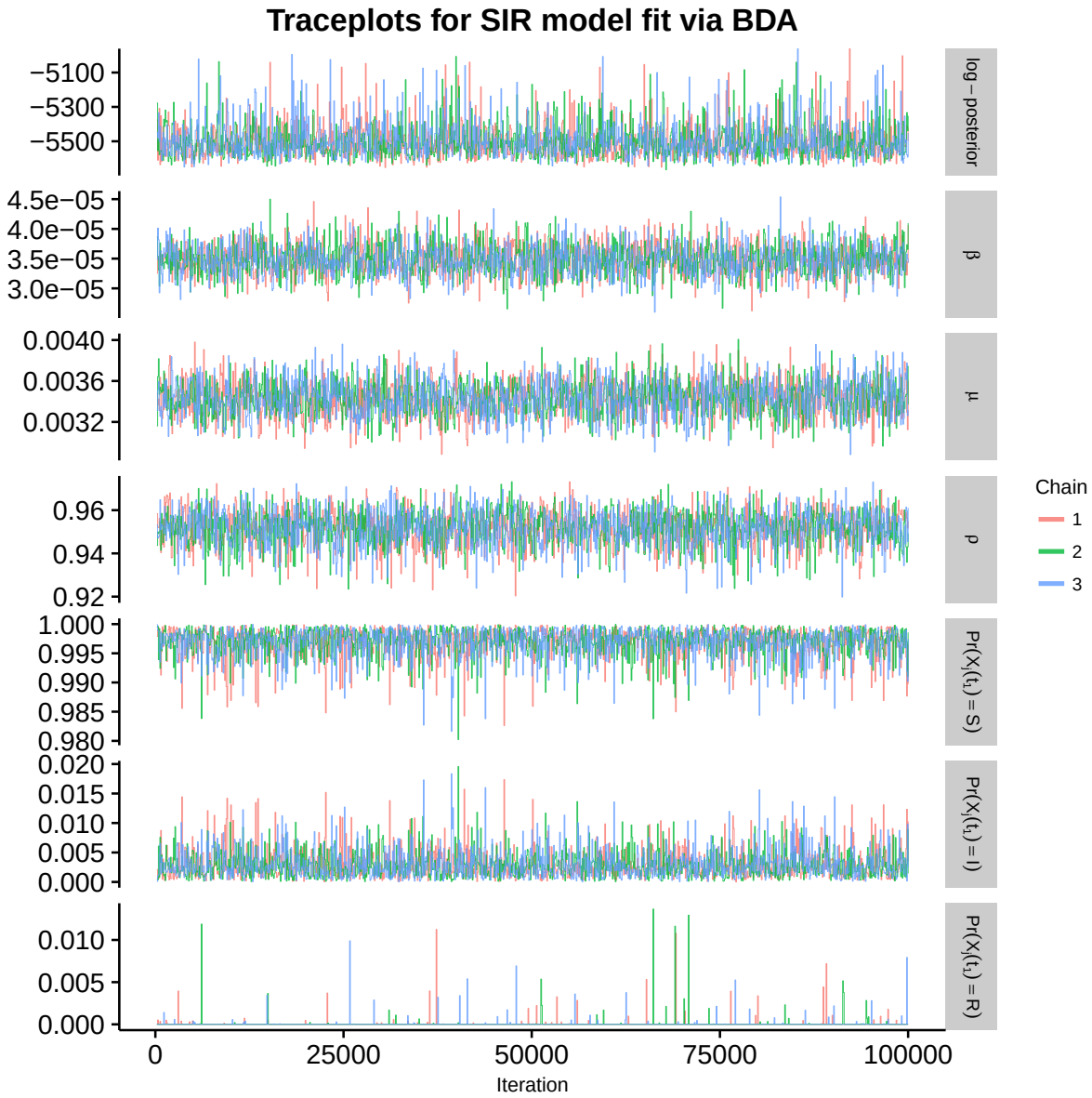


Figure S13: Traceplots of the log–posterior and model parameters for the SIR model fit using BDA following an initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

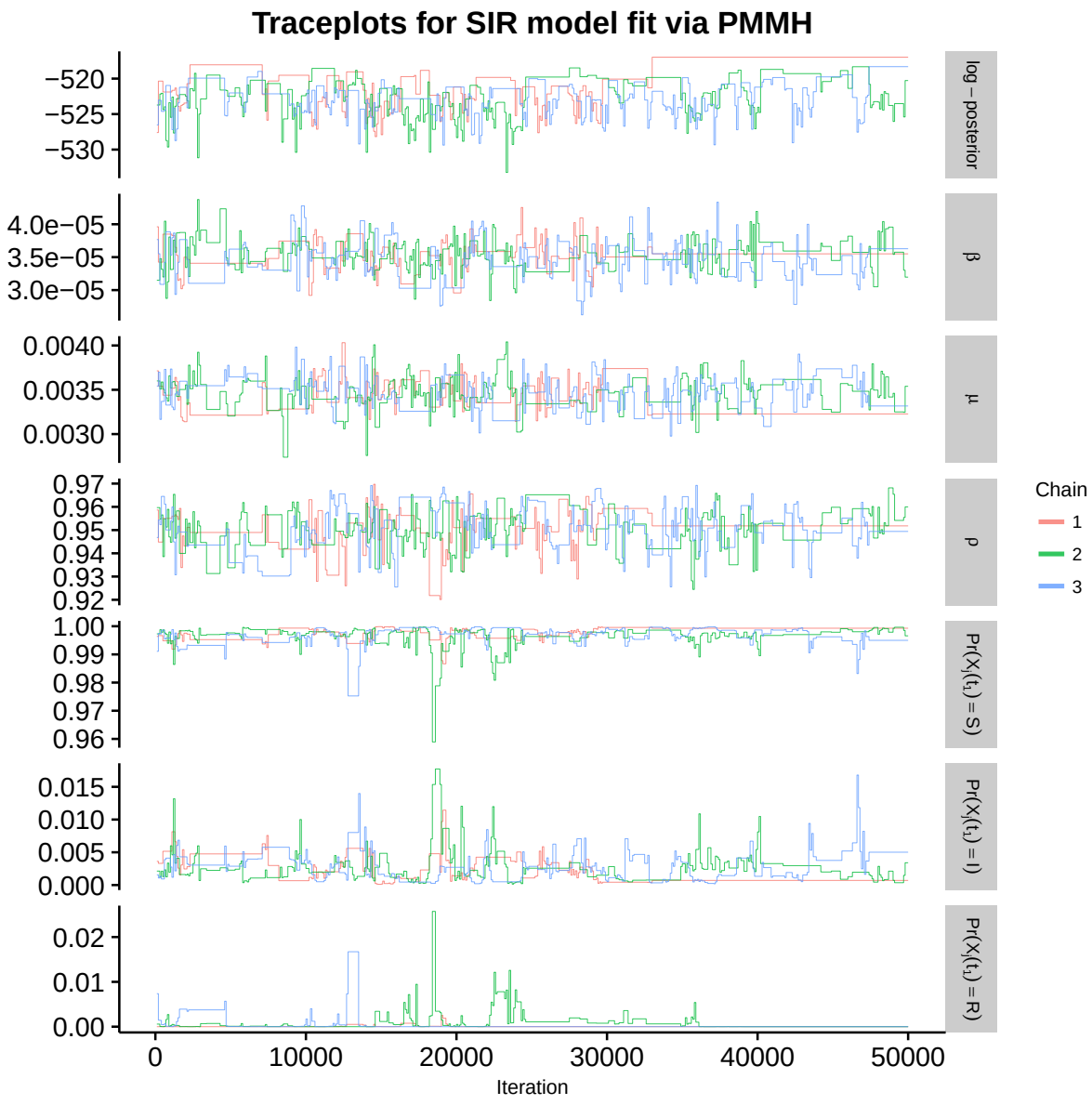


Figure S14: Traceplots of the log–posterior and model parameters for the SIR model fit using PMMH with 2,500 particles, following a tuning run of 5,000 iterations used to estimate the covariance matrix for the RWMH. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

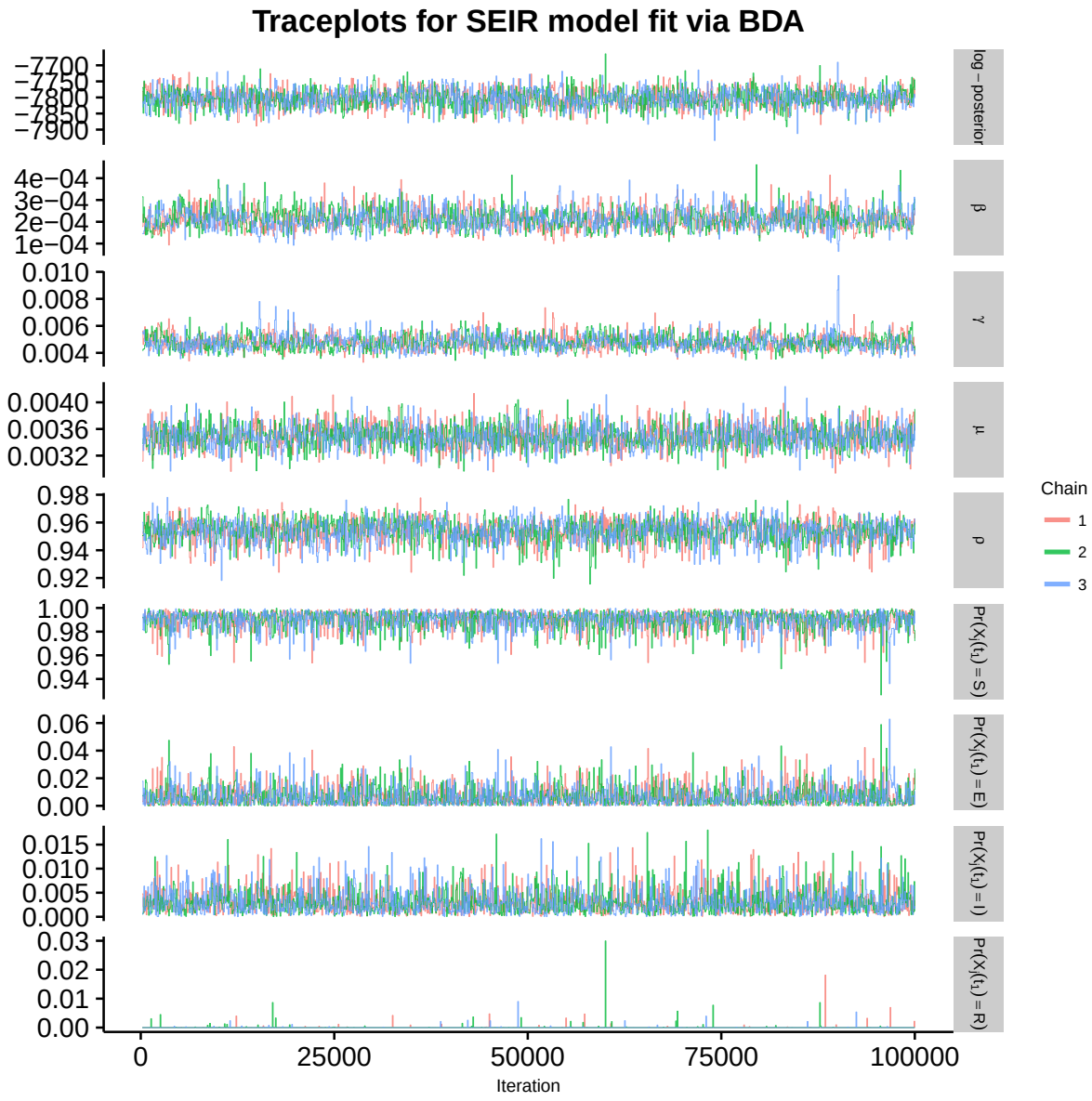


Figure S15: Traceplots of the log–posterior and model parameters for the SEIR model fit using BDA following an initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, γ is the rate at which an exposed individual becomes infectious, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

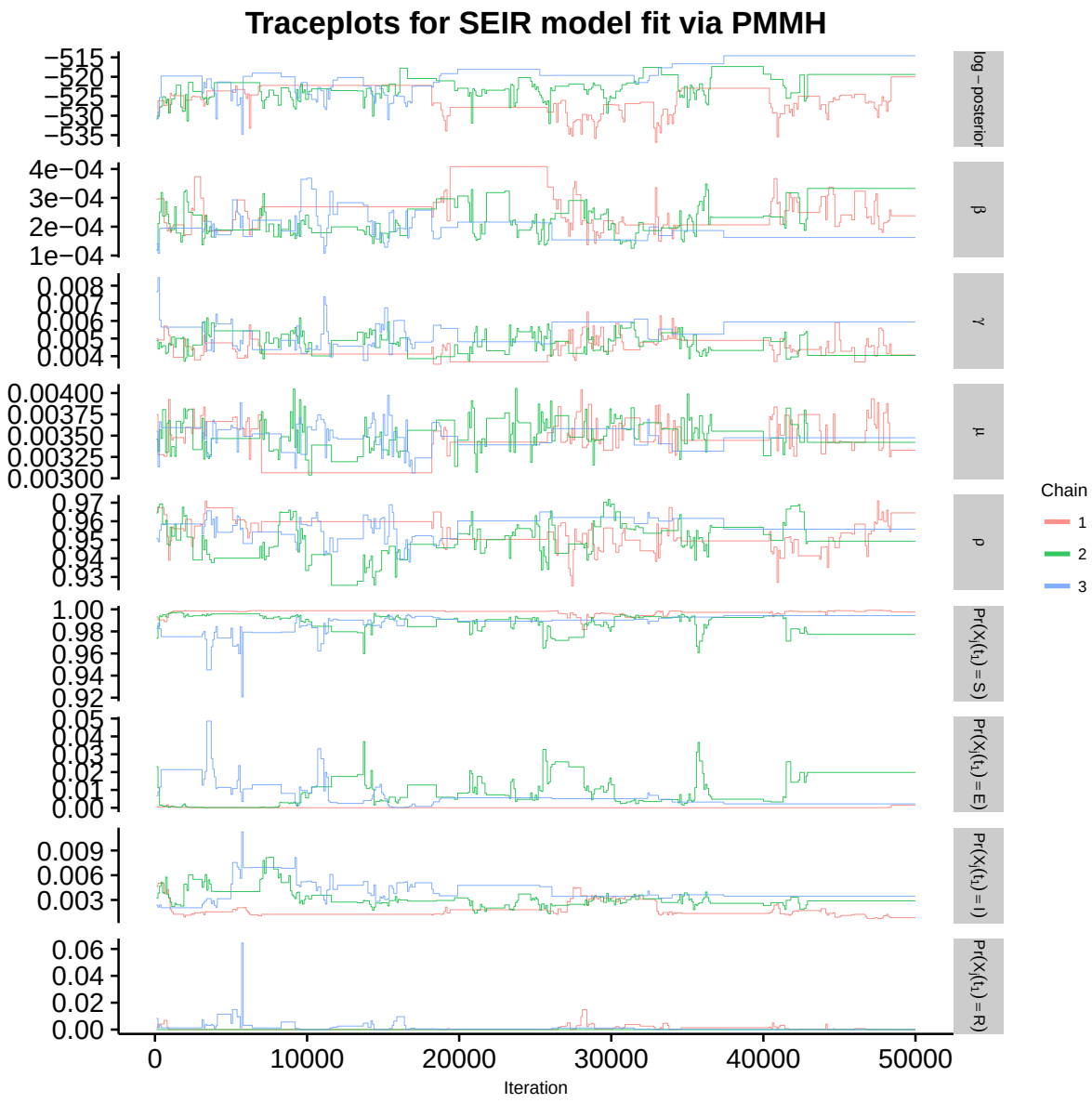


Figure S16: Traceplots of the log–posterior and model parameters for the SEIR model fit using PMMH with 2,500 particles, following a tuning run of 5,000 iterations used to estimate the covariance matrix for the RWMH. β denotes the per–contact infectivity rate, γ is the rate at which an exposed individual becomes infectious, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

SIR model

Parameter	Prior distribution
R_0	Beta′(0.6, 0.7, 1, 4)
β	Gamma(0.6, 10000)
μ	Gamma(0.7, 100)
$\mathbf{p}_{t_1}$	Dirichlet(90, 0.5, 0.01)
ρ	Beta(10, 1)

SEIR model

R_0	Beta′(0.6, 0.7, 1, 4)
β	Gamma(0.6, 10000)
γ	Gamma(0.5, 100)
μ	Gamma(0.7, 100)
$\mathbf{p}_{t_1}$	Dirichlet(90, 0.5, 0.5, 0.01)
ρ	Beta(10, 1)

Table S11: Prior distributions for the SIR and SEIR model and measurement process parameters for the models fit to the dataset simulated under time–varying SEIR dynamics. The prior for R_0 is the induced prior implied by β and μ. The per–contact infectivity rate is β, the rate at which an exposed individual becomes infectious is γ, the recovery rate is μ, the binomial sampling probability is ρ, and the initial state probabilities are $\mathbf{p}_{t_1}$.

SIR model

Parameter	Posterior median (95% Credible interval)
R_0	4.05 (3.40, 4.81)
β	0.000035 (0.000030, 0.000040)
μ	0.0034 (0.0031, 0.0038)
ρ	0.95 (0.93, 0.97)

SEIR model

Parameter	Posterior median (95% Credible interval)
R_0	23.80 (15.10, 36.98)
β	0.00021 (0.00013, 0.00032)
γ	0.0047 (0.0038, 0.0061)
μ	0.0035 (0.0032, 0.0038)
ρ	0.95 (0.94, 0.97)

Table S12: Posterior median estimates and 95% credible intervals for SIR and SEIR model parameters fit under a binomial emission distribution to the epidemic simulated with time–varying SEIR dynamics.

S11 Simulation 3 — Inference under Population Size Misspecification — Details

We simulated an outbreak under SIR dynamics, with $R_0 = \beta N/\mu = 3.5$, in a population of 1,250 individuals. Roughly 0.2% of the population was initially infected, and 95% were initially susceptible. The mean infectious period was $1/\mu = 7$ days. Prevalence was observed at weekly intervals, with detection probability $\rho = 0.3$, over a one year period.

We ran three chains for 100,000 iterations each under the following assumed population sized: 150, 300, 500, 900, 1100, 1200, 1250, 1300, 1400. We sampled the paths for 10% of the subjects, chosen uniformly at random, per MCMC iteration. We discarded the first 500 iterations from each chain as burn-in. Diffuse priors were specified for all model parameters, with the prior for the per–contact infectivity rate depending on the assumed population size (summarized in Table S13).

Param.	Prior distribution
R_0	Beta′(0.00042 $\times \frac{1250}{N}$, 0.35, 1, 2 / N)
β	Gamma(0.00042 $\times \frac{1250}{N}$, 1)
μ	Gamma(0.35, 2)
$\mathbf{p}_{t_1}$	Dirichlet(100, 1, 5)
ρ	Beta(1,1)

Table S13: Prior distributions for SIR model and measurement process parameters. The prior for R_0 is the induced prior implied by β and μ. The per–contact infectivity rate is β, the recovery rate is μ, the binomial sampling probability is ρ, and the initial state probabilities are $\mathbf{p}_{t_1}$. The prior for β was scaled in accordance with the assumed population size.

S12 Simulation 4 — Effect of Prior Specification on Inference — Setup and Additional Results

S12.1 Simulation details

We ran three MCMC chains for each of the SIR models fit under the prior regimes that are specified in Table S14 along with the true parameter values under which the data were simulated. Each chain was run for 100,000 MCMC iterations with 75 subject–paths per iteration. The first 100 iterations of each were discarded as burn–in, after which the samples from all three chains for each model were combined to form the posterior sample.

	Prior Distribution			
Parameter	Regime 1	Regime 2	Regime 3	Regime 4
$R_0 = 1.84$	Beta$'$(3, 3, 1, 1.526)	Beta$'$(0.3, 0.1, 1, 0.6)	Beta$'$(3, 3, 1, 1.526)	Beta$'$(0.3, 0.1, 1, 0.6)
$\beta = 0.00035$	Gamma(3, 10000)	Gamma(0.3, 1000)	Gamma(3, 10000)	Gamma(0.3, 1000)
$\mu = 0.14$	Gamma(3, 20)	Gamma(0.1, 0.8)	Gamma(3, 20)	Gamma(0.1, 0.8)
$\rho = 0.2$	Beta(21, 75)	Beta(21, 75)	Beta(1,1)	Beta(1,1)

Table S14: True parameter values and prior distributions under four different prior regimes. The prior for R_0 is the implied prior induced by the priors for β and μ. In regimes one and three, the central 80% of the prior mass for R_0 lay between 1.25 and 4.56, while in regimes two and four, 80% of the prior mass lay between 3.8×10^{-4} and 2.7×10^{4}. In regimes one and two, 80% of the prior mass for ρ lay between 0.17 and 0.27, while in regimes three and four the prior mass for ρ was uniformly distributed between 0 and 1. We used the same mildly informative Dirichlet(9, 0.2, 0.5) prior for $\mathbf{p}_{t_1}$ in all prior regimes.

S12.2 Convergence diagnostics

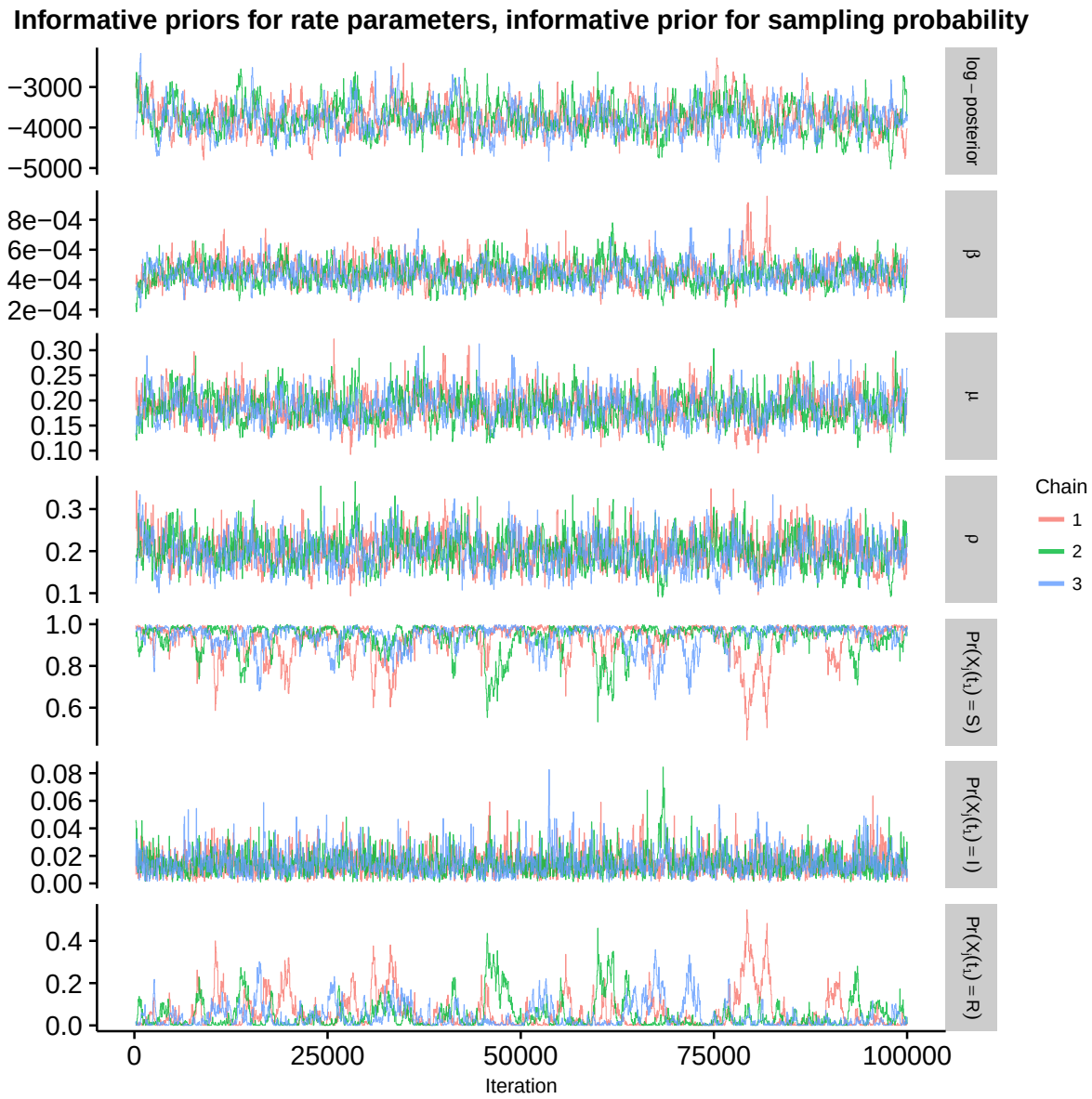


Figure S17: Traceplots of the log–posterior and model parameters for the SIR model fit under informative priors for all model parameters. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

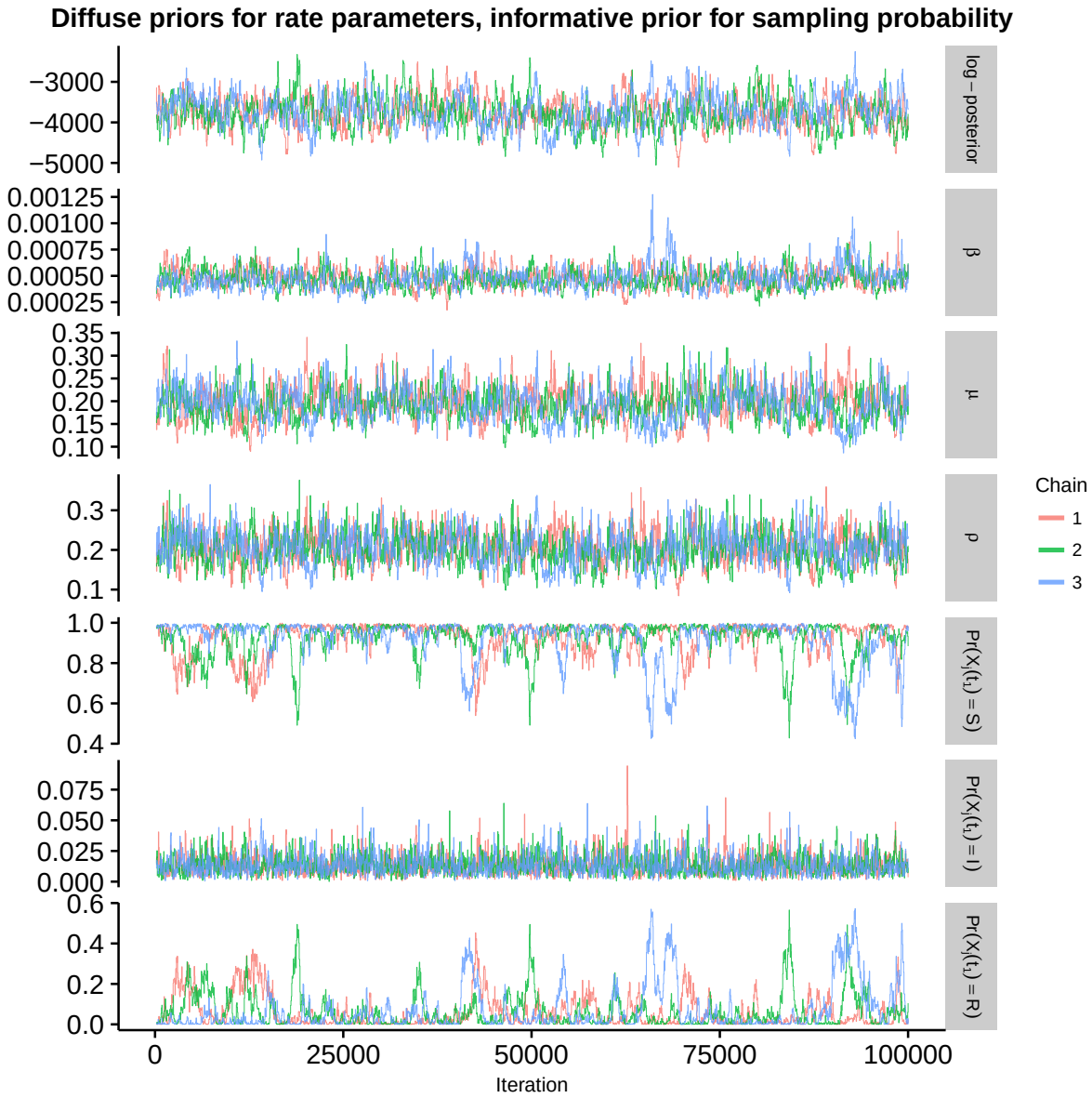


Figure S18: Traceplots of the log–posterior and model parameters for the SIR model fit under diffuse priors for the rate parameters and an informative prior for the binomial sampling probability. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

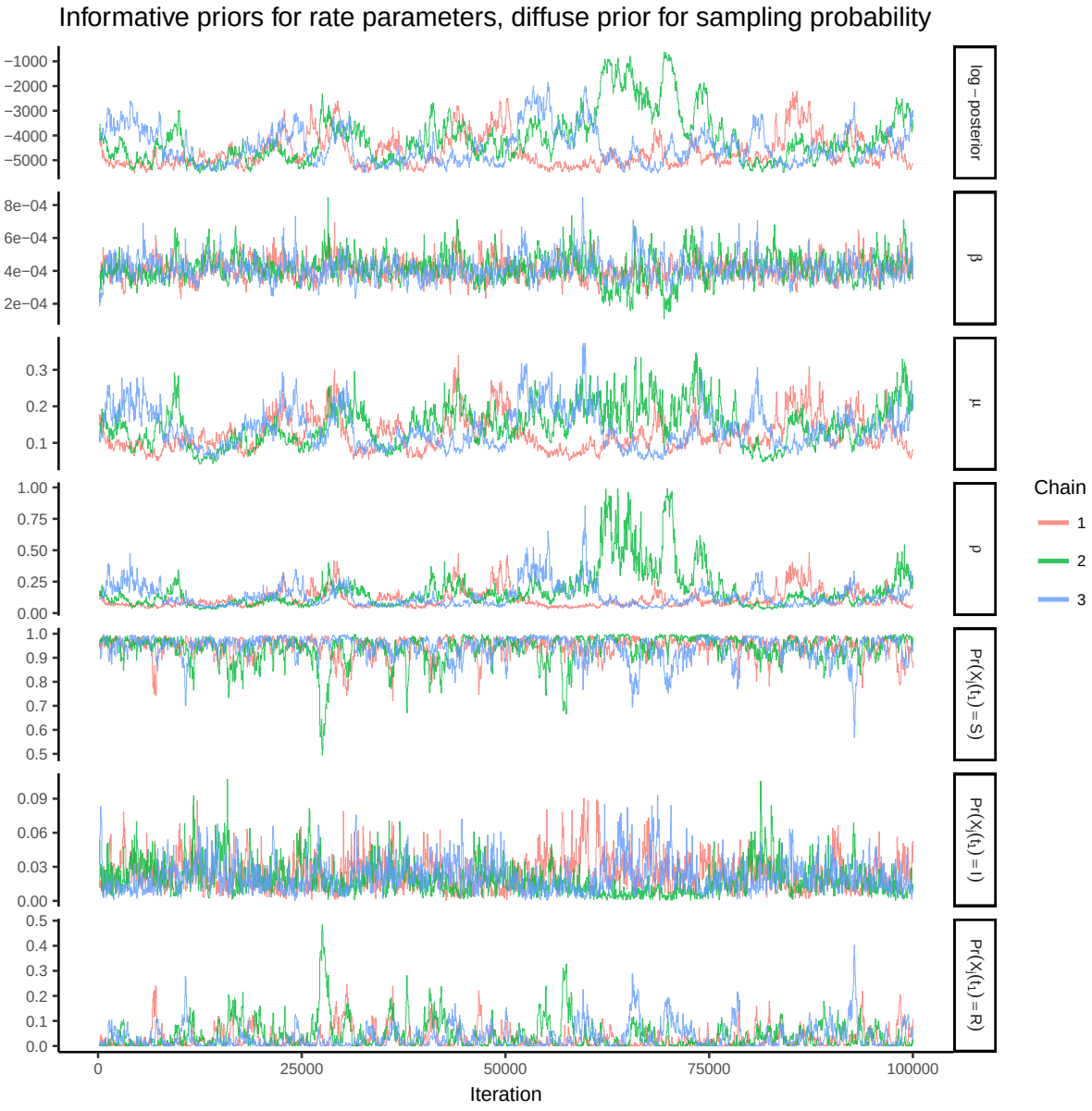


Figure S19: Traceplots of the log–posterior and model parameters for the SIR model fit under informative priors for the rate parameters and a diffuse prior for the binomial sampling probability. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

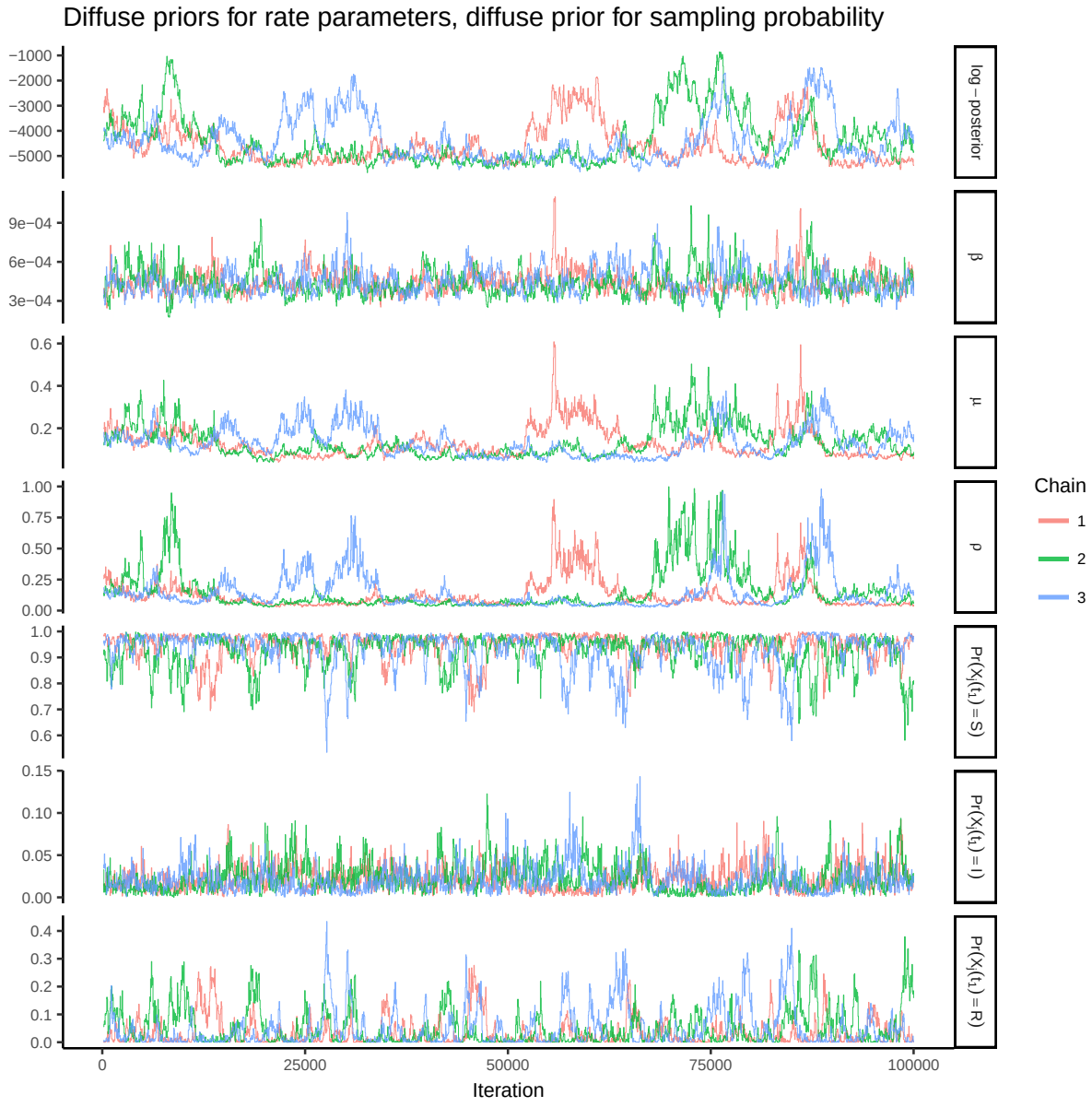


Figure S20: Traceplots of the log–posterior and model parameters for the SIR model fit under diffuse priors for all model parameters. β denotes the per–contact infectivity rate, μ is the recovery rate, ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

S13 Setup, additional results, and MCMC diagnostics for British boarding school example

We ran three MCMC chains per model to fit the SIR and SEIR models to the British boarding school dataset, for 100,000 iterations per chain. We sampled the paths for 100 subjects, chosen uniformly at random, per MCMC iteration, and discarded, as burn-in, the first 100 iterations of each chain for the SIR model, and the first 5,000 iterations of each chain for the SEIR model. Prior distributions, along with posterior medians and credible intervals are given in tables S15 and S16. The induced prior for R_0 is highly diffuse due to the diffuse prior on the per–contact infectivity rate. The prior distribution for the recovery rate and the rate at which exposed individuals became infectious reflected prior knowledge of the natural history of influenza. The prior for the detection probability has roughly 90% of its mass above 0.3, but is arguably quite diffuse given that it is known that over 90% of the boys were eventually infected.

We also fit the SIR and SEIR models using PMMH with paths for 5,000 particles simulated approximately via a multinomial modification of τ–leaping over two hour increments. The same priors were used as for the chains fit using BDA. Parameters were updated via random walk Metropolis–Hastings on transformed scales with a proposal covariance matrix that was estimated from an initial run of 2,000 MCMC iterations. We applied a log transformation to the rate parameters, a logit transformation to the binomial sampling probability, and a generalized logit transformation to the initial state probabilities. Results for PMMH are not reported since the MCMC never converged (see traceplots below).

Parameter	Prior Distribution	Posterior Median (95% BCI)
R_0	Beta′(0.001, 1, 1, 1526)	3.89 (3.40, 4.47)
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, 1.00)
$\Pr(X_j(t_1) = S)$		0.99 (0.98, 0.99)
$\Pr(X_j(t_1) = I)$	Dirichlet(900,3,9)	0.003 (0.001, 0.007)
$\Pr(X_j(t_1) = R)$		0.009 (0.004, 0.017)

Table S15: Prior distributions and posterior estimates for parameters of the SIR model with binomial emissions fit to the British boarding school outbreak data. The per–contact infectivity rate is β, the recovery rate is μ, and the binomial sampling probability is ρ. The prior for R_0 is the implied prior induced by the priors for β and μ. Effective sample size were β: 11,304; μ: 16,238; ρ: 3,920; $p_{S_{t_1}}$: 26,989; $p_{I_{t_1}}$: 284,431; $p_{R_{t_1}}$: 22,761.

S13.1 Boarding school example — MCMC diagnostics

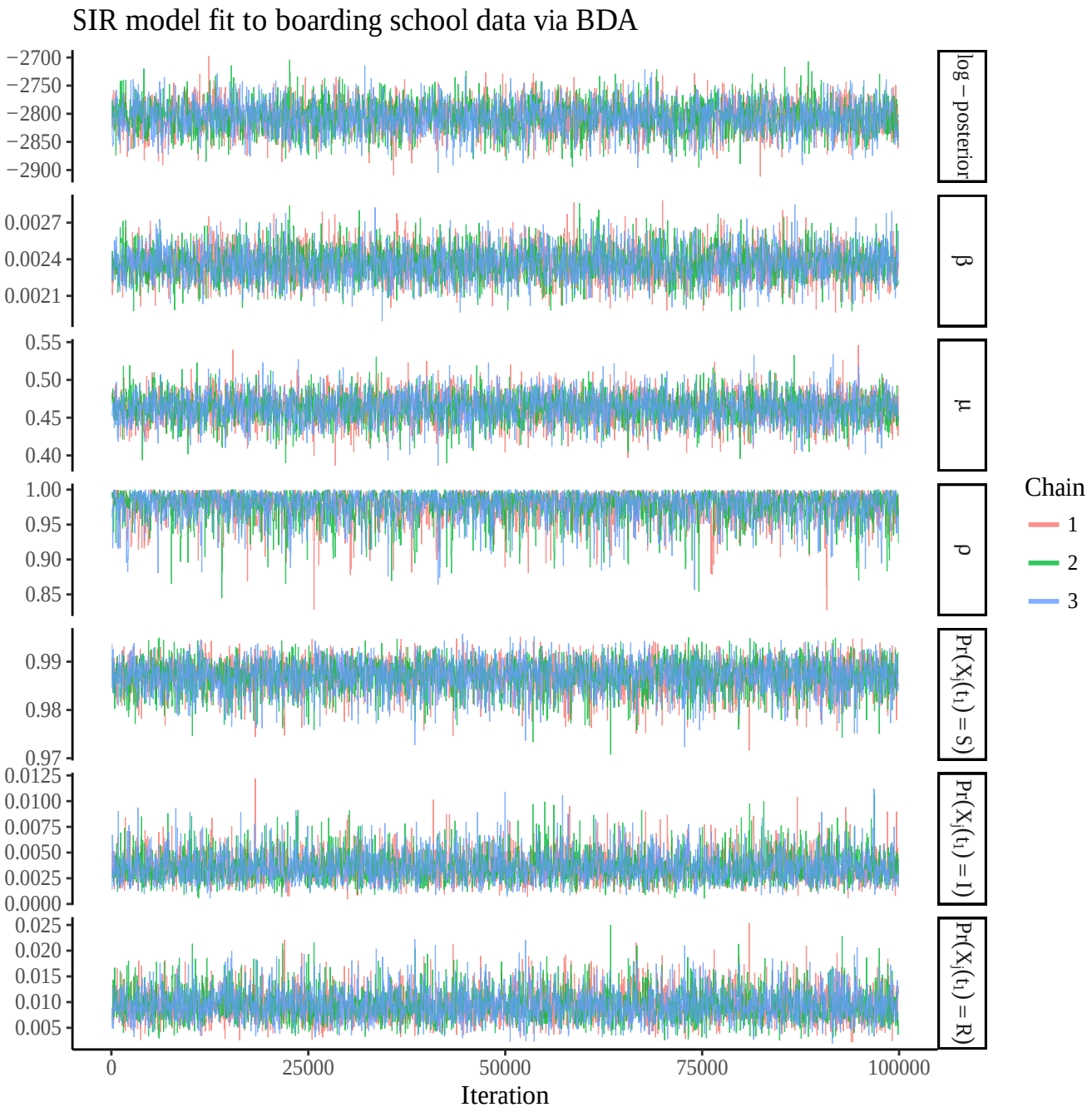


Figure S21: Traceplots of the log–posterior and model parameters for the SIR model fit under binomial emissions using BDA following an initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

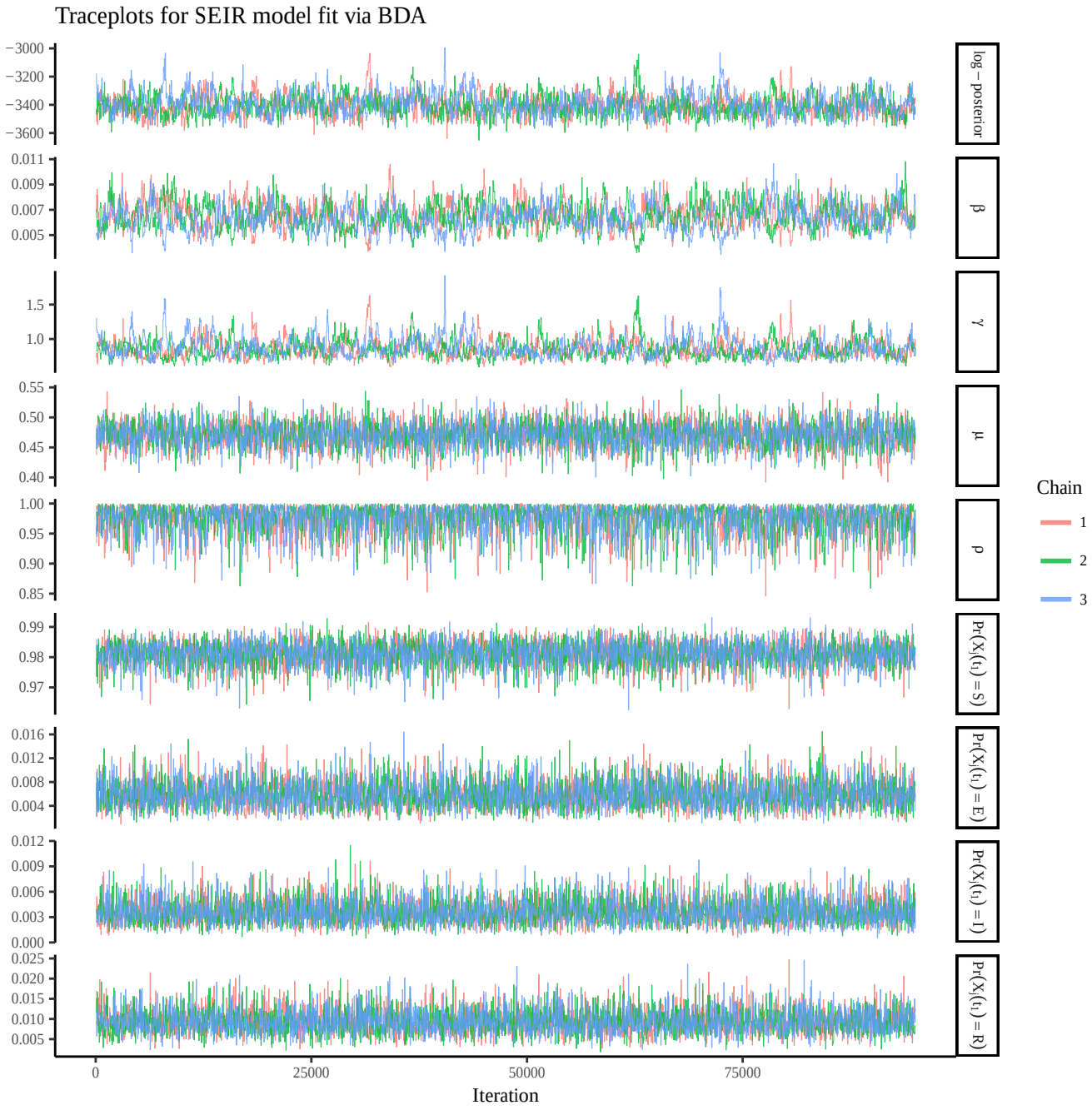


Figure S22: Traceplots of the log–posterior and model parameters for the SEIR model fit under binomial emissions via BDA following an initial burn–in of 5,000 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, γ is the rate at which immunity is lost, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

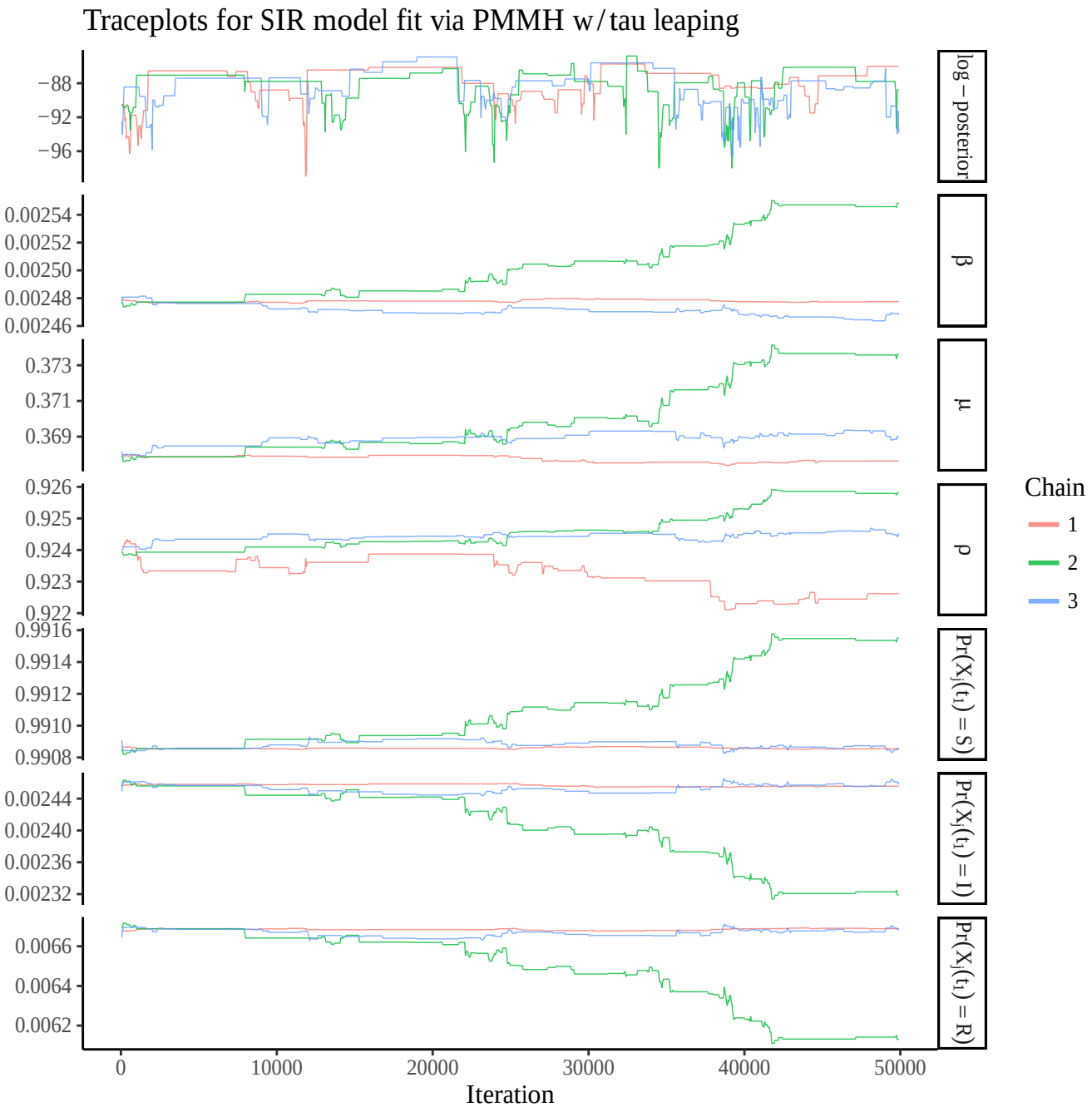


Figure S23: Traceplots of the log–posterior and model parameters for the SIR model fit under binomial emissions using PMMH with 5,000 particles per chain and a time–step of 2 hours in the approximate τ–leaping algorithm, following a tuning run of 2,000 iterations to estimate the RWMH covariance matrix and in initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

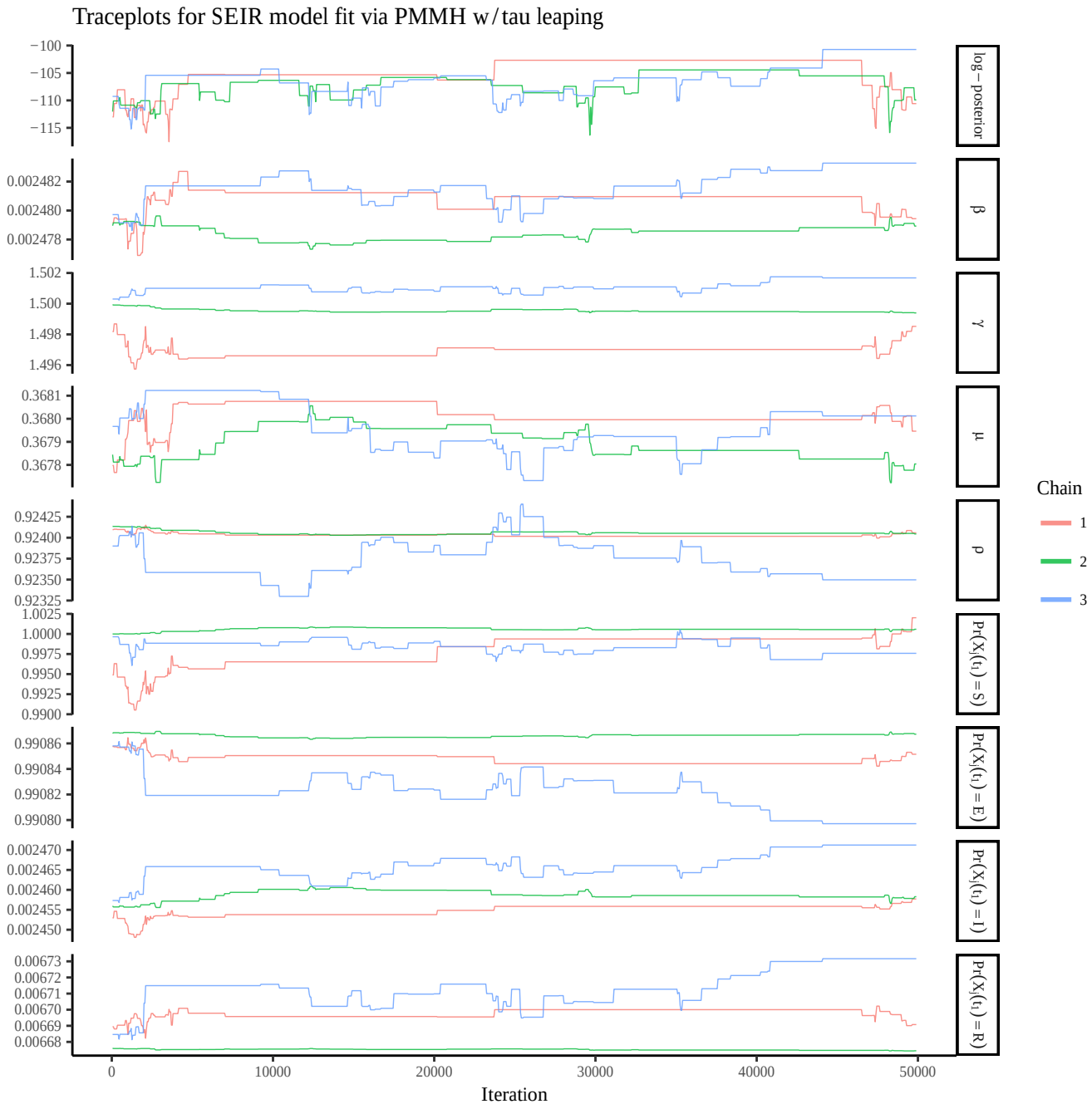


Figure S24: Traceplots of the log–posterior and model parameters for the SEIR model fit under binomial emissions using PMMH with 5,000 particles per chain and a time–step of 2 hours in the approximate τ–leaping algorithm, following a tuning run of 2,000 iterations to estimate the RWMH covariance matrix and in initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, γ is the rate at which immunity is lost, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

Parameter	Prior Distribution	Posterior Median (95% BCI)
R_0	Beta$'$(0.001, 1, 1, 1526)	3.89 (3.40, 4.47)
β	Gamma(0.001, 1)	0.0064 (0.0046, 0.0086)
γ	Gamma(0.001, 1)	0.84 (0.66, 1.19)
μ	Gamma(1,2)	0.47 (0.43, 0.51)
ρ	Beta(1,2)	0.98 (0.91, 1.00)
$\Pr(X_j(t_1) = S)$		0.98 (0.97, 0.99)
$\Pr(X_j(t_1) = E)$	Dirichlet(900, 6,3,9)	0.006 (0.002, 0.01)
$\Pr(X_j(t_1) = I)$		0.003 (0.001, 0.007)
$\Pr(X_j(t_1) = R)$		0.009 (0.004, 0.016)

Table S16: Prior distributions and posterior estimates for parameters of the SEIR model with binomial emissions fit to the British boarding school outbreak data. The per–contact infectivity rate is β, the rate at which an exposed individual becomes infectious is γ, the recovery rate is μ, and the binomial sampling probability is ρ. The prior for R_0 is the implied prior induced by the priors for β and μ. Effective sample size were β: 679; γ: 658; μ: 10,069; ρ: 3,244; $p_{S_{t_1}}$: 26,868; $p_{I_{t_1}}$: 26,168; $p_{R_{t_1}}$: 273,613.

Table S17:

S13.2 Supplementary analysis of the British boarding school example under negative binomial emissions

The PMMH MCMC runs in which SIR and SEIR models were fit to the boarding school data under a binomial emission distribution were plagued by severe particle degeneracy (Figures S23 and S24). The binomial emission distribution requires that the latent prevalence always be at least as great as the observed prevalence. However, this seemed to be a very stringent criterion with such a high case detection rate. That this criterion was so stringent is suggestive of non–trivial model misspecification. We attempted to confirm this by simulating a dataset that resembled the boarding school data. One possible data generating mechanism that yielded to similar prevalence counts resulted from an outbreak evolving under SEIR dynamics that varied over three epochs (Figure S25). That even a model with this simple set of time–varying dynamics would undoubtedly still be misspecified with respect to the real world circumstances in the boarding school is suggestive of a non–trivial level of model misspecification for both the simple SIR and SEIR models that we attempted to fit. Still, the inability of PMMH to fit simple, easily interpretable, SEMs to this data under binomial emissions is a severe limitation.

We fit an alternative set of SIR and SEIR models to the data using BDA and PMMH in which the observed prevalence was modeled as a negative binomial sample of the true prevalence, parameterized by its mean and overdispersion. This is a somewhat unrealistic emission distribution because it allows for the observed prevalence to be greater than the true prevalence. That this tended to occur more often in the later parts of the epidemic when boys were being discharged from the infirmary was particularly odd. However, the negative binomial emission distribution allows us to avoid degeneracy in the collection of PMMH particles by doing away with the constraint that the latent prevalence be no smaller than the observed prevalence. Parameters were assigned the same priors

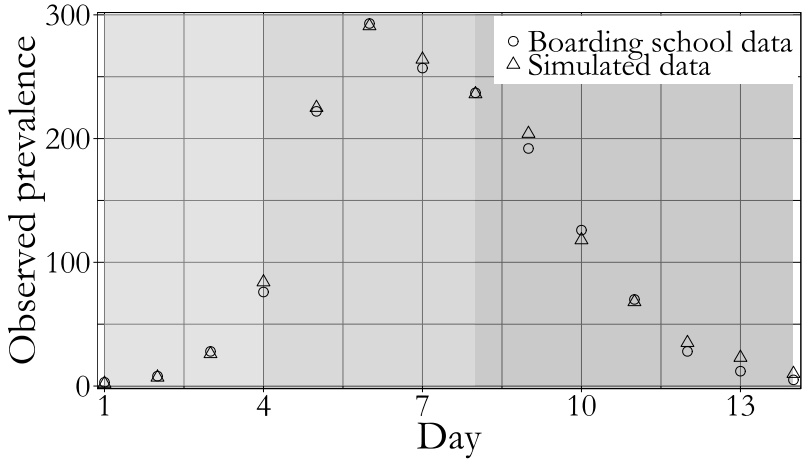


Figure S25: British boardings school data and data simulated under SEIR dynamics with time–varying dynamics over three epochs (indicated by different shaded regions). The simulated dataset was generated using the following parameters: in the first epoch (days 1-4), $\beta = 0.0035$, $\gamma = 1.25$, $\mu = 0.3$. In the second epoch (days 4-8), $\beta = 0.065$, $\gamma = 0.51$, $\mu = 0.41$. In the third epoch (days 8-14), $\beta = 0.06$, $\gamma = 2.5$, $\mu = 0.54$. The data were a binomial sample of the true prevalence with detection probability $\rho = 0.98$. There were three exposed individuals and two infected individuals at the beginning of day 1.

given in Tables S15 and S16, and the negative binomial overdispersion parameter, ϕ, was assigned a Gamma(1, 0.1) prior parameterized by rate. When fitting the model with BDA, we sampled new values for the rate parameters and initial state probabilities from their univariate full conditional distributions via Gibbs sampling. New values for the negative binomial sampling probability and the overdispersion parameter were sampled using multivariate random walk Metropolis–Hastings on the logit scale for ρ and on the log scale for ϕ. An empirical covariance matrix for the RWMH was estimated from an initial run of 10,000 iterations and scaled until the acceptance rate was between 15%–50%. We ran three chains per model for 100,000 iterations each, updating the paths of 100 subjects per MCMC iteration, and discarding the first 10,000 iterations as burn–in. We also ran three chains for 50,000 iterations each using PMMH for each of the models, with 500 particles per chain for the SIR model and 5,000 particles per chain for the SEIR model. Particle paths were simulated approximately using τ–leaping over a time step of 2 hours. Parameters were updated via multivariate RWMH whose covariance matrix was estimated from an initial tuning run of 2,000 iterations. Rate parameters and the overdispersion parameter were updated on the log scale, the negative binomial sampling probability was updated on the logit scale, and the initial state probabilities were updated on the generalized logit scale. We discarded the first 1,000 of each PMMH chain as burn–in.

Although the posterior median estimates under binomial and negative binomial emissions for the SIR and SEIR dynamics and detection rate are generally quite similar, the posterior credible intervals are considerably wider when the data modeled as a negative binomial sample of the true prevalence. This manifests both in the widths of the credible intervals for the latent process (Figure S26), and the credible intervals for the model parameters (Figure S27). This is not unexpected given that the negative binomial distribution is substantially more flexible than the binomial distribution. In comparing the posterior estimates obtained using BDA and PMMH under negative binomial emissions, we find that the estimates are essentially identical for the SIR model. For the SEIR

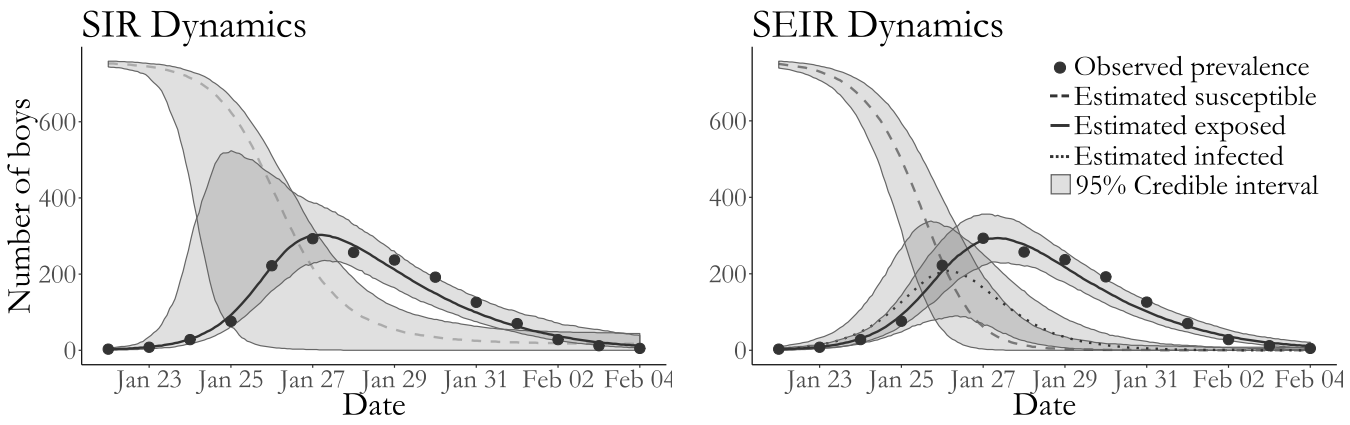


Figure S26: Boarding school data, pointwise posterior median estimates and pointwise 95% credible intervals under negative binomial emissions (grey shaded areas) for the numbers of infected boys (solid line) and susceptible boys (dashed line). Posterior estimates based on a thinned sample, with every 250^{th} configuration retained.

model, estimates of the dynamics are generally similar, though not to the same degree as those for the SIR model. We notice that the credible intervals for the mean infectious period and the negative binomial detection probability obtained using PMMH are substantially wider than those obtained using BDA. Upon closer inspection of the traceplots of the model parameters, it is clear that the negative binomial overdispersion parameter in the PMMH chains did not converge (Figure S29).

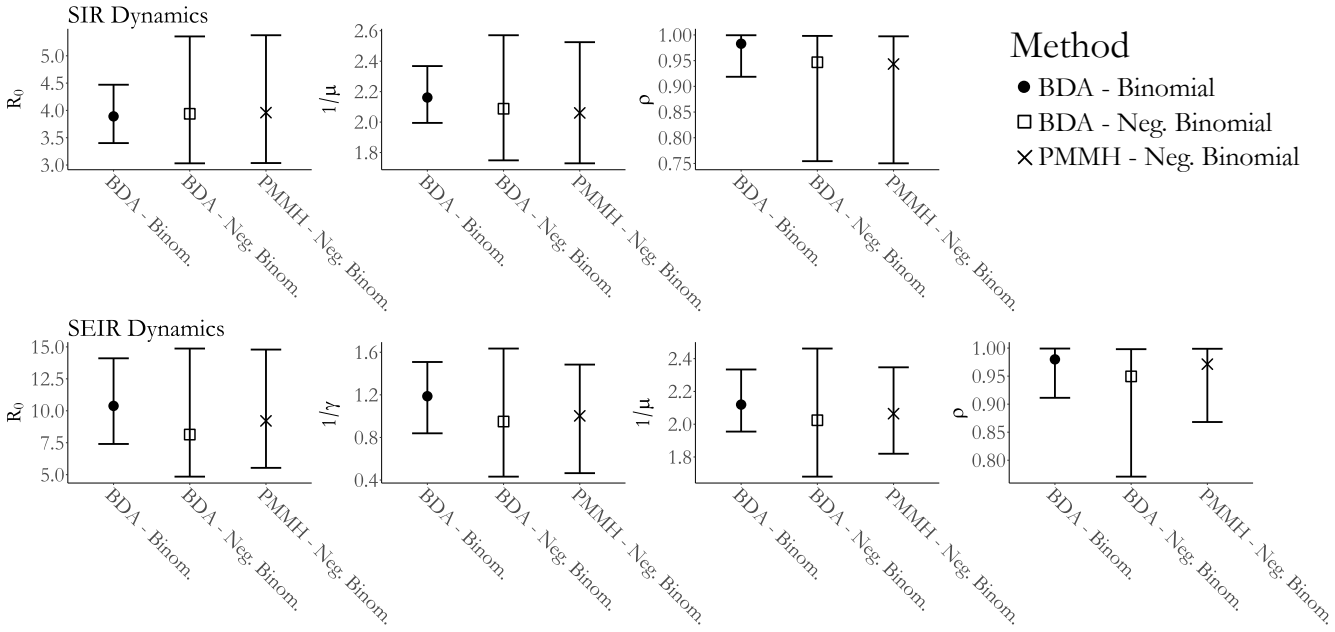


Figure S27: Posterior medians and 95% credible intervals for SIR and SEIR models fit with BDA and PMMH to the British boarding school data under binomial and negative binomial emission distributions.

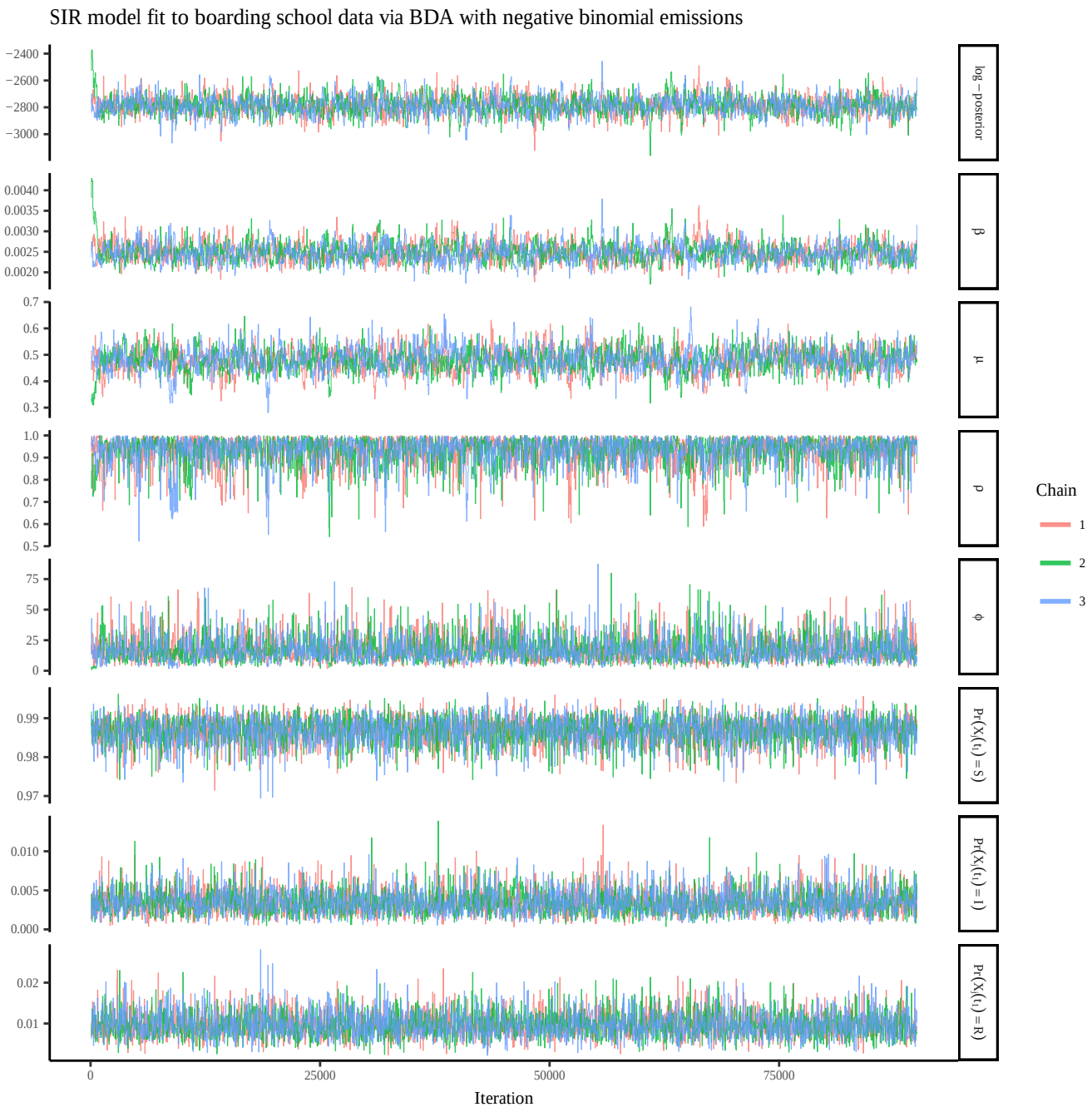


Figure S28: Traceplots of the log–posterior and model parameters for the SIR model fit under negative binomial emissions using BDA following an initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

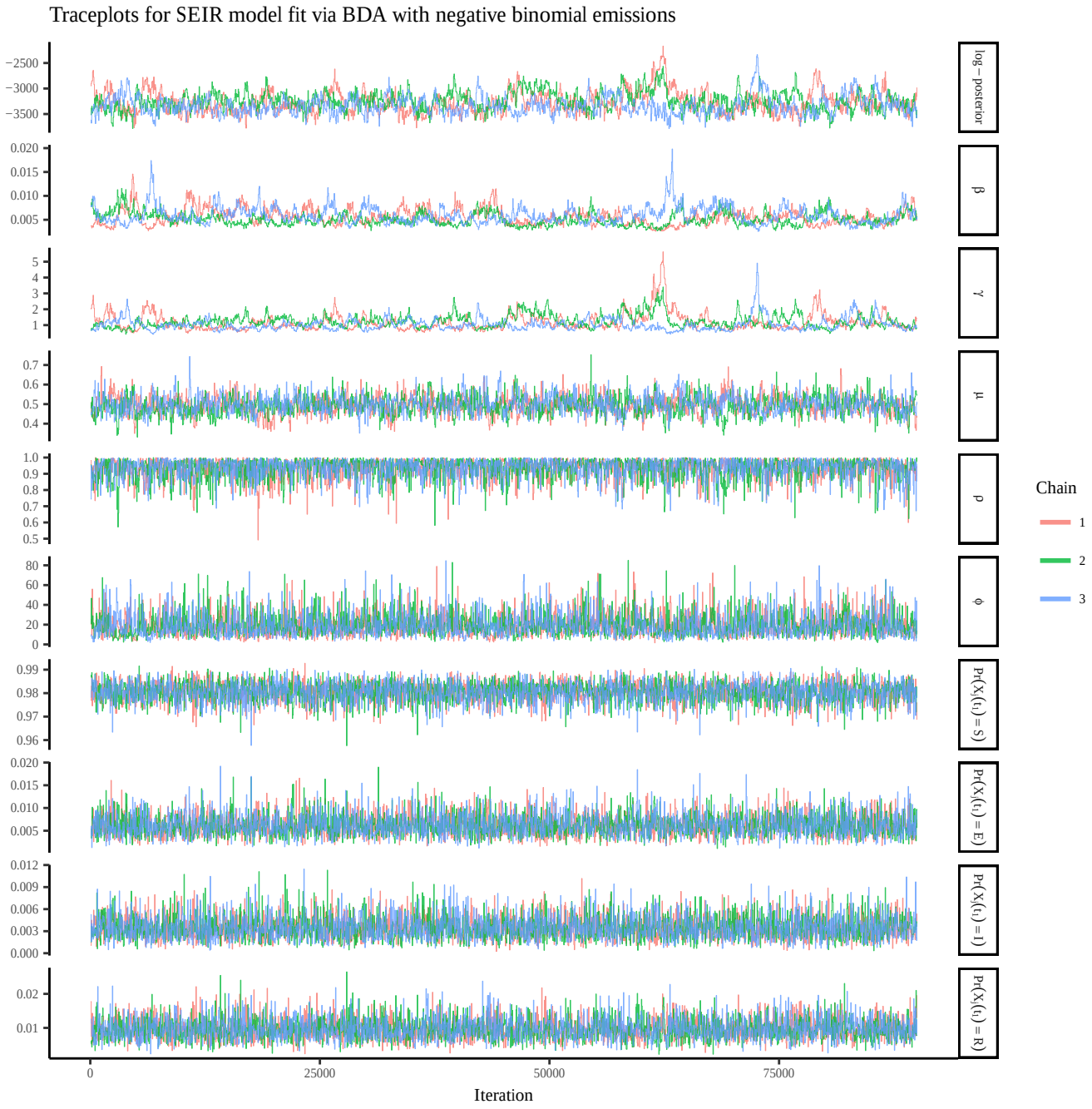


Figure S29: Traceplots of the log–posterior and model parameters for the SEIR model fit under negative binomial emissions via BDA following an initial burn–in of 5,000 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, γ is the rate at which immunity is lost, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

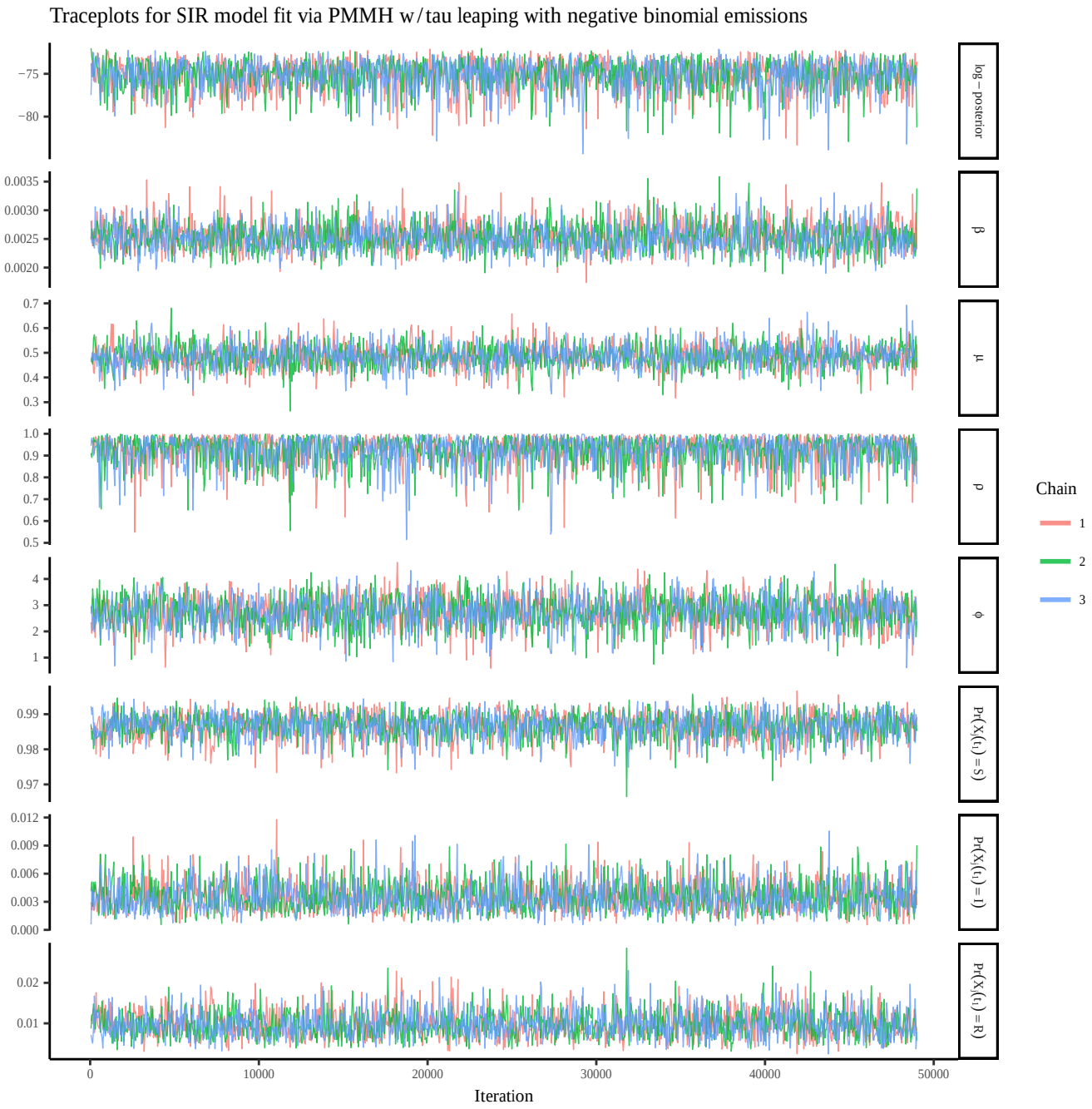


Figure S30: Traceplots of the log–posterior and model parameters for the SIR model fit under negative binomial emissions using PMMH with 5,000 particles per chain and a time–step of 2 hours in the approximate τ–leaping algorithm, following a tuning run of 2,000 iterations to estimate the RWMH covariance matrix and in initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.

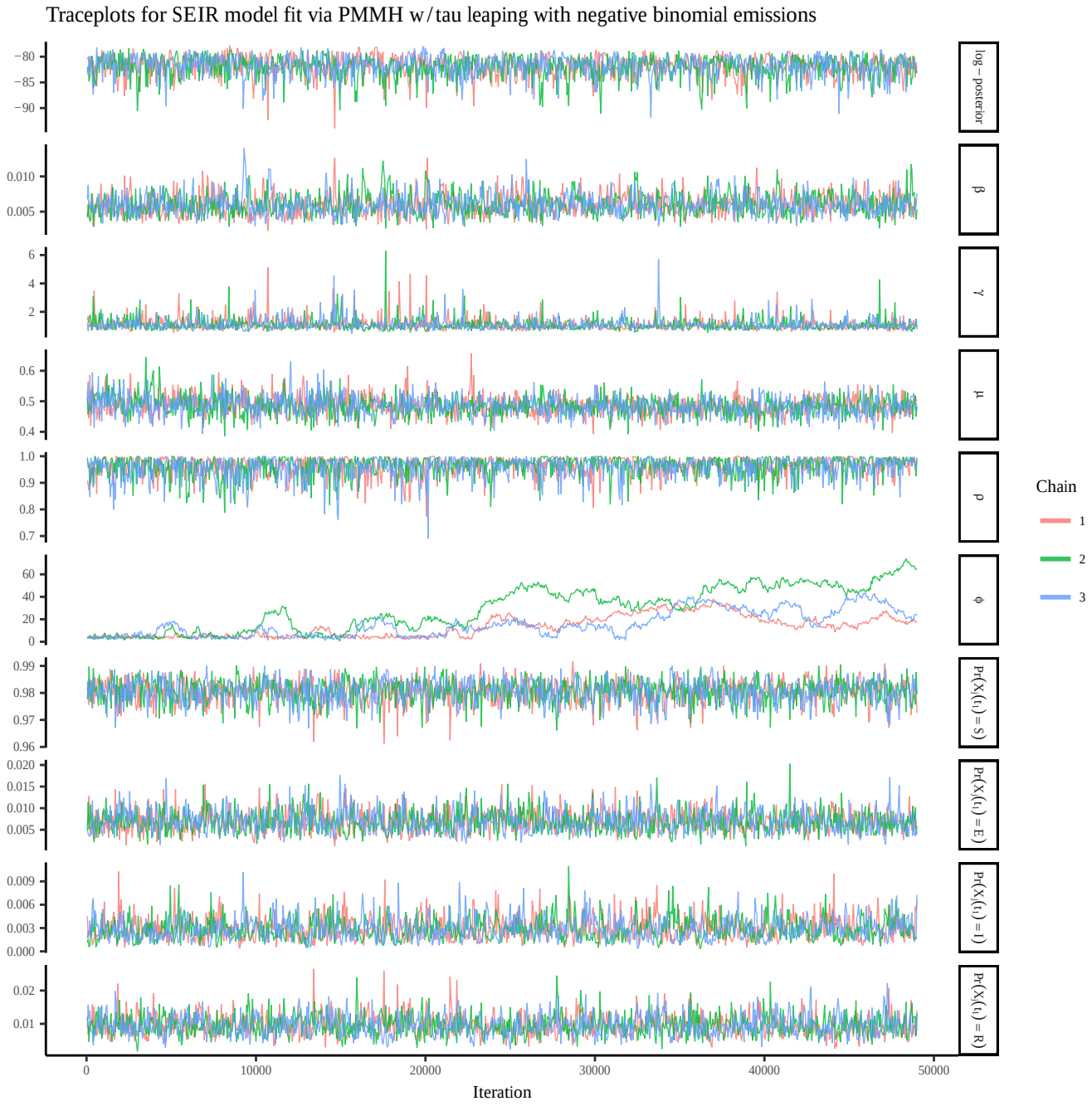


Figure S31: Traceplots of the log–posterior and model parameters for the SEIR model fit under negative binomial emissions using PMMH with 5,000 particles per chain and a time–step of 2 hours in the approximate τ–leaping algorithm, following a tuning run of 2,000 iterations to estimate the RWMH covariance matrix and in initial burn–in of 100 iterations. β denotes the per–contact infectivity rate, μ is the recovery rate, γ is the rate at which immunity is lost, and ρ is the binomial sampling probability. Traceplots are thinned to display every 50^{th} iteration.