Uniformly Ergodic Data-Augmented MCMC for Fitting Stochastic Epidemic Models to Incidence Data

Raphaël Morsomme[1*] and Jason Xu[1]

[1]Department of Statistical Science, Duke University

October 12, 2021

Abstract

Stochastic epidemic models provide an interpretable probabilistic description of the spread of a disease through a population. Yet, fitting these models in missing data settings is a notably difficult task; in particular, when the epidemic process is only partially observed, the likelihood of many classic models becomes intractable. To remedy this issue, this article introduces a data-augmented MCMC algorithm for fast and exact inference under the stochastic SIR model given discretely observed incidence counts. In a Metropolis-Hastings step, the algorithm jointly proposes event times of the augmented data according to a stochastic process whose dynamics closely resemble those of the SIR process, and from which we can efficiently generate an epidemic that is compatible with the observed data. Not only is the algorithm fast, but since the augmented data are generated from a very faithful approximation of the target epidemic model, the sampler can update large portions of the augmented data per iteration without lowering the acceptance rate prohibitively. We find that the method explores the high-dimensional latent space efficiently, and show that the Markov chain underlying the algorithm is uniformly ergodic. The proposed algorithm enables exact inference and scales to outbreaks in populations of hundreds of thousands, even on a single laptop. We validate its performance via thorough simulation experiments, consider and a case study on the 2014 Ebola outbreak in Western Africa.

Keywords— Compartmental model; incidence data; exact posterior inference; data-augmented MCMC; likelihood-based methods; uniform ergodicity

*raphael.morsomme@duke.edu

1 Introduction

The efficient control of a disease outbreak requires an understanding of the mechanisms underlying its spread. Mechanistic compartmental models, which describe the transition of individuals between various states, have a long mathematical modeling tradition in epidemiology [15]. Due to their interpretability, they are commonly used to describe the dynamics of an outbreak and typically serve as the main source of information for predicting the course of an outbreak and identifying interventions that could be effective. Originally, deterministic versions of the models were employed by mathematicians and epidemiologists. These models are simple to analyze, but fail to capture the inherent randomness in the spread of a disease. For instance, they cannot be used to estimate the probability of a large-scale outbreak or its expected duration, and do not allow for uncertainty quantification when used within inferential procedures. Stochastic epidemic models (SEM), on the other hand, incorporate the random nature of infections and recoveries and therefore provide more realistic descriptions of the spread of a disease, and in turn more reliable inference from observed data.

Conducting inference on SEMs is, however, a notably difficult task. Challenges stem from the fact that the observed data typically provide incomplete information on a process that evolves continuously through time, making the likelihood of the model intractable. Were the infection and recovery times of each individual of the population observed, conducting inference on SEMs would be straightforward, but this is rarely the case. In practice, one often has access to either incidence data such as weekly counts of new infections, or prevalence data such as the numbers of people infectious at discrete reporting times. For instance, the 2013–2016 outbreak of Ebola virus reported over $10,000$, mainly in three Western African countries (Guinea, Liberia and Sierra Leone). This outbreak was characterized by the large number of infections that occurred after individuals were tested positive for the virus, giving rise to noisy incidence data. Numerous infections happened at the hospitals where infected individuals received a treatment as well as at the funerals of people deceased from the virus[1]. It therefore seems appropriate to model a positive test for the virus as an indication that the individual became infectious some time prior to the test rather than as a removal time to do: stronger transition tying back to goal of partial observed inference.

More specifically, the marginal likelihood of such partially observed data becomes a bottleneck, as it requires a large integration step that accounts for all possible configurations of the missing. In particular, direct computation of this likelihood requires the transition probabilities between observation times for which no closed form is available. to do: cite recent papers from Simon Spencer's group too that do so for discrete-time models, and mention classical matrix exponentiation being intractable. Numerical methods to obtain transition probabilities have only recently been developed in [12, 11], but the high computational costs limit their algorithm to moderate sized outbreaks.

[1] In most regions impacted by the Ebola virus, touching the body of the deceased person during funerals is a tradition. Since the Ebola virus is transmitted via bodily fluids such as sweat, numerous infections occurred at funerals.

In this article, we make use of data augmentation to explore configurations of the missing data through latent variables. We show how a data-augmented MCMC algorithm can target the exact joint posterior of model parameters and these latent variables under the stochastic SIR model, a commonly used SEM, from discretely observed infection incidence counts. Rather than marginalizing directly, our approach accounts for the latent space via sampling. The algorithm updates the event times of the augmented data by jointly proposing infection and removal times from a stochastic process whose dynamics closely resemble those of the SIR process, and from which we can efficiently generate an epidemic that is compatible with the observed data. Its success lies in the design of this surrogate process serving as an efficient proposal scheme within a Metropolis-Hastings algorithm. Since the latent spaces are generated from a very faithful approximation of the target epidemic model, the algorithm can update a large portion of the event times per iteration while maintaining a relatively high acceptance rate, thereby exploring the high-dimensional latent space efficiently.

The remainder of the article is structured as follows. Section 2 provides background information on the inference task. It introduces the stochastic SIR model and explains why conducting inference under partially observed data is a difficult task. Previous works addressing this problem are also presented. Section 3.1 describes the data-augmented MCMC algorithm proposed in this article and assesses how closely it approximation the SIR process. Section 4.1 presents the results of a simulation study examining the performance of the algorithm, while Section 4.2 describes an analysis of the 2014 Ebola outbreak in Guéckédou, Guinea. Finally, Section 5 discusses the findings and concludes the article.

2 Background and Prior Work

2.1 The Stochastic SIR Model

Our point of departure is to consider the stochastic SIR model – also referred to as the general epidemic model – which offers a parsimonious and interpretable representation of the mechanistic, population-level dynamics of an epidemic. The stochastic SIR model is a compartmental model in which the individuals of a population transition through three types or compartments: susceptible (S), infectious (I) and removed (R). The only possible moves are from S to I (infections) and from I to R (removals). A susceptible individual becomes infected through contact with an infectious individual. Once infected, she is immediately infectious and remains so for some period of time after which she is removed from the process without the possibility of reinfection. In this formulation, demographic dynamics such as births and deaths of individuals are ignored since they usually occur at a much slower rate than infections and removals.

Assuming a closed population with a fixed size n, the stochastic SIR model consists of a continuous-time vector-valued process

$$\mathbf{X} = \{\mathbf{X}(t), t > 0\} \in \chi_{\mathbf{X}} \tag{1}$$

with

$$\mathbf{X}(t) = (X_1(t), \ldots, X_n(t)) \in \{s, i, r\}^n \tag{2}$$

where the agent-level subprocess

$$X_j(t) = \begin{cases} s, & t \in [0, \tau_j^I] \\ i, & t \in (\tau_j^I, \tau_j^R] \\ r, & t \in (\tau_j^R, \infty) \end{cases}, \quad i = 1, \ldots, n$$

denotes the status (compartment) of individual j at time t with τ_j^I and τ_j^R respectively the infection and removal times of individual j. If individual j never becomes infectious, then we set $\tau_j^I = \tau_j^R = \infty$ and $X_j(t) = s$ for $t \in [0, \infty)$. Here $\chi_{\mathbf{X}}$ denotes the set of valid trajectories compatible with the evolution of a disease—that is, the set of trajectories in which no infection occurs whenever the infectious compartment is depleted :

$$\chi_{\mathbf{X}} = \{X : X(t) \in \{s, r\}^n \Rightarrow X(t+u) = X(t), \forall u > 0\}. \tag{3}$$

The stochastic SIR model is specified by the rates at which individuals move from one compartment to another. If we assume a homogeneously mixing population where contacts between individuals occur independently at some rate constant β, then the contacts between two given individuals are said to follow a Poisson process with rate β. This parameter can be interpreted as the *infection rate*: when a susceptible individual comes into contact with an infectious individual, she immediately becomes infectious. If we also make the common assumption that the infectious periods follow independent exponential distributions with rate γ, then the process (1) is a time–homogeneous continuous time Markov chain, whose instantaneous transition rates are given by the matrix $\Lambda = [\lambda_{\mathbf{x},\mathbf{x}'}]$ with

$$\lambda_{\mathbf{x},\mathbf{x}'} = \begin{cases} \beta I(t), & \mathbf{x} \text{ and } \mathbf{x}' \text{ only differ at position } j \text{ with } x_j = s \text{ and } x'_j = i, \\ \gamma, & \mathbf{x} \text{ and } \mathbf{x}' \text{ only differ at position } j \text{ with } x_j = i \text{ and } x'_j = r, \\ 0, & \text{otherwise.} \end{cases}$$

where $I(t) = \#\{x_j(t) = i\}$ is the total number of infectious individuals at time t. Thus, the individual-level infection and removal rates at time t are respectively $\beta I(t)$ and γ.

2.2 Inference with Complete Data

When the Markov process (1) is completely observed until time t_{end}, we obtain the following likelihood

$$\begin{aligned} L(\theta; \mathbf{X}) &= \prod_{j \in \mathcal{I}} \beta I(\tau_j^I) \prod_{k \in \mathcal{R}} \gamma \exp\left\{ - \int_0^{t_{end}} \beta I(t) S(t) + \gamma I(t) dt \right\} \\ &= \beta^{n_I} \gamma^{n_R} \prod_{j \in \mathcal{I}} I(\tau_j^I) \exp\left\{ - \int_0^{t_{end}} \beta I(t) S(t) + \gamma I(t) dt \right\} \end{aligned} \tag{4}$$

describing the complete data trajectory. To establish notation, $\theta = (\beta, \gamma) \in \chi_\theta$ are the model parameters and $\chi_\theta = (0, \infty) \times (0, \infty)$ denotes the parameter space, the index sets $\mathcal{I} = \{j \in (1, \cdots, n) : \tau_j^I \in (0, t_{end}]\}$ and $\mathcal{R} = \{j \in (1, \cdots, n) : \tau_j^R \in (0, t_{end}]\}$ denote the individuals that are respectively infected and removed during the observation interval $(0, t_{end}]$, $n_I = |\mathcal{I}|$ and $n_R = |\mathcal{R}|$ are the numbers of observed infections and removals, and $S(t) = \#\{X_i(t) = S\}$ is the number of susceptible individuals at time t.

It is straightforward to conduct inference in this continuously observed scenario. In particular, the likelihood (4) belongs to the exponential family, and maximum likelihood estimates of its parameters can be expressed in terms of the sufficient statistics defined above as

$$\hat{\beta}_{MLE} = \frac{n_I}{\int_0^{t_{end}} I(t)S(t)dt}, \quad \hat{\gamma}_{MLE} = \frac{n_R}{\int_0^{t_{end}} I(t)dt}.$$

Since the functions I and S are constant between event times, the two integrals correspond to finite sums which are straightforward to compute. Furthermore, in a Bayesian context, inference is facilitated by the conjugacy with the Gamma distribution: if we let the parameters follow independent Gamma prior distributions

$$\beta \sim Ga(a_\beta, b_\beta), \qquad \gamma \sim Ga(a_\gamma, b_\gamma) \tag{5}$$

where $Ga(a, b)$ denotes the parametrization with mean a/b and variance a/b^2, then the posterior distributions of the parameters are independent Gamma distributions

$$\beta|\mathbf{X} \sim Ga\left(a_\beta + n_I, b_\beta + \int_0^{t_{end}} I(t)S(t)dt\right) \tag{6}$$

and

$$\gamma|\mathbf{X} \sim Ga\left(a_\gamma + n_R, b_\gamma + \int_0^{t_{end}} I(t)dt\right) \tag{7}$$

from which one can easily generate independent values with a Monte Carlo sampler to explore the posterior distribution of parameters $\pi(\theta|\mathbf{X})$.

2.3 Inference with Incomplete Data

In practice, inference is complicated by the fact that the process (1) is only partially observed. When we do not observe all epidemic events, which is virtually always the case with real data, we do not have access to the sufficient statistics needed to evaluate the likelihood (4). Various types of partially observed data typify real data and have been considered in the literature, such as observing only the removal times [8, 20] or the number of infectious individuals in the population at discrete points in time [5]. For instance, the former arise in animal experiments in which positive cases are immediately isolated from the rest of the population and no longer contribute to disease spread. In such cases, exact times of removals are known, but such information is unavailable in typical observational studies.

In this article, we focus on partially observed *incidence* counts of infections at discrete points in time. Such data arise when an infectious individual may be identified some time after onset, and may continue to

infect others even after being tested positive for the disease. Consider an observation schedule $t_{0:K}$ ($K \geq 1$) with $0 = t_0 < t_1 < \cdots < t_K = t_{end}$. The observed data consist of the K-dimensional vector $\mathbf{Y} = I_{1:K}$ where $I_k = \#\{\tau_j^I \in (t_{k-1}, t_k]\}$ is the number of infections during the k^{th} time interval. For simplicity, we further assume that the initial configuration of the population $\mathbf{X}(0)$ is known.

If we adopt a Bayesian approach, the posterior distribution of the parameters given the observed data is formally related to the joint posterior via integration:

$$\pi(\theta|\mathbf{Y}) \ \propto \ L(\theta; \mathbf{Y})\pi(\theta) \ = \ \int_{\chi_{\mathbf{X}}^*} L(\theta; \mathbf{x})\pi(\theta)d\mathbf{x} \tag{8}$$

where $\pi(\theta)$ is the prior distribution on the parameters θ and $\chi_{\mathbf{X}}^* \subset \chi_{\mathbf{X}}$ is the set of trajectories of the process (1) that are compatible with the observed data $\mathbf{Y}$. The partial data likelihood $L(\theta; \mathbf{Y})$ therefore consists of a high-dimensional integral over all epidemic paths compatible with the observed data. This marginalization step has no closed form solution, and presents computational challenges even for a population of moderate size.

2.4 Prior Work

Researchers have explored several approaches for conducting inference on partially observed stochastic epidemic processes. Early approaches include the use of martingale-based equations [3, 26, 23]. However, such methods are difficult to apply to dynamic models with partially observed data and are therefore not suitable to the stochastic SIR model with incomplete data. More recently, researchers have based their computation on a simpler process that approximates the model's dynamics and whose likelihood in the presence of partially observed data is tractable. Popular approximations include chain binomial models [10, 1], diffusion processes [4, 6] and Gaussian processes [14]. While these approximation-based approaches bypass the intractability of the partial data likelihood 8, the assumptions on which they are based are questionable when the population is small, and, as a result, the dynamic of the stochastic process differs from its asymptotic behavior.

Since the popularization of Markov chain Monte Carlo methods (MCMC) in the field of statistics [24, 7, 25], researchers have developed sampling-based methods to directly work with the SIR likelihood instead of an approximation thereof. MCMC algorithms fall into two categories: model-based forward simulation and data augmented MCMC (DA-MCMC). Particle filtering [16] is an example of the former category that is extremely popular among practitioners. Its plug-and-play feature makes it applicable to a wide variety of models. However, model-based forward simulation suffers from two drawbacks: simulating data from a model as complex as the SEM that is compatible with the observed data is prohibitively slow, and these methods can fail to converge when the model does not fit the data well. Approximate Bayesian Computation (ABC) [17] offers a solution to the latter problem, but its inference is based on an approximation of the model's likelihood. As a result, the inference is inevitably biased, and it is difficult to quantify how closely the approximate posterior resembles the original target distribution.

The second family of MCMC-based inferential methods treats the unobserved event times as nuisance parameters; that is, the observed data are augmented with latent data that consist of the times and types of unobserved epidemic events. Researchers have developed algorithms to explore this latent space efficiently. Early attempts employed reversible-jump MCMC [9] to explore models with different numbers of unobserved events [8, 20]. Their algorithm augments the observed data which consist of the recovery times with the unobserved infection times and explores the latent space by inserting, deleting or uniformly moving an infection time in each iteration of the MCMC algorithm. More recently, Fintzi et al. [5] proposed a MCMC algorithm to conduct inference with discretely observed infection counts prevalence counts. They constructed latent data consisting of the infection and recovery times of each individual. The algorithm explores this latent space by updating the event times of an individual in each iteration of the algorithm according to the exact dynamics of the stochastic process. These DA-MCMC methods suffer from poor mixing in the presence of large epidemics. Since they update a single element of the latent data per iteration, the resulting Markov chains fail to explore the latent space efficiently. Although the update of multiple event times per iteration in [21] and the non-centered parameterization in [18] improves mixing, the gains are modest and the Markov chains still suffer from a high level of auto-correlation.

3 Exact Inference with a Data-Augmented Approach

We adopt a data-augmented MCMC approach that bridges the challenging partially observed setting to the tractable complete data likelihood by way of latent variables. In particular, our method hinges on the efficacy of a carefully designed proposal process for the latent variables that faithfully approximates the original process dynamics. The sampler targets the exact posterior of the model given partially observed incidence data and enjoys the efficiency of fast proposals in the high-dimensional latent space as well as fast computations involving the complete data likelihood.

As shown in Section 2.2, the complete data likelihood (4) is amenable to computation. This suggests augmenting the observed data $\mathbf{Y}$ with latent data $\mathbf{Z}$ such that the likelihood $L(\theta; (\mathbf{Y}, \mathbf{Z}))$ has the closed form (4) and constructing a Markov chain $\{(\theta^{(m)}, \mathbf{Z}^{(m)})\}_{m=0}^{M}$ with state space $\chi = \chi_\theta \times \chi_{\mathbf{Z}}$ whose stationary distribution is the joint posterior distribution $\pi(\theta, \mathbf{X}|\mathbf{Y})$ [8, 20, 5]. We consider the latent data $\mathbf{Z} = \left\{(z_j^I, z_j^R)\right\}_{j=1}^{n} \in \chi_{\mathbf{Z}}$ which consist of the times and types of unobserved epidemic events, where the infection and removal times for individual j are

$$z_j^I \begin{cases} \in [0, t_{end}) & \text{if individual } j \text{ becomes infectious before } t_{end} \\ = \infty & \text{if individual } j \text{ does not become infectious before } t_{end} \end{cases}$$

and

$$z_j^R \begin{cases} \in (z_j^I, t_{end}] & \text{if individual } j \text{ is removed before } t_{end} \\ = \infty & \text{if } z_j^I = \infty \text{ or individual } j \text{ is removed after } t_{end}. \end{cases}$$

The set $\chi_{\mathbf{Z}} \subset (\bar{\mathbb{R}}^+ \times \bar{\mathbb{R}}^+)^n$ consists of the latent data $\mathbf{Z}$ compatible with the observed data $\mathbf{Y}$ and the progression of an epidemic (see Equation 3). Since $\mathbf{Z}$ contains all the information present in $\mathbf{Y}$, we can write

$$L(\theta; (\mathbf{Y}, \mathbf{Z})) = L(\theta; \mathbf{Z}).$$

The data-augmented MCMC algorithm that we propose alternates between updates of the parameters θ given the current state of the latent data $\mathbf{Z}$, and updates of the latent data given the current state of the parameters. A one-step transition from $\mathbf{x}_0 = (\theta_0, \mathbf{z}_0)$ to $\mathbf{x}_1 = (\theta_1, \mathbf{z}_1)$ therefore looks like

$$\mathbf{x}_0 = (\theta_0, \mathbf{z}_0) \to (\theta_1, \mathbf{z}_0) \to (\theta_1, \mathbf{z}_1) = \mathbf{x}_1.$$

The transition kernel P of the Markov chain $\{\mathbf{x}^{(m)}\}_m$ is thus a composition of two kernels P_θ and $P_{\mathbf{z}}$ which alternatively updates θ while keeping $\mathbf{z}$ fixed, and then updates $\mathbf{z}$ while keeping θ fixed. Due to the Gamma conjugacy of the complete-data likelihood mentioned in Section 2.2, drawing new parameter values poses no challenge: we can simply employ a Gibbs sampler to directly draw new values for $\theta = (\beta, \gamma)$ from the two independent full conditional posterior distributions given by Equations 6 and 7. In these expressions, $n_I, n_R, \int S(t)I(t)dt, \int I(t)dt$ are sufficient statistics from the latent data $\mathbf{z}$. The latent data is updated in a Metropolis-Hastings step whose kernel is

$$P_{\mathbf{z}}((\theta, \mathbf{z}_0), (\theta, \mathbf{z}_1)) = q(\mathbf{z}_1|\theta)\alpha((\theta, \mathbf{z}_0), (\theta, \mathbf{z}_1))dz_1$$

where the proposal density q will be defined in the following section and

$$\alpha\left(\left(\theta^{(m)}, \mathbf{Z}^{(m-1)}\right), \left(\theta^{(m)}, \mathbf{Z}^{(m)}\right)\right) = \min\left\{1, \frac{L\left(\theta^{(m)}; \mathbf{Z}^{(m)}\right) q\left(\mathbf{Z}^{(m-1)}|\theta^{(m)}\right)}{L\left(\theta^{(m)}; \mathbf{Z}^{(m-1)}\right) q\left(\mathbf{Z}^{(m)}|\theta^{(m)}\right)}\right\}$$

corresponds to the Metropolis-Hasting acceptance ratio [25]. Algorithm 1 provides the details of the Data-Augmented MCMC.

3.1 Efficient Proposal Process for Latent Data

Though forward simulation of the stochastic SIR process is straightforward, simulating trajectories *conditionally* on the observed incidence data $\mathbf{Y}$ is a notoriously difficult task [13]. Instead, we will generate the latent variables conditionally on $\mathbf{Y}$ from a surrogate process, and accept or reject them in a Metropolis-Hastings step.

To this end, we consider a stochastic process whose dynamics closely resemble those of the SIR process and which is designed for efficient simulation of epidemic trajectories compatible with the incidence data $\mathbf{Y} = T_{1:K}$. We refer to this surrogate process as the *piecewise decoupled SIR* process (PD-SIR). Similarly to the SIR process, the PD-SIR process corresponds to a compartmental model in which individuals move from the compartments S to I and from I to R. The removal dynamics are identical under both processes: infection periods follow independent exponential distributions with rate γ, resulting in a constant individual-level removal rate γ.

Algorithm 1 Data-Augmented MCMC

Require: $\theta^{(0)}$

$\mathbf{Z}^{(0)} \sim q(\mathbf{Z}|\theta^{(0)})$ (generate the initial latent space)

for $j = 1, \ldots, N$ **do**

$\beta^{(j)}|Z^{(j-1)} \sim Ga\left(a_\beta + n_I^{(j-1)}, b_\beta + \int_0^{t_{end}} I^{(j-1)}(t)S^{(j-1)}(t)dt\right)$ (Gibbs update)

$\gamma^{(j)}|Z^{(j-1)} \sim Ga\left(a_\gamma + n_R^{(j-1)}, b_\gamma + \int_0^{t_{end}} I^{(j-1)}(t)dt\right)$ (Gibbs update)

$\theta^{(j)} \leftarrow (\beta^{(j)}, \gamma^{(j)})$

$\mathbf{Z}^\star \sim q(\mathbf{Z}|\theta^{(j)})$ (generate latent data from the PD-SIR process)

$\alpha = \min\left\{1, \frac{L(\theta^{(j)}; \mathbf{Z}^\star) q(\mathbf{Z}^{(j-1)}|\theta^{(j)})}{L(\theta^{(j)}; \mathbf{Z}^{(j-1)}) q(\mathbf{Z}^\star|\theta^{(j)})}\right\}$

$u \sim U(0, 1)$

if $u < \alpha$ **then**

$\mathbf{Z}^{(j)} \leftarrow \mathbf{Z}^*$

else

$\mathbf{Z}^{(j)} \leftarrow \mathbf{Z}^{(j-1)}$

end if

end for

The infection dynamics, however, vary slightly. In the SIR process, the population-level infection rate at time t is $\mu_T(t) = \beta S(t) I(t)$ and so, from the perspective of each susceptible individual, the individual-level infection rate is $\mu(t) = \frac{\mu_T(t)}{S(t)} = \beta I(t)$. Note that μ varies after every event since the value of I changes after an infection or a removal. In contrast, in the PD-SIR process, the infection rate is kept constant over small time intervals. Consider a resetting schedule $r_{0:L}$ $(L \geq 1)$ with $0 = r_0 < r_1 < \cdots < r_L$. The individual-level infection rate of the PD-SIR process is defined as

$$\tilde{\mu}(t) = \beta I(r_{k-1}), \forall t \in [r_{k-1}, r_k)$$

where $I(r_{k-1})$ denotes the number of infectious individuals at the beginning of the time interval. That is, the infection rate is kept constant over short periods of time. The infection rate μ and the evolution of the variable I are therefore decoupled during each interval, and the infection rate is reset at the beginning of the intervals. Over a single interval, the PD-SIR process is equivalent to a two-type branching process approximation of the SIR dynamics [11]. I really like how this is written but I wonder if there's some simple figure we can make to illustrate the model and difference between SIR at a glance =¿ line plot of mu over time with colored vertical line for every event. Done..

Now, by letting the resetting schedule $r_{0:L}$ coincide with the observation schedule $t_{0:K}$,

$$L = K, \text{ and } r_k = t_k, \quad k = 1, \ldots, K,$$

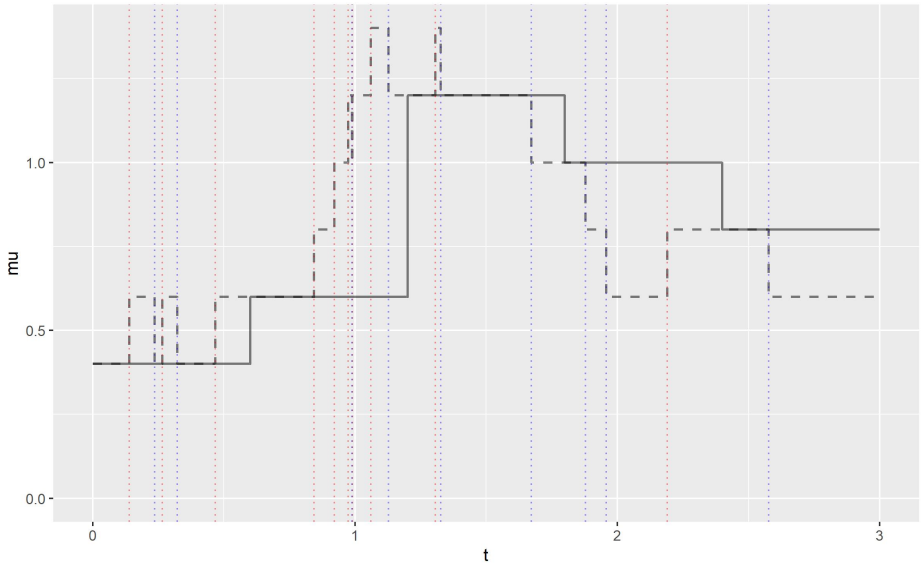


Figure 1: Infection dynamics of the SIR (dashed line) and PD-SIR (solid line) processes in a small population ($(S(0), I(0)) = (10, 2)$). $(\beta, \gamma) = (0.2, 1)$ and the resetting schedule of the PD-SIR contains $L = 5$ intervals of equal length. The vertical dotted lines indicate the infection (red) and removal (blue) times.

we can straightforwardly simulate realizations from the PD-SIR process conditionally on the observations $I_{1:K}$. By construction, under the PD-SIR model the population of susceptible individuals $S(t)$ follows a linear death process during each time interval with death rate $\mu_k = \beta I(t_{k-1})$. Generating event times from a linear death process conditionally on the number of events happening by a certain time can be done extremely efficiently as shown by the following theorem whose proof is presented in Appendix A.

Theorem 3.1. *Consider a linear death process with death rate μ and let $d_{1:T} \in (t_l, t_u]$ be the times of the N deaths occurring between times t_l and t_u. Then*

$$d_j \overset{d}{=} X_{(j)}, \quad j = 1, \ldots, N$$

where $X_{(j)}$ is the j^{th} order statistics of N i.i.d. random variables following a truncated exponential distribution with rate μ, lower bound t_l and upper bound t_u.

We can use this result to generate latent data $\mathbf{z} = \left\{(z_i^I, z_i^R)\right\}_{i=1}^n$ compatible with $\mathbf{Y} = I_{1:K}$ from the PD-SIR process as follows. For each time interval $[t_{k-1}, t_k)$, we compute $\mu_k = \beta I(t_{k-1})$ and $U_k = I(0) + \sum_{l=1}^{k-1} I_l$, the total number of infections that have happened before t_{k-1}. The index set $\mathfrak{I}_k = (U_k + 1, \ldots, U_k + I_k)$ therefore denotes the I_k individuals infected during the kth interval. Note that $I(t_{k-1})$ only depends on past events and can therefore be computed given the PD-SIR process up to time t_{k-1}. Algorithm 2 provides a simple recursion to compute this variable. For $j \in \mathfrak{I}_k$, by following Theorem 3.1, we draw the

infection times z_j^I as the order statistics of I_k i.i.d. truncated exponential random variables with rate μ_k, lower bound t_{k-1} and upper bound t_k. For the same indices j, we then generate the removal times z_j^R from the removal dynamics of the SIR. To accomplish this, we propose the removal time of individual j from the mixed distribution

$$(1-p_j)\delta_\infty + p_j \mathsf{TruncExp}(\gamma; z_j^I, t_{end})$$

placing point mass $(1-p_j)$ at ∞ and continuous mass on $(z_j^I, t_{end}]$, where δ_∞ corresponds to the Dirac distribution with mass 1 on the element ∞,

$$p_j = 1 - \exp\{-\gamma(t_{end} - z_j^I)\} = P(z_j^R < t_e nd | z_j^I)$$

is the cumulative distribution function of an exponential distribution with rate γ, and $\mathsf{TruncExp}(\gamma; z_j^I, t_{end})$ denotes a truncated exponential distribution with rate γ bounded between z_j^I and t_{end}. By construction, this scheme generates latent data from the PD-SIR process that are compatible with the observed data. Algorithm 2 provides a summary of the procedure in pseudo-code.

Based on this construction, we see that the density q of the PD-SIR process is

$$\begin{aligned} q(\mathbf{z}|\theta) = & \prod_{k=1}^{K} \prod_{j \in \mathcal{I}_k} \mathsf{TruncExp}(\tau_j^I; \mu_k, t_{k-1}, t_k) \\ & \times \prod_{i=1}^{n} (1-p_i)^{\mathbf{1}(z_i^R = \infty)} \left(p_i \mathsf{TruncExp}(z_i^I; \gamma, z_i^I, t_{end})\right)^{\mathbf{1}(z_i^R \le t_{end})} \\ = & \prod_{k=1}^{K} \prod_{j \in \mathcal{I}_k} \mathsf{TruncExp}(\tau_j^I; \mu_k, t_{k-1}, t_k) \\ & \times \prod_{i=1}^{n} (1-p_i)^{\mathbf{1}(z_i^R = \infty)} \left(p_i \mathsf{TruncExp}(z_i^I; \gamma, z_i^I, t_{end})\right)^{\mathbf{1}(z_i^R \le t_{end})} \end{aligned}$$

where

$$\mathsf{TruncExp}(x; \mu, l, u) = \frac{\mu \exp\{-\mu x\}}{\exp\{-\mu l\} - \exp\{-\mu u\}}, \quad x \in (l, u)$$

denotes the density of a truncated exponential distribution with parameters as notated previously.

3.2 Quality of Approximation

Since the proposal for the latent data is independent from the current configuration of the latent data, the efficiency of the DA-MCMC algorithm to explore the latent space directly depends on how similar the surrogate process is to the target process. In particular, the Markov chain will have good mixing properties if the PD-SIR process is a close approximation of the SIR process. The PD-SIR process only differs from the SIR process in its infection dynamics, the removal dynamics being identical in the two processes. Figure 2 compares the trajectories of the compartments S, I and R of a SIR process of moderate size ($(S(0) = 1000, I(0), R(0)) = (1000, 10, 0)$ with $(\beta, \gamma) = (0.003, 1)$ and $t_{end} = 6$) and those of four PD-SIR processes constrained to be compatible with the observed incidence data $I_{1:K}$ from the SIR process for

Algorithm 2 Generating a PD-SIR process conditionally on discretely observed infection incidence counts $I_{1:K}$

Require: $I_{1:K}, \theta = (\beta, \gamma)$, $I(0)$

for $j = 1, \ldots, I(0)$ **do**

$z_j^I \leftarrow 0$ (by the memoryless property of the exponential distribution).

$p_j \leftarrow 1 - \exp\{-\gamma(t_{end} - 0)\}$

$z_j^R \sim (1 - p_j)\delta_\infty + p_j \mathsf{TruncExp}(\gamma, 0, t_{end})$

end for

for $k = 1, \ldots, K$ **do**

$\mu_k \leftarrow \beta I(t_{k-1})$

$X_{1:I_k} \sim \mathsf{TruncExp}(\mu_k, t_{k-1}, t_k)$ i.i.d.

$U_k \leftarrow I(0) + \sum_{l=1}^{k-1} I_l$

$\mathfrak{I}_k \leftarrow (U_k + 1, \ldots, U_k + I_k)$

for $j \in \mathfrak{I}_k$ **do**

$z_j^I \leftarrow X_{(j)}$ (infection times)

$p_j \leftarrow 1 - \exp\{-\gamma(t_{end} - z_j^I)\}$

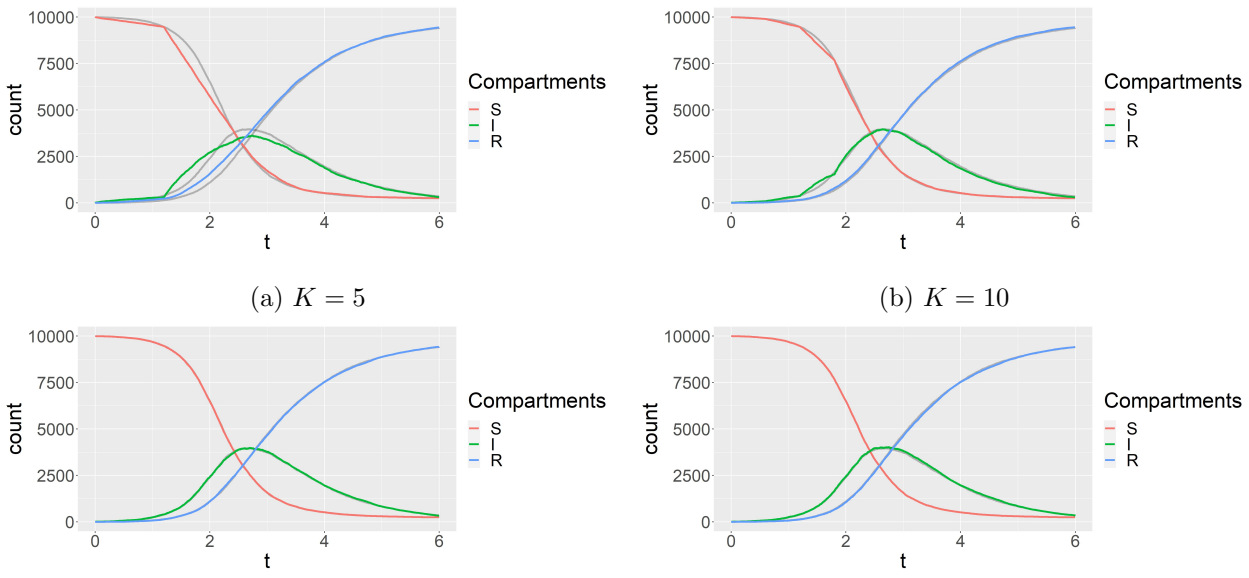
$z_j^R \sim (1 - p_j)\delta_\infty + p_j \mathsf{TruncExp}(\gamma, z_j^I, t_{end})$ (removal times)

end for

$R_k \leftarrow \#\{i : \tau_i^R \in (t_{k-1}, t_k]\}$ (number of removals in the k^{th} interval)

$I(t_k) \leftarrow I(t_{k-1}) + I_k - R_k$

end for



(a) $K = 5$ (b) $K = 10$

(c) $K = 50$ (d) $K = 1000$

Figure 2: Trajectories of the compartments S, I and R in a SIR (in colors) and PD-SIR processes (in grey) with the same infection incidence data $T_{1:K}$.

$K \in (5, 10, 50, 1,000)$. We see that the PD-SIR is qualitatively close to the SIR, even for small values of K. Unsurprisingly, the quality of the approximation improves as K increases. The fact that the PD-SIR is a faithful approximation of the SIR leads to a relatively high acceptance rate in the Metropolis-Hastings step for the latent data. As a result, our algorithm makes larger jumps in the high-dimensional latent space and therefore explores it more efficiently.

The following characteristics of our DA-MCMC algorithm are worth noting. First, initializing the Markov chain only requires setting $\theta^{(0)} = (\beta^{(0)}, \gamma^{(0)})$ since $\mathbf{Z}^{(0)}$ can be generated from the PD-SIR process conditionally on $\theta^{(0)}$ and $\mathbf{Y}$ alone. Second, since the PD-SIR closely approximates the SIR model, the acceptance rate in the Metropolis-Hastings step for the latent data is relatively high. For large population, however, the acceptance rate may be too low, thereby hindering the mixing of the Markov chain. To address this issue, we propose to update the event times of only a fraction $0 < \rho \leq 1$ of the individuals in the population. Smaller values for ρ result in smaller jumps in the latent space and a larger acceptance rate. Third, if all event times are updated ($\rho = 1$), then the proposed and current latent data are independent conditionally on the current values of the parameters: $\mathbf{Z}^{(k-1)} \perp \mathbf{Z}^{*} | \theta^{(k)}$. This contrasts with existing DA-MCMC algorithms which only update a small fraction of the latent data per iteration. Fourth, not only is proposing latent data from the RD-SIR process extremely fast since all random variables can be generated via the inverse-CDF method, but ensuring that these data are compatible with the observed data can be done at no additional cost, making the proposal scalable to large outbreaks. Finally, we show in the following section that the algorithm generates a Markov chain that is *uniformly ergodic*, the strongest form of ergodicity for

a Markov chain [25].

3.3 Uniform Ergodicity

It is straightforward to show that the Markov chain $\{(\theta^{(m)}, \mathbf{z}^{(m)})\}_m$ is harris ergodic and converges to the target posterior distribution $\pi(\theta, \mathbf{z}|\mathbf{Y})$. For any π-integrable function g, the estimator

$$\bar{g}_m := \frac{1}{m}\sum_{i=0}^{m-1} g(\theta^{(m)}, \mathbf{z}^{(m)})$$

is therefore consistent for $E_\pi g$, regardless of the starting values of the chain. A Markov chain is uniformly ergodic if for some finite M and positive constant $r < 1$ the n-step transition kernel P^n satisfies

$$\|P^n(x, .) - \pi(.)\|_{TV} \leq Mr^n, \forall x \in \chi$$

where $\|\mu\|_{TV}$ denotes the total variation norm of the measure μ. Uniform ergodicity ensures the existence of a Central Limit Theorem for $\bar{g}_m$ whenever $E_\pi|g|^2 < \infty$ (ref), that is,

$$\sqrt{m}(\bar{g}_m - E_\pi g) \Rightarrow N(0, \sigma^2)$$

for some positive, finite σ^2. Furthermore, uniform ergodicity implies that the batch means and spectral variance estimators of σ^2 are consistent (ref).

Toward establishing uniform ergodicity of our DA-MCMC algorithm, we will show that the state space $\chi = \chi_\theta \times \chi_{\mathbf{z}}$ of the Markov chain $\{(\theta^{(m)}, \mathbf{z}^{(m)})\}_m$ is a *small set* for the transition kernel P. That is, there exists a probability measure ν on the σ-algebra $\sigma(\chi)$ such that

$$\beta\nu(.) \leq P^m(x, .), \forall x \in \chi.$$

for some constants $m \geq 1$ and $\beta > 0$. By Proposition 2 in [25], this implies that the Markov is uniformly ergodic.

To derive this result, we will make use of three lemmas. The appendix contains a detailed a proof of the latter two. Figures for lemma 3.1 and 3.2?

Lemma 3.1. *Let $Ga(x; a, b)$ denote the density of the gamma distribution with shape a and rate b evaluated at x. Then*

$$\inf_{0\leq\beta\leq B} Ga(x; a, b+\beta) = \begin{cases} Ga(x; a, b), & x < x_a^* \\ Ga(x; a, b+B), & x \geq x_a^* \end{cases} \tag{9}$$

where $x_a^* = \frac{a}{B}\log\left(1 + \frac{B}{b}\right)$. Moreover,

$$\inf_{0\leq\alpha\leq A} Ga(x; a+\alpha, b) = \begin{cases} Ga(x; a, b), & x > x_b^* \\ Ga(x; a+A, b), & x \leq x_b^* \end{cases} \tag{10}$$

where $x_b^* = \frac{1}{b}\left[\frac{\Gamma(a+A)}{\Gamma(a)}\right]^{1/A}$.

Proof. Equation 9 is proven in Jones, Hobert (2004). For Equation 10, note that x_b^* is the only positive solution to $Ga(x; a, b) = Ga(x; a + A, b)$. Now, for all $0 < x \leq x_b^*$ and all $0 \leq \alpha \leq A$, we have

$$\begin{aligned}
\frac{Ga(x; a + A, b)}{Ga(x; a + \alpha, b)} &= b^{A-\alpha} x^{A-\alpha} \frac{\Gamma(a + \alpha)}{\Gamma(a + A)} \\
&\leq b^{A-\alpha} \left(\frac{1}{b} \left[\frac{\Gamma(a + A)}{\Gamma(a)} \right]^{1/A} \right)^{A-\alpha} \frac{\Gamma(a + \alpha)}{\Gamma(a + A)} \\
&= \left[\frac{\Gamma(a + A)}{\Gamma(a)} \right]^{(A-\alpha)/A} \frac{\Gamma(a + \alpha)}{\Gamma(a + A)} \\
&= \left(\frac{\Gamma_{a,a+A}}{\Gamma_{a+\alpha,a+A}} \right)^{A-\alpha} \\
&\leq 1,
\end{aligned}$$

where $\Gamma_{d,e} = \left(\frac{\Gamma(e)}{\Gamma(d)} \right)^{\frac{1}{e-d}}$ is a geometric average, and where the last inequality holds because $\Gamma_{a,a+A} \leq \Gamma_{a+\alpha,a+A}$. The case $x > x_b^*$ can be shown similarly. □

Lemma 3.1 can be used to minorize a gamma density of the form

$$Ga(x; a + \alpha, b + \beta) = \frac{(b + \beta)^{a+\alpha}}{\Gamma(a + \alpha)} x^{a+\alpha-1} \exp\{-x(b + \beta)\}, \quad 0 \leq \alpha \leq A, 0 \leq \beta \leq B. \tag{11}$$

over α and β for each value of x. To this end, we establish the following technical lemma.

Lemma 3.2. *For a fix $x > 0$, the density $Ga(x; a + \alpha, b + \beta)$ is minimized by $(\alpha, \beta) \in \{(A, 0), (0, B)\}$, the minimizing set of values depending on x. In particular,*

$$\inf_{\substack{0 \leq \alpha \leq A \\ 0 \leq \beta \leq B}} Ga(x; a + \alpha, b + \beta) = \begin{cases} Ga(x; a + A, b), & x < x_a \text{ or } x < x_{a+A}^* \vee x_b^* \\ Ga(x; a, b + B), & x_{a+A}^* < x \text{ or } x_a^* \wedge x_{b+B}^* < x \\ \min\{Ga(x; a + A, b), Ga(x; a, b + B)\}, & x_a^* \wedge x_b^* \leq x < x_{a+A}^* \vee x_{b+B}^* \end{cases}$$

The final lemma can be used to minorize truncated exponential densities.

Lemma 3.3.

$$\min_{0 \leq \mu \leq M} \mathit{TruncExp}(x; \mu, l, u) = \mathit{TruncExp}(x; M, l, u)$$

We can now proceed with the proof of uniform ergodicity.

Proposition 3.1. *The transition kernel $P = P_\theta P_{\mathbf{z}}$ of our DA-MCMC algorithm satisfies a minorization condition $M(1, \delta, \chi, \nu)$ and, as a result, is uniformly ergodic.*

Proof. First, note that the transition kernel P is a composition of two kernels

$$P = P_\theta P_{\mathbf{z}}$$

where the kernel P_θ updates the parameters θ while keeping the latent data $\mathbf{z}$ fixed and $P_\mathbf{z}$ update $\mathbf{z}$ while keeping θ fixed. A one-step transition therefore corresponds to the following scheme

$$\mathbf{x}_1 = (\theta_1, \mathbf{z}_1) \to (\theta_2, \mathbf{z}_1) \to (\theta_2, \mathbf{z}_2) = \mathbf{x}_2.$$

The kernel P has a density p with respect to the product measure $\mu = \lambda\mu_\mathbf{z}$ with λ the Lebesgue measure on χ_θ, the space of θ, and $\mu_\mathbf{z}$ a σ-finite measure on $\chi_\mathbf{z}$, the space of $\mathbf{z}$.

$$\begin{aligned} P(\mathbf{x}_1, d\mathbf{x}_2) &= p(\mathbf{x}_1, \mathbf{x}_2)\mu(d\mathbf{x}_2) \\ &= p((\theta_1, \mathbf{z}_1), (\theta_2, \mathbf{z}_2))\lambda(d\theta_2)\mu_\mathbf{z}(d\mathbf{z}_2) \\ &= p_\theta((\theta_1, \mathbf{z}_1), (\theta_2, \mathbf{z}_1))p_\mathbf{z}((\theta_2, \mathbf{z}_1), (\theta_2, \mathbf{z}_2))\mu_\mathbf{z}(d\mathbf{z}_2)\lambda(d\theta_2) \end{aligned}$$

where

$$\begin{aligned} p_\theta((\theta_1, \mathbf{z}_1), (\theta_2, \mathbf{z}_1)) &= \pi(\theta_2|\mathbf{z}_1) \\ &= \pi(\beta_2|\mathbf{z}_1)\pi(\gamma_2|\mathbf{z}_1) \end{aligned}$$

is the density of the transition kernel P_θ with respect to λ and corresponds to the product of the full conditional distributions of β and γ since the kernel P_θ is a Gibbs sampler, and

$$\begin{aligned} p_\mathbf{z}((\theta_2, \mathbf{z}_1), (\theta_2, \mathbf{z}_2)) &= q_\mathbf{z}((\theta_2, \mathbf{z}_1), (\theta_2, \mathbf{z}_2))\alpha((\theta_2, \mathbf{z}_1), (\theta_2, \mathbf{z}_2)) \\ &= q(\mathbf{z}_2; \theta_2) \min\left\{1, \frac{\pi(\theta_2, \mathbf{z}_2)q(\mathbf{z}_1; \theta_2)}{\pi(\theta_2, \mathbf{z}_1)q(\mathbf{z}_2; \theta_2)}\right\} \\ &= \min\left\{q(\mathbf{z}_2; \theta_2), \frac{\pi(\theta_2, \mathbf{z}_2)q(\mathbf{z}_1; \theta_2)}{\pi(\theta_2, \mathbf{z}_1)}\right\} \end{aligned}$$

is the density of the transition kernel $P_\mathbf{z}$ with respect to $\mu_\mathbf{z}$ and corresponds to one step of the Metropolis-Hasting algorithm where a new configuration of the latent data $\mathbf{z}_2$ is proposed conditionally on the current value of the parameters θ_2, but independently of the current configuration $\mathbf{z}_1$.

To show that χ is a small state, it is sufficient to show that there exists a function k such that

$$p(\mathbf{x}_1, \mathbf{x}_2) \geq k(\mathbf{x}_2) > 0, \qquad \forall \mathbf{x}_1 \in \chi$$

for all $\mathbf{x}_2 \in \chi$. The density p

$$\begin{aligned} p((\theta_1, \mathbf{z}_1), (\theta_2, \mathbf{z}_2)) &= \min\left\{\pi(\theta_2|\mathbf{z}_1)g(\mathbf{z}_2; \theta_2), \frac{\pi(\theta_2|\mathbf{z}_1)\pi(\theta_2, \mathbf{z}_2)}{\pi(\theta_2, \mathbf{z}_1)}g(\mathbf{z}_1; \theta_2)\right\} \\ &= \min\left\{\pi(\theta_2|\mathbf{z}_1)g(\mathbf{z}_2; \theta_2), \frac{\pi(\theta_2, \mathbf{z}_2)}{\pi(\mathbf{z}_1)}g(\mathbf{z}_1; \theta_2)\right\} \\ &= \min\left\{\pi(\theta_2|\mathbf{z}_1)g(\mathbf{z}_2; \theta_2), \pi(\theta_2|\mathbf{z}_2)g(\mathbf{z}_1; \theta_2)\right\} \end{aligned}$$

depends on $x_1 = (\theta_1, \mathbf{z}_1)$ only through the full conditional $\pi(\theta_2|\mathbf{z}_1)$ and the proposal density $g(\mathbf{z}_1; \theta_2)$.

It therefore suffices to show that there exist functions k_1 and k_2 such that

$$\pi(\theta|\mathbf{z}_1) \geq k_1(\theta) > 0 \qquad \text{and} \qquad g(\mathbf{z}_1; \theta) \geq k_2(\theta) > 0.$$

First,

$$\begin{aligned}\pi(\theta|\mathbf{z}_1) &= \pi(\beta|\mathbf{z}_1)\pi(\gamma|\mathbf{z}_1) \\ &= Ga(\beta; a_\beta + n_T, b_\beta + SI_1)Ga(\gamma; a_\gamma + n_{J1}, b_\gamma + I_1) \\ &\geq h_\beta(\beta)h_\gamma(\gamma) \\ &= k_1(\theta) > 0\end{aligned}$$

where, by Proposition 3.1,

$$h_\beta(\beta) = \begin{cases} Ga(\beta; a_\beta + n_T, b_\beta), & \beta < \frac{a_\beta + n_T}{t_{end}(n_T + I_0)n} \log\left(1 + \frac{t_{end}(n_T+I_0)n}{b_\beta}\right) \\ Ga(x; a_\beta + n_T, b_\beta + n(n_T + I_0)t_{end}), & \text{else} \end{cases}$$

since $n_T = \sum_k T_k$ is known and

$$SI_1 = \int_0^{t_{end}} S_1(t)I_1(t)dt \in [0, n(n_T + I_0)t_{end}];$$

and, by Proposition 3.2,

$$h_\gamma(\gamma) = \min\{Ga(\gamma; a_\gamma + n_T + I_0, b_\gamma), Ga(x; a_\gamma, b_\gamma + (n_T + I_0)t_{end})\}$$

since $0 \leq n_1^J \leq n_T + I_0$ and

$$I_1 = \int_0^{t_{end}} I_1(t)dt \in [0, (n_T + I_0)t_{end}].$$

Second, each factor in

$$g(\mathbf{z}_1; \theta) = \prod_{i=1}^{n_T} \mathsf{TruncExp}(z_i^T; \beta I_k, t_{k(i)-1}, t_{k(i)})(1 - p_i)^{1\{\mathbf{z}_i^J = \infty\}}(p_i \mathsf{TruncExp}(z_i^J; \gamma, z_i^T, t_{end}))^{1\{\mathbf{z}_i^J \leq t_{end}\}}$$

can be minorized as follows. First,

$$\begin{aligned}\mathsf{TruncExp}(z_i^T; \beta I_k, t_{k(i)-1}, t_{k(i)}) &\geq \mathsf{TruncExp}(t_{k(i)}; \beta I_k, t_{k(i)-1}, t_{k(i)}) \\ &\geq \mathsf{TruncExp}(t_{k(i)}; \beta(n_T + I_0), t_{k(i)-1}, t_{k(i)}),\end{aligned}$$

with the last inequality following from Proposition 3.3, second,

$$1 - p_i = \exp\{-\gamma(t_{end} - z_i^T)\} \geq \exp\{-\gamma(t_{end} - t_{k(i)-1})\},$$

third,

$$p_i \mathsf{TruncExp}(z_i^J; \gamma, z_i^T, t_{end}) = \mathsf{Exp}(z_i^J - z_i^T; \gamma)1(z_i^T < z_i^J \leq t_{end}) \geq \mathsf{Exp}(t_{end} - t_{k(i)-1}; \gamma)$$

where the equality holds since $p_i = P(z_i^J \leq t_{end} | z_i^J > z_i^T)$ is equal to the normalizing constant of the truncated exponential.

Taken together, these inequalities give

$$\begin{aligned}g(\mathbf{z}_1; \theta) &\geq \prod_{i=1}^{n_T} \mathsf{TruncExp}(t_{k(i)}; \beta(I_0 + n_T), t_{k(i)-1}, t_{k(i)}) \min\{\exp\{-\gamma(t_{end} - t_{k(i)-1})\}, \mathsf{Exp}(t_{end} - t_{k(i)-1}; \gamma)\} \\ &= k_2(\theta) > 0\end{aligned}$$

which completes the proof. □

The kernel

$$P_\theta((\theta_0, \mathbf{z}), (\theta_1, \mathbf{z})) = Ga\left(\beta_1; a_\beta + n_I, b_\beta + \int_0^{t_{end}} S(t)I(t)dt\right) Ga\left(\gamma_1; a_\gamma + n_R, b_\gamma + \int_0^{t_{end}} I(t)dt\right) d\beta_1 d\gamma_1$$

therefore corresponds to a Gibbs kernel where $n_I, n_R, \int S(t)I(t)dt, \int I(t)dt$ are sufficient statistics from the latent data $\mathbf{z}$

4 Performance on simulated and real epidemic data

4.1 Simulation Study

In this section, the convergence properties of the DA-MCMC algorithm are examined with a simulation study. Each data set in this study is simulated from the stochastic SIR process and the prior distributions on the parameters are the independent weakly informative distributions $\beta \sim Ga(0.1, 1)$ and $\gamma \sim Ga(1, 1)$.

First, the mixing property of the Markov chain is assessed in a medium-size population whose initial configuration is $(S(0), I(0)) = (1000, 10)$. The parameters are $(\beta, \gamma) = (0.003, 1)$ and the process is observed until time $t_{end} = 6$ when the pandemic has completed most of its course but is not over yet (see Figure 3, where $I(t_{end}) = 55$). The numbers of infections in $K = 10$ time intervals of equal length are observed and correspond to $T_{1:K} = (40, 111, 193, 259, 178, 93, 29, 19, 9, 6)$. The Markov chain is run for $100,000$ iterations from the initial values $(\beta^{(1)}, \gamma^{(1)}) = (0.0003, 0.1) = (\beta/10, \gamma/10)$ and the event times of 202 ($\rho = 0.2$) individuals are updated in each Metropolis-Hastings step for the latent data. The algorithm takes in 2.5 minutes to complete on a personal laptop. Figure 4 shows the traceplots of the parameters β and γ, the log likelihood and the basic reproduction number $R_0 = S(0)\beta/\gamma$, which corresponds to the average number of secondary infections generated by an infectious individual in a susceptible population. We observe that the Markov chain quickly migrates from the low density region of the space where $(\beta^{(1)}, \gamma^{(1)})$ is located to a high density region where it mixes well. The acceptance rate of the latent spaces proposed in the Metropolis-Hastings step is 0.11. Excluding the first $10,000$ iterations of the Markov chain as a burn-in, the posterior means of β, γ and R_0 are respectively 0.00304, 0.995 and 3.07.

Second, the impact of the tuning parameter ρ on the performance of the MCMC algorithm is evaluated. In particular, we examine the acceptance rate of the Metropolis-Hastings step, the total run time of the algorithm and the auto-correlation function of the parameters. Three SIR processes with different population sizes $S(0) = (10^2, 10^3, 10^4)$ are simulated. We set $I(0) = 10$, $\gamma = 1$, $t_{end} = 3$ and $K = 50$ for all simulations and $\beta = (0.001, 0.01, 0.1)$ respectively for the three different population sizes. These values for β ensure that the three processes have comparable dynamics. For each process, the DA-MCMC algorithm is run for $50,000$ iterations with different values of $\rho = (10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1)$. The algorithm is run only if there is at least one event time updated per iteration, that is, if $\rho(S_0 + I_0) > 1$. Figure **??** presents the acceptance rate across the range of ρ. We observe that the acceptance rate decreases as the number of event times updated per iteration and the population size increase. Figure **??** shows the run time of the MCMC

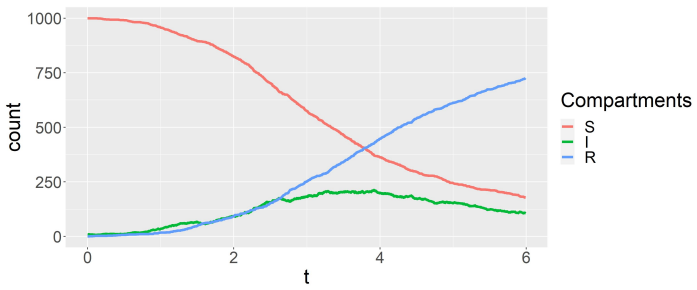
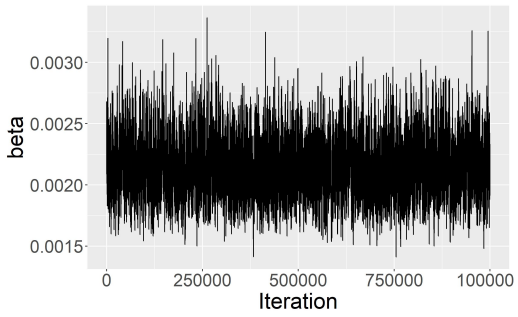
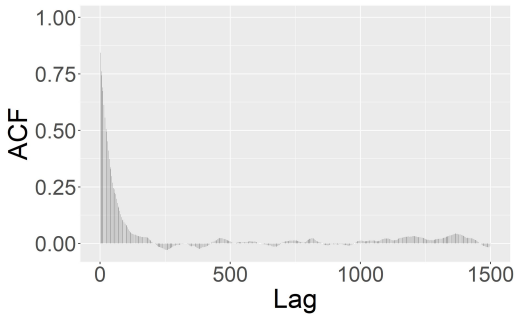


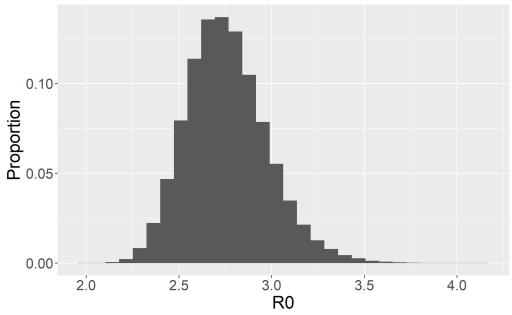
Figure 3: Trajectories of the S, I and R compartments of a SIR process for a medium-size population ($S_0 = 1,000$).



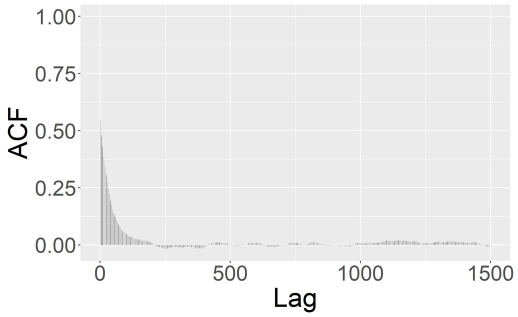
(a) β



(b) γ



(c) Log likelihood



(d) R_0

Figure 4: Traceplots of the Markov chain (including burn-in).

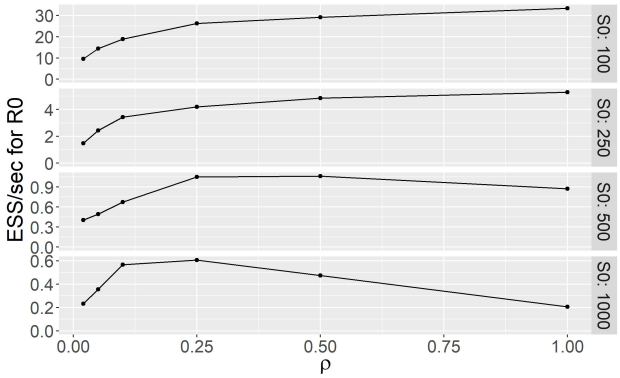


Figure 5: ESS/sec for R_0

algorithm for these simulations. The run time increases with the population size and with ρ. Finally, Figure 6 shows the effect of ρ on the auto correlation function for the parameter β for the medium-size population process ($S(0) = 1,000$) for three values of ρ. We observe that updating half of the augmented data per iteration yields the lowest auto-correlation function across the values of ρ considered. This suggests that, in this setting, updating the entire latent space in the Metropolis-Hastings step results in too many rejections while updating a tenth or less prevents the chain from efficiently exploring the latent space.

4.2 Ebola Outbreak in Guéckédou, Guinea

We now turn to a case study concerning the Ebola outbreak in Western Africa. Between the end of 2013 and 2015, Guinea, along with several neighboring countries, experienced the largest outbreak of the Ebola virus disease in history. The virus, which has a fatality rate of 70%, was responsible for the death of almost $2,000$ people in Guinea alone during the outbreak. The outbreak is believed to have originated in the Guéckédou prefecture of Guinea at the end of November 2013 [2]. Weekly infection incidence counts are available for each prefecture for the 73 weeks between the end of December 2013 and May 2015. We fit the stochastic SIR model to these incidence counts for the Guéckédou prefecture using the MCMC algorithm proposed in this article. For simplicity, we assume that the population of the prefecture forms a closed population, that the model's parameters remained constant throughout the outbreak and that the reported infections counts are exact. Applying our MCMC algorithm to stochastic epidemic models where these assumptions are relaxed is the subject of ongoing research. The *effective* population size is set to $n = 150,000$, roughly half of the total population in 2014. Note that as long as the number of infections is small compared to the total population size, the exact population size does not need to be known to estimate the parameters of the model. Furthermore, since the first documented infection occurred late November, one month before the first reported incidence count [2], the number of infectious individuals at the beginning of the observation period is set to $I(0) = 10$, with the units of time corresponding to "days" in the analysis.

The Markov chain is run for $50,000$ iterations and the event times of 10% of the individuals are

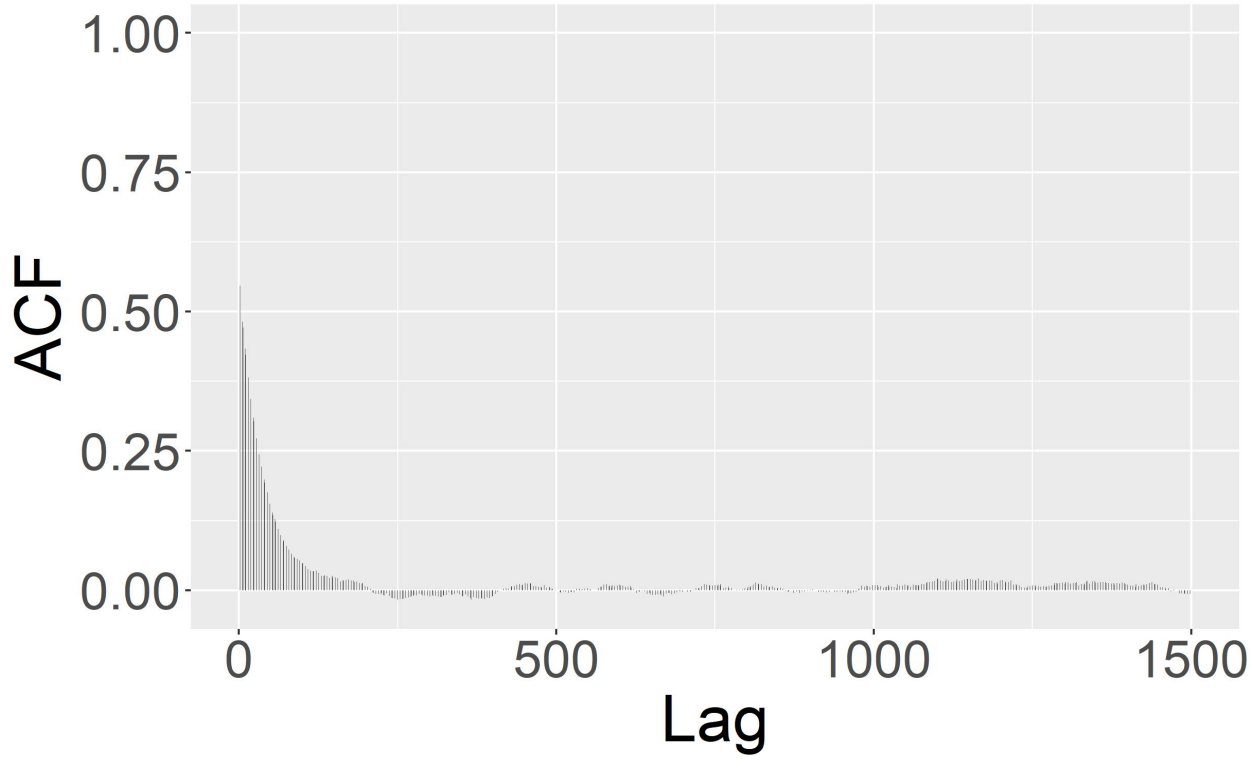
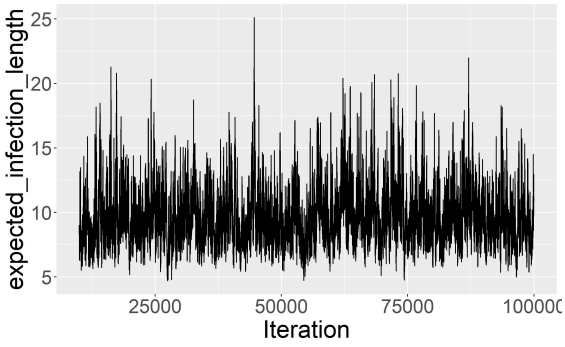


Figure 6: $\rho = 0.1$

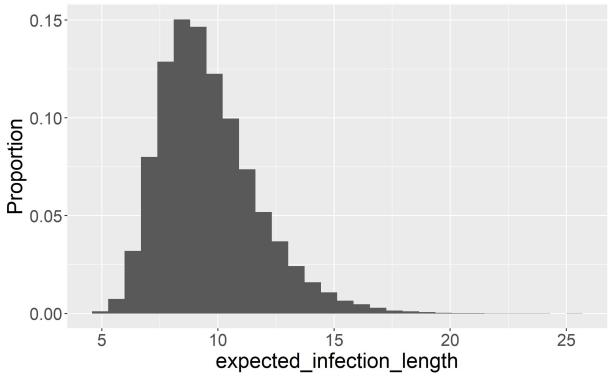
updated each iteration. The initial values of the parameters are set to $(\beta^{(1)}, \gamma^{(1)}) = (10^{-7}, 0.05)$ and the conjugate distributions in Equation 5 are used as priors with $a_\beta = b_\beta = a_\gamma = b_\gamma = 0.01$ to make them weakly informative. Even for such a large population, the total run time of the algorithm was less than 100 minutes on a personal laptop. The Metropolis-Hastings step for the latent space proposals achieves a healthy 22.1% acceptance rate, and the first $10,000$ iterations of the Markov chain are discarded as a burn-in. Figure 7 shows the joint posterior distributions of (β, γ), whose posterior means are respectively $1.06 * 10^{-6}$ and 0.109, as well as the marginal distribution of the basic reproduction number R_0. This value for γ indicates that people remain infectious for around 9 days on average, which is consistent with existing literature. As noted by numerous authors, conditionally on partially observed data, the parameters β and γ appear positively correlated. The posterior distribution of R_0 is unimodal, relatively symmetric, and centered around 1. A basic reproduction number close to 1 is consistent with an outbreak that lasted several months without infecting the majority of the population.

5 Discussion and Conclusion

The method proposed in this article enables the classical Metropolis-Hastings algorithm to be an efficient method to conduct full posterior inference of the stochastic SIR model given only incidence data. Existing



(a) Joint posterior density of $\theta = (\beta, \gamma)$. The 50% and 90% confidence regions are shown.



(b) Posterior density of R_0.

Figure 7: Posterior densities of parameters of interest (excluding burn-in of $10,000$ iterations).

attempts using Markov chain Monte Carlo either mix poorly or do not scale to populations of more than a few hundred individuals, while methods relying on forward simulation are limited to moderate-sized outbreaks and may suffer from degeneracy issues on a case-by-case basis depending on the amount of missing data. Moreover, the simplifying assumptions in [6] or the approximate-Bayesian-computation framework of [17], which compromise targeting the exact posterior for computational reasons, generate biased estimates that may lead to spurious conclusions. Instead, our data-augmented algorithm enables fast and exact inference, even for large outbreaks, and leverages a well-studied, transparent MCMC framework.

Central to the success of the DA-MCMC algorithm is an efficient proposal scheme that swiftly explores the latent space of epidemic paths that are compatible with the observed data. The PD-SIR process possesses three features that yield such an efficient algorithm. First, generating a PD-SIR process is extremely fast; it only requires the simulation of truncated exponential distributions, which can efficiently be realized via the inverse-CDF method. Moreover, the individual infection rates are only updated K times, as opposed to after each event as in the SIR model. Second, generating a PD-SIR process that is constrained to be compatible with the observed incidence data can be done at no additional cost; in contrast, generating a SIR process compatible with the observed data would be prohibitively slow [13]. Third, the dynamics of the PD-SIR process closely resemble those of the SIR process: the removal dynamics are identical and, for short resetting intervals, the infection dynamics are also very similar in the two processes.

The first two features make the algorithm extremely fast. While existing data-augmented MCMC algorithms have only be applied to populations of a few hundred of individuals, the analysis of the Ebola outbreak in Guéckédou shows that our algorithm can be applied to populations of up to $150,000$ individuals, generating tens of thousands of posterior samples in a reasonable amount of time. The latter feature of the PD-SIR process enables the algorithm to update a large portion of the augmented data in each iteration while maintaining a healthy acceptance rate. As a result, the Markov chain makes large jumps in the latent space and has very good mixing properties. In contrast, existing DA-MCMC algorithms keep most of the

latent space fixed across iterations [8, 20, 5] and their Markov chains therefore mix much more slowly.

The algorithm possesses a tuning parameter ρ that determines the portion of the augmented data being updated each iteration. Larger values for ρ enable the chain to make larger steps in the latent space but can result in a low acceptance rate, while smaller values of ρ result in a higher acceptance rate but constrain the chain to make small jumps. Depending on the size of the population, different values of the tuning parameter are optimal. To find a value that optimizes the mixing properties of the chain, one can use several short runs of the algorithm with different values for ρ and select the value that yields the lowest auto-correlation function (ACF). Since the run time of the algorithm increases with ρ, one could also look at the effective sample size per unit of time. For instance, for populations of a few hundred individuals, we observed that updating the entire augmented data ($\rho = 1$) yields the lowest ACF. In this case, the current and proposed augmented data are independent conditionally on the current value of the parameters.

The DA-MCMC algorithm proposed in this article is specific to the stochastic SIR model, which is arguably a simplistic representation of the spread of disease. The model relies on assumptions such as perfect reporting, constant infection rates, homogeneously mixing population and exponentially-distributed infectious periods. In future work, we will present applications of our DA-MCMC algorithm to processes where these assumptions are relaxed. In particular, we are considering extensions to epidemic models with under-reporting, a varying infection rate, stratified populations and non-Markovian dynamics, as well as the general stochastic SIR model of [22].

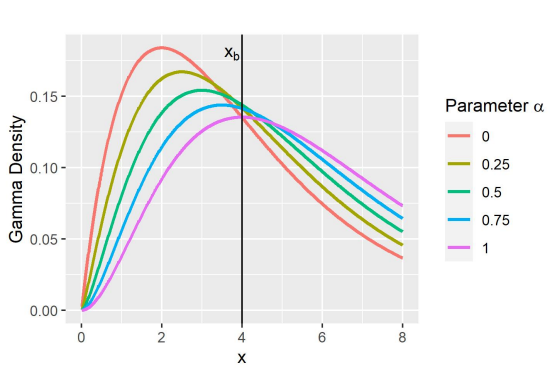
A Distribution of Death Times in Linear Death Process

Theorem 3.1 was first proved by [19]. Since this is a result that has rarely been mentioned in the literature, we provide a proof in this appendix from (cite Ross textbook).

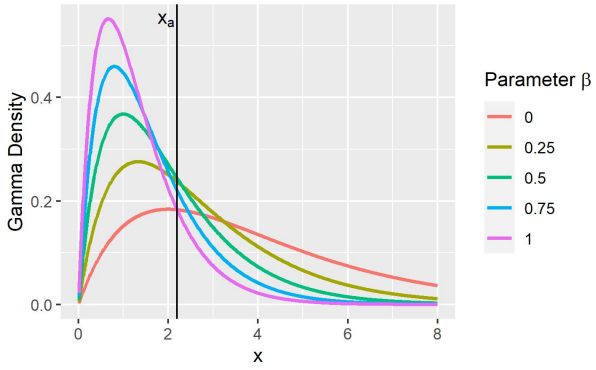
Consider a linear pure death process with n particles and individual death rate μ. Let T_i be the time of the ith death. Then $W_1 = T_1 \sim \mathsf{Exp}(n\mu)$ and $W_i = T_i - T_{i-1} \sim \mathsf{Exp}((n-1)\mu)$ independently. Let N be the number of deaths by time t. Then,

$$
\begin{aligned}
&f(T_1 = t_1, \ldots, T_N = t_N | N) \\
&\propto f(T_1 = t_1, \ldots, T_N = t_N, T_{N+1} > t) 1\{T_N < t\} \\
&\propto f(W_1 = t_1, W_2 = t_2 - t_1, \ldots, W_N = t_N - t_{N-1}, W_{N+1} > t - t_N) 1\{T_N < t\} \\
&\propto \exp\{-n\mu t_1\} \exp\{-(n-1)\mu(t_2 - t_1)\} \ldots \exp\{-(n-N)\mu(t - t_N)\} 1\{T_N < t\} \\
&\propto \exp\{-\mu t_1\} \exp\{-\mu t_2\} \ldots \exp\{-\mu t_N\} 1\{T_N < t\}
\end{aligned}
$$

which corresponds to the kernels of independent exponential distribution truncated above by t.



(a) $\alpha \in [0,1]$, $\beta = 0$



(b) $\alpha = 0$, $\beta \in [0,1]$

Figure 8: Example of minorization of the density of a gamma distribution $Ga(2+\alpha, 0.5+\beta)$ for α and β separately.

B Proofs

Proof of Lemma 3.2:

Proof. Given $x > 0$, 9 shows that for a fixed α, $\beta \in \{0, B\}$ is the minimizer of 11, and 10 shows that for a fixed β, $\alpha \in \{0, A\}$ is the minimizer of 11. This implies that for a fixed $x > 0$, 11 is minimized by $(\alpha, \beta) \in \{(0,0), (A,0), (0,B), (A,B)\}$.

A case-by-case analysis of the nine possibilities

$$(x < x_a^*; x_a^* < x < x_{a+A}^*; x_{a+A}^* < x) \times (x < x_{b+B}^*; x_{b+B}^* < x < x_b^*; x_b^* < x)$$

is presented in Table 1 and shows that it is sufficient to consider $(\alpha, \beta) \in \{(A,0), (0,B)\}$.

	$x < x_a^*$	$x_a^* \le x < x_{a+A}^*$	$x_{a+A}^* < x$
$x < x_b^*$	$(a+A, b)$	$(a+A, b)$	
$x_b^* \le x < x_{b+B}^*$	$(a+A, b)$	ad hoc	$(a, b+B)$
$x_{b+B}^* < x$		$(a, b+B)$	$(a, b+B)$

Table 1: Values of (α, β) that minimize $Ga(x; a+\alpha, b+\beta)$ for different values x.

The two empty entries in Table 1 correspond to impossible configurations. Indeed, $x_b^* > x_a^*$ and $x_{a+A}^* > x_{b+B}^*$ for all a, A, b, B since

$$\frac{x_b^*}{x_a^*} = \frac{\left(\frac{\Gamma(a+A)}{\Gamma(a)}\right)^{1/A} b^{-1}}{aB^{-1}\log(1+B/b)} = \frac{\left(\frac{\Gamma(a+A)}{\Gamma(a)}\right)^{1/A}}{a} \frac{B/b}{\log(1+B/b)} = \frac{\Gamma_{a,a+A}}{a} \frac{B/b}{\log(1+B/b)} \ge 1$$

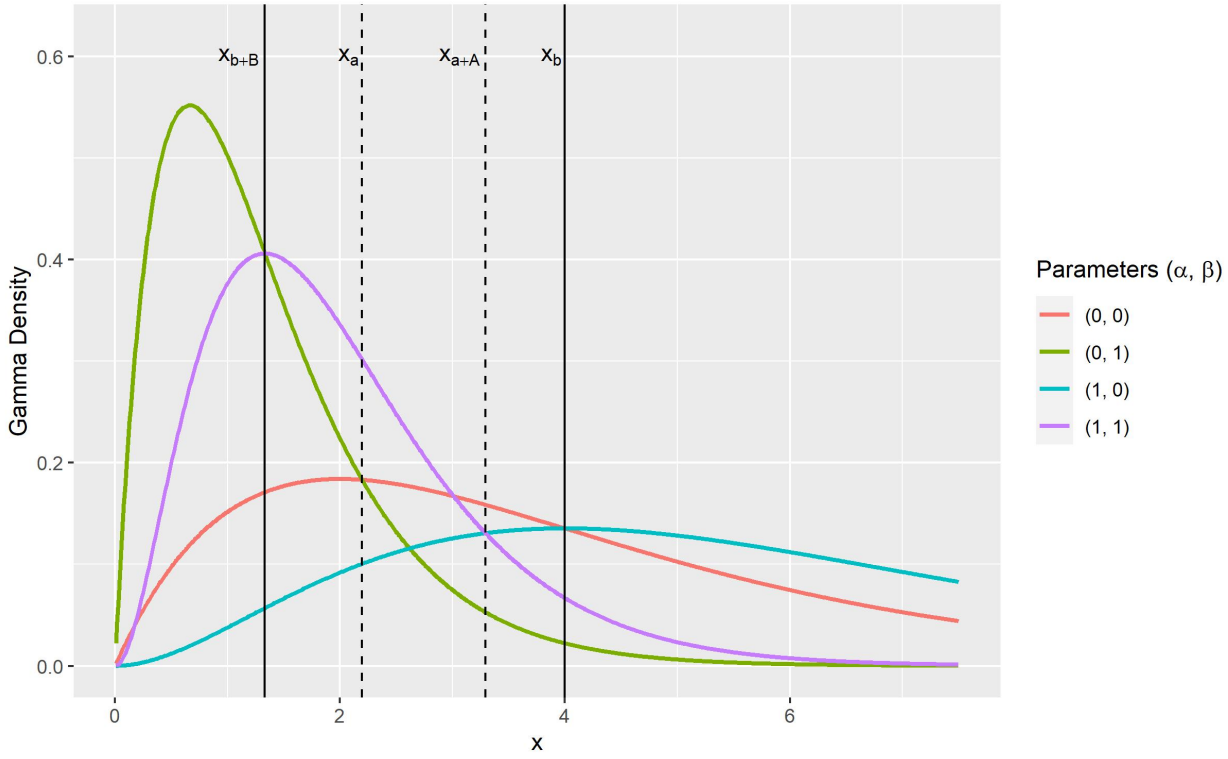


Figure 9: Example of minorization of the density of a gamma distribution $Ga(2+\alpha, 0.5+\beta)$ for α and β jointly.

and

$$\frac{x^*_{b+B}}{x^*_{a+A}} = \frac{\left(\frac{\Gamma(a+A)}{\Gamma(a)}\right)^{1/A}(b+B)^{-1}}{(a+A)B^{-1}\log(1+B/b)} = \frac{\Gamma_{a,a+A}}{a+A}\frac{B}{b+B}\frac{1}{\log(1+B/b)} \leq 1$$

where the inequalities hold since $a \leq \Gamma_{a,a+A} \leq a+A$ and $\frac{y}{\log(1+y)} \leq 1$.

If $x^*_a \wedge x^*_{b+B} \leq x < x^*_{a+A} \vee x^*_b$, which corresponds to the middle entry of the center column in Table 1, then one needs to directly check which set of values in $\{(0,0),(A,0),(0,B),(A,B)\}$ minimizes Equation 11. In fact, it is sufficient to consider only $\{(A,0),(0,B)\}$ since

$$Ga(x, a+A, b) \leq \begin{cases} Ga(x;a,b), & x < x^*_b \\ Ga(x;a+A,b+B), & x < x^*_{a+A} \end{cases}$$

and

$$Ga(x, a, b+B) \leq \begin{cases} Ga(x;a,b), & x > x^*_a \\ Ga(x;a+A,b+B), & x > x^*_{b+B} \end{cases}$$

□

Proof of Lemma 3.3

Proof. Remember that $\mathsf{TruncExp}(u;\mu,0,u) = \frac{\mu\exp\{-u\mu\}}{1-\exp\{-u\mu\}}$ is the density of an exponential distribution bounded above by u.

$$\frac{d}{d\mu}\mathsf{TruncExp}(u;\mu,0,u) = \left(\frac{\exp\{-u\mu\}}{1-\exp\{-u\mu\}}\right)\left[1-\mu u-\mu u\left(\frac{\exp\{-u\mu\}}{1-\exp\{-u\mu\}}\right)\right]$$

Now let $g(\mu) = 1-\mu u-\mu u\left(\frac{\exp\{-u\mu\}}{1-\exp\{-u\mu\}}\right)$. By L'Hôpital's rule, $g(0)=0$. Moreover, for $\mu>0$, we have

$$\begin{aligned} g'(\mu) &= -\mu\left(1+\frac{\exp\{-u\mu\}}{1-\exp\{-u\mu\}}\right)\left[1-u\mu\left(\frac{\exp\{-u\mu\}}{1-\exp\{-u\mu\}}\right)\right] \\ &= -\mu\left(1+\frac{\exp\{-u\mu\}}{1-\exp\{-u\mu\}}\right)\left[1-\frac{u\mu}{\exp\{u\mu\}-1}\right] \\ &\leq 0 \end{aligned}$$

where the inequality follows from $\frac{u\mu}{\exp\{u\mu\}-1} \leq 1$. This implies that $g(\mu)\leq 0$ for $\mu>0$. We therefore have $\frac{d}{d\mu}\mathsf{TruncExp}(u;\mu,0,u)\leq 0$ and the result of the proposition follows. □

References

[1] Helen Abbey. An examination of the reed-frost theory of epidemics. *Human biology*, 24(3):201, 1952.

[2] Sylvain Baize, Delphine Pannetier, Lisa Oestereich, Toni Rieger, Lamine Koivogui, N'Faly Magassouba, Barrè Soropogui, Mamadou Saliou Sow, Sakoba Keita, Hilde de Clerck, et al. Emergence of zaire ebola virus disease in guinea. *New England Journal of Medicine*, 371(15):1418–1425, 2014.

[3] Niels G. Becker. On a general stochastic epidemic model. *Theoretical Population Biology*, 11(1):23–36, 1977.

[4] Simon Cauchemez and Neil 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(25):885–897, 2008.

[5] Jonathan Fintzi, Xiang Cui, Jon Wakefield, and Vladimir N. Minin. Efficient data augmentation for fitting stochastic epidemic models to prevalence data. *Journal of Computational and Graphical Statistics*, 26(4):918–929, 2017.

[6] Jonathan Fintzi, Jon Wakefield, and Vladimir N. Minin. A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. *arXiv preprint arXiv:2001.05099*, 2020.

[7] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[8] Gavin J. Gibson and Eric Renshaw. Estimating parameters in stochastic compartmental models using markov chain methods. *Mathematical Medicine and Biology: A Journal of the IMA*, 15(1):19–40, 1998.

[9] Peter J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[10] Major Greenwood. On the statistical measure of infectiousness. *Epidemiology & Infection*, 31(3):336–351, 1931.

[11] Lam Si Tung Ho, Forrest W. Crawford, Marc A. Suchard, et al. Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease. *The Annals of Applied Statistics*, 12(3):1993–2021, 2018.

[12] Lam Si Tung Ho, Jason Xu, Forrest W. Crawford, Vladimir N. Minin, and Marc A. Suchard. Birth/birth-death processes and their computable transition probabilities with biological applications. *Journal of mathematical biology*, 76(4):911–944, 2018.

[13] Asger Hobolth and Eric 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(3):1204, 2009.

[14] Roman Jandarov, Murali Haran, Ottar Bjørnstad, and Bryan Grenfell. Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, pages 423–444, 2014.

[15] William Ogilvy Kermack and Anderson G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[16] Aaron A. King, Dao Nguyen, and Edward L. Ionides. Statistical inference for partially observed markov processes via the r package pomp. *arXiv preprint arXiv:1509.00503*, 2015.

[17] Trevelyan J. McKinley, Ian Vernon, Ioannis Andrianakis, Nicky McCreesh, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein, Richard G. White, et al. Approximate bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statistical science*, 33(1):4–18, 2018.

[18] Peter Neal and Gareth Roberts. A case study in non-centering for data augmentation: stochastic epidemics. *Statistics and Computing*, 15(4):315–327, 2005.

[19] Marcel F. Neuts and Sidney I. Resnick. On the times of births in a linear birthprocess. *Journal of the Australian Mathematical Society*, 12(4):473–475, 1971.

[20] Philip D. O'Neill and Gareth O. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129, 1999.

[21] 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(107):20150225, 2015.

[22] Norman C. Severo. Generalizations of some stochastic epidemic models. *Mathematical biosciences*, 4(3-4):395–402, 1969.

[23] Aidan Sudbury. The proportion of the population never hearing a rumour. *Journal of applied probability*, pages 443–446, 1985.

[24] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical association*, 82(398):528–540, 1987.

[25] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[26] Ray Watson. An application of a martingale central limit theorem to the standard epidemic model. *Stochastic Processes and Their Applications*, 11(1):79–89, 1981.