RT2 Report:
Bayesian Inference for Stochastic Epidemics

James Neill
Supervisors: Chris Jewell and Lloyd Chapman

May 2023

1 Introduction

Stochastic epidemic modelling is the application of statistical models to spread of infectious pathogens through a population. Allen (2017) states that “stochastic modelling of epidemics is important when the number of infectious individuals is small or when the variability in transmission, recovery, births, deaths, or the environment impacts the epidemic outcome”. These models can then be used to derive information about real-world epidemics.

In this report, the main model we will be investigating is the SIR model. The SIR model splits the population into three categories (susceptible, infective, and recovered), where individuals can move between the categories at constant but unknown rates. Our goal is to gain information on these rates, given that we have data on the times individuals recover from the pathogen. However, we generally do not have data on the time when individuals are infected with the pathogen, resulting in a missing data problem.

In order to conduct inference about the model parameters and infection times, we will be using a Bayesian approach; specifically, we will be using Markov chain Monte Carlo methods, such as the Metropolis-Hastings algorithm and the Gibbs sampler. These algorithms are used to calculate an approximate sample from a distribution that cannot normally be sampled from.

There are many practical applications of the methods described in this report, including modelling the spread of smallpox (Stockdale et al., 2017), antibiotic-resistant microbes (Kypraios et al., 2010), parasitic diseases (Chapman et al., 2018), influenza (Osthus et al., 2017), invasive species (Cook et al., 2007), and COVID-19 (Mbuvha and Marwala, 2020).

In Section 2 we introduce the standard stochastic SIR model, simulate from the model, and derive the likelihood of the times of infection and recovery. In Section 3 we introduce Markov chain Monte Carlo methods, with details on the Metropolis-Hastings algorithm and the Gibbs sampler. In Section 4 we use Markov chain Monte Carlo for Bayesian inference on the SIR model with missing data, and use this method on a simulated dataset. Extensions to the SIR model are discussed in Section 5, including a variable infection rate and a non-fixed population. In Section 6 we explore current open research areas, and we conclude with a discussion in Section 7.

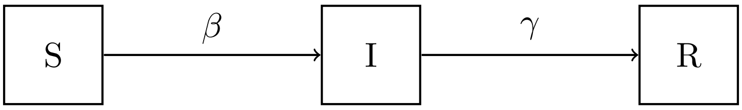
2 The SIR Model

2.1 Defining the Model

We first investigate a simple model: the *standard stochastic SIR epidemic*, as defined in Chapter 7 of Held et al. (2019). This model considers the spread of a pathogen through a population where we assume all individuals are potentially vulnerable to the pathogen. We also assume the population has a fixed size N – this means there is no immigration or emigration, no births, and no deaths (other than from the pathogen).

We split the population into three categories:

- Susceptible (S) – individuals who have not yet been infected with the pathogen.
- Infective (I) – individuals who currently are infected with the pathogen and can transmit it to others.
- Recovered (R) – individuals who have previously been infected with the pathogen, but no longer transmit it, and are now immune. This includes both individuals who have recovered to a healthy state and individuals who have died from the pathogen.



Let $S(t)$, $I(t)$, and $R(t)$ be the number of individuals at time t who are susceptible, infective, and recovered respectively. Unless otherwise specified, we start at $t = 0$ with one infective individual and all other individuals susceptible: $S(0) = N - 1$, $S(0) = 1$, and $R(0) = 0$. Since the total population is fixed, we have $S(t) + I(t) + R(t) = N$ for all t.

Each susceptible individual moves from state S to state I at constant rate β (for each infective individual they contact), and each infective individual moves from state I to state R at constant rate γ. This means the overall rate of change between S and I at time t is $\beta S(t)I(t)/N$, and the overall rate of change between I and R at time t is $\gamma I(t)$. This corresponds to the *deterministic general epidemic* defined by the following differential equations:

$$\begin{aligned}
\frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\frac{\beta S(t)I(t)}{N}, \\
\frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \frac{\beta S(t)I(t)}{N} - \gamma I(t), \\
\frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma I(t).
\end{aligned}$$

We assume both the time until infection and the time between infection and recovery are exponentially distributed. This means that at any time t, the time until the next infection is distributed as $\mathrm{Exp}\left(\frac{\beta S(t)I(t)}{N}\right)$, and the time until the next recovery is distributed as $\mathrm{Exp}(\gamma I(t))$. The model reaches a steady state when $I(t) = 0$, since there are no more infective individuals to spread the pathogen.

We also define the *basic reproduction number* $R_0 := \beta/\gamma$. Since each individual infectious period is exponentially distributed with rate γ, the average length of infection is $1/\gamma$, and so R_0 is the average number of infections each infective individual causes during their infectious period. This means a major outbreak will only occur if $R_0 > 1$.

2.2 Simulating from the Model

We can simulate from the SIR model using Gillespie's algorithm, originally formulated in Gillespie (1976). At each time when an event occurs, we generate the time until the next infection from $\text{Exp}\left(\frac{\beta S(t)I(t)}{N}\right)$ and the time until the next recovery from $\text{Exp}(\gamma I(t))$. The event that corresponds to the lesser time is is selected as the event which occurs next, and then time moves on by that time amount.

Algorithm 1 SIR Model Simulation

```
Require: Population size $N \in \mathbb{N}$, transmission rate $\beta \in (0, \infty)$, recovery rate $\gamma \in (0, \infty)$.
Let $t = 0$.
Let $S(0) = N - 1$, $I(0) = 1$, $R(0) = 0$.
while $I(t) \neq 0$ do
    Generate $t_i$ from $\text{Exp}\left(\frac{\beta S(t)I(t)}{N}\right)$.
    Generate $t_r$ from $\text{Exp}(\gamma I(t))$.
    if $t_i < t_r$ then
        Let $S(t + t_i) = S(t) - 1$ and $I(t + t_i) = I(t) + 1$.
        Let $t = t + t_i$.
    end if
    if $t_r < t_i$ then
        Let $I(t + t_r) = I(t) - 1$ and $R(t + t_r) = R(t) + 1$.
        Let $t = t + t_r$.
    end if
end while
```

We know that the minimum of two exponential random variables with rates λ_1 and λ_2 is also an exponential random variable, with rate $\lambda_1 + \lambda_2$. This means we can change our algorithm: at each time when an event occurs, we generate the time until the next event from

$$\text{Exp}\left(\frac{\beta S(t)I(t)}{N} + \gamma I(t)\right),$$

and determine the value of α, where

$$\alpha = \frac{\frac{\beta S(t)I(t)}{N}}{\frac{\beta S(t)I(t)}{N} + \gamma I(t)}.$$

The type of event that occurs is an infection event with probability α, and a recovery event otherwise; then time moves on by the generated amount.

Algorithm 2 SIR Model Simulation

```
Require: Population size $N \in \mathbb{N}$, transmission rate $\beta \in (0, \infty)$, recovery rate $\gamma \in (0, \infty)$.
Let $t = 0$.
Let $S(0) = N - 1,\ I(0) = 1,\ R(0) = 0$.
while $I(t) \neq 0$ do
    Generate $t^*$ from Exp $\left(\frac{\beta S(t)I(t)}{N} + \gamma I(t)\right)$.
    Let $\alpha = \left(\frac{\beta S(t)I(t)}{N}\right) / \left(\frac{\beta S(t)I(t)}{N} + \gamma I(t)\right)$.
    Generate $u$ from $U(0, 1)$.
    if $u < \alpha$ then
        Let $S(t + t^*) = S(t) - 1$ and $I(t + t^*) = I(t) + 1$.
    end if
    if $u > \alpha$ then
        Let $I(t + t^*) = I(t) - 1$ and $R(t + t^*) = R(t) + 1$.
    end if
    Let $t = t + t^*$.
end while
```

Algorithm 2 is an improvement over Algorithm 1 because we only sample from an exponential distribution once per iteration (rather than twice). Sampling from the exponential distribution using the inverse c.d.f. method involves taking the logarithm of a generated value. This adds a very small amount of computation time per sample, but can become large depending on the amount of time needed to simulate from the model. This means the computation time of any simulation is less when using Algorithm 2.

2.3 Simulated Example

We now use Algorithm 2 to simulate the SIR model for a fixed population of 100 individuals. We let $\beta = 2$ and $\gamma = 1$. In Figure 1 we see how $S(t)$, $I(t)$, and $R(t)$ change over time. The final state of the system is reached at $t = 9.09$ (to 3s.f.), with 76 recovered individuals, and 26 individuals who were never infected.

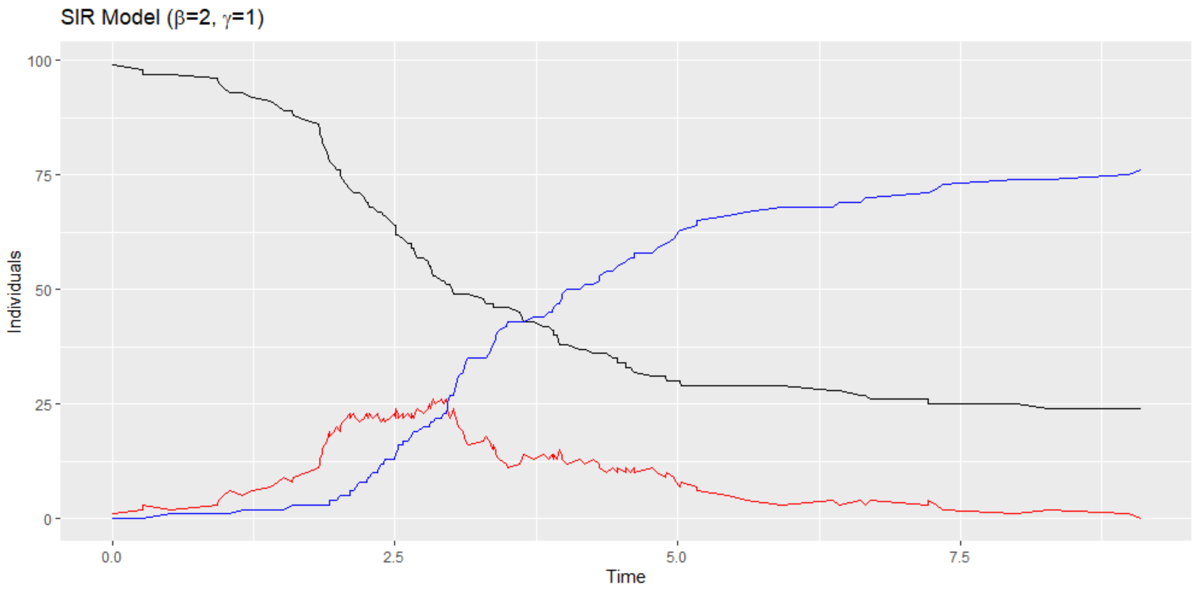


Figure 1: Simulation of the SIR model, with $S(t)$ in black, $I(t)$ in red, and $R(t)$ in blue.

2.4 Event Times Likelihood

Consider the standard stochastic SIR epidemic on a population with fixed size N, of which n individuals have become infective and recovered when the model reaches its steady state (of no infective individuals). We let $\mathbf{i}$ be the set of infection event times, $\mathbf{r}$ be the set of recovery event times, and $\mathbf{t}$ be the set of both infection and recovery event times. We order the recovery times such that $r_1 \leq r_2 \leq \cdots \leq r_N$, and let i_j be the infection time for the individual who recovers at time r_j. We start the model at the moment of the first infection, so $i_1 = 0$. If $n < N$, then not all individuals become infective, so we let $i_{n+1}, \dots, i_N, r_{n+1}, \dots, r_N = \infty$.

We now derive the joint likelihood of $\mathbf{i}$ and $\mathbf{r}$, given parameters β and γ,

$$L(\mathbf{i}, \mathbf{r}|\beta, \gamma) = \prod_{j=1}^{n} L(i_j|\beta, \gamma) \prod_{j=1}^{n} L(r_j|\beta, \gamma).$$

First we consider the recovery part of the likelihood. We know that the length of the infection period for each individual j is $r_j - i_j$, and that this time period is exponentially distributed with rate γ. Using the p.d.f. of the exponential distribution we have

$$\begin{aligned}\prod_{j=1}^{n} L(r_j|\beta, \gamma) &= \prod_{j=1}^{n} \left(\gamma \exp\left(-\gamma(r_j - i_j)\right)\right), \\ &= \gamma^n \exp\left(-\gamma \sum_{j=1}^{n} (r_j - i_j)\right).\end{aligned}$$

Now we consider the infection part of the likelihood. We know that $i_1 = 0$, so we only need to consider the likelihood for $j \geq 2$. We know that the length of the pre-infection period for each individual j is i_j. By considering the number of infectious individuals throughout the pre-infection period, the total exposure-time is $\sum_{k:t_k \leq i_j} I(t_{k-1})(t_k - t_{k-1})$. Using the p.d.f. of the exponential distribution, we have

$$\begin{aligned}\prod_{j=1}^{n} L(i_j|\beta, \gamma) &= \prod_{j=2}^{n} \left(\frac{\beta}{N}(I(i_j) - 1) \exp\left(-\frac{\beta}{N}(I(i_j) - 1) \sum_{k:t_k \leq i_j} I(t_{k-1})(t_k - t_{k-1})\right)\right), \\ &= \frac{\beta^{n-1}}{N^{n-1}} \left(\prod_{j=2}^{n} (I(i_j) - 1)\right) \exp\left(-\frac{\beta}{N}\sigma(\mathbf{i}, \mathbf{r})\right),\end{aligned}$$

where

$$\sigma(\mathbf{i}, \mathbf{r}) = \sum_{j=1}^{n} \sum_{k=1}^{N} \left(\min\{r_j, i_k\} - \min\{i_j, i_k\}\right),$$

the total exposure-time of all individuals pre-infection. This formula is derived from individual j can only being able to infect individual k for the length of time where both j is infective $(i_j < t < r_j)$ and k has not yet been infected $(t < i_k)$, summed over all j and k.

This means the overall likelihood is

$$L(\mathbf{i}, \mathbf{r}|\beta, \gamma) = \frac{\beta^{n-1}}{N^{n-1}} \gamma^n \left(\prod_{j=2}^{n} (I(i_j) - 1)\right) \exp\left(-\frac{\beta}{N}\sigma(\mathbf{i}, \mathbf{r}) - \gamma \sum_{j=1}^{n} (r_j - i_j)\right),$$

as seen in Chapter 9 of Held et al. (2019)

However, in most real-world cases when we have data of the recovery times $\mathbf{r}$, we will not have data on the infection times $\mathbf{i}$. Instead, we must treat $\mathbf{i}$ as an additional parameter to be estimated, as developed in O'Neill and Roberts (1999).

We now use a Bayesian approach to find the joint posterior distribution $\pi(\mathbf{i}, \beta, \gamma|\mathbf{r})$. Using Bayes' rule, we have

$$\pi(\mathbf{i}, \beta, \gamma|\mathbf{r}) \propto L(\mathbf{i}, \mathbf{r}|\beta, \gamma)\pi_0(\beta, \gamma),$$

where $\pi_0(\beta, \gamma)$ is the joint prior distribution of β and γ. We will assume β and γ have independent prior distributions, both following a gamma distribution:

$$\begin{aligned} \beta &\sim \text{Gamma}(a_\beta, b_\beta), \\ \gamma &\sim \text{Gamma}(a_\gamma, b_\gamma). \end{aligned}$$

This means we have joint posterior distribution

$$\begin{aligned} \pi(\mathbf{i}, \beta, \gamma|\mathbf{r}) &\propto L(\mathbf{i}, \mathbf{r}|\beta, \gamma)\pi_0(\beta)\pi_0(\gamma), \\ &\propto \beta^{n-1}\gamma^n \left(\prod_{j=2}^{n} (I(i_j) - 1)\right) \exp\left(-\frac{\beta}{N}\sigma(\mathbf{i}, \mathbf{r}) - \gamma\sum_{j=1}^{n}(r_j - i_j)\right) \beta^{a_\beta - 1} \exp\left(-b_\beta\beta\right) \gamma^{a_\gamma - 1} \exp\left(-b_\gamma\gamma\right), \\ &\propto \beta^{n+a_\beta-2}\gamma^{n+a_\gamma-1} \left(\prod_{j=2}^{n} (I(i_j) - 1)\right) \exp\left(-\frac{\beta}{N}\sigma(\mathbf{i}, \mathbf{r}) - \gamma\sum_{j=1}^{n}(r_j - i_j) - b_\beta\beta - b_\gamma\gamma\right). \end{aligned}$$

Clearly this is an intractable distribution that we cannot sample from, so we look to methods that can give an approximate sample.

3 Markov Chain Monte Carlo Methods

3.1 MCMC Introduction

Markov chain Monte Carlo (MCMC) is a collection of methods used to approximate a sample from a distribution that cannot normally be sampled from. In using MCMC methods we construct a Markov chain that converges to the target distribution $\pi(\boldsymbol{\theta})$. The states of this chain are then used as approximate samples to $\pi(\boldsymbol{\theta})$. The main source used for this section is Brooks (1998).

3.2 The Metropolis-Hastings Algorithm

The first MCMC method we investigate is the Metropolis-Hastings algorithm, originally developed in Metropolis et al. (1953) and Hastings (1970). In this algorithm, instead of sampling from the intractable $\pi(\boldsymbol{\theta})$, we repeatedly sample from a *proposal distribution* $q(\boldsymbol{\theta}|\boldsymbol{\eta})$, which is chosen such that it is easy to sample from. In each step k of the algorithm, we sample from $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{k-1})$, and let the generated value be $\boldsymbol{\theta}_k$ with probability α (otherwise, we let $\boldsymbol{\theta}_k$ be $\boldsymbol{\theta}_{k-1}$). We calculate α at each step based on π and q.

Algorithm 3 General Metropolis-Hastings

```
Require: Number of iterations $M$, initial value $\boldsymbol{\theta}_0$, proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\eta})$.
for $k \in \{1, \dots, M\}$ do
    Generate $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{k-1})$.
    Let
        $\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_{k-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{k-1})}\right\}.$
    With probability $\alpha$, let $\boldsymbol{\theta}_k = \boldsymbol{\theta}^*$. Otherwise, let $\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1}$.
end for
```

If the proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\eta})$ does not depend on $\boldsymbol{\eta}$, then we write $q(\boldsymbol{\theta}|\boldsymbol{\eta}) = q(\boldsymbol{\theta})$ and call this case the *independence sampler*. If $q(\boldsymbol{\theta}|\boldsymbol{\eta}) = q(\boldsymbol{\eta}|\boldsymbol{\theta})$, then the calculation of α simplifies to $\min\{1, \pi(\boldsymbol{\theta}^*)/\pi(\boldsymbol{\theta}_{k-1})\}$ – we call this case the *random walk Metropolis (RWM)*.

Let $K(\boldsymbol{\theta}, \boldsymbol{\eta})$ be the probability density of moving from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$ in the Markov chain constructed in this algorithm. Let $\alpha(\boldsymbol{\theta}, \boldsymbol{\eta})$ be the probability of accepting a move from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$. Then clearly we have $K(\boldsymbol{\theta}, \boldsymbol{\eta}) = q(\boldsymbol{\eta}|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}, \boldsymbol{\eta})$.

To prove that the Markov chain we have constructed has $\pi(\boldsymbol{\theta})$ as its stationary distribution, we will show that the *detailed balance* equation holds: for any points $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ we have

$$\pi(\boldsymbol{\theta})K(\boldsymbol{\theta}, \boldsymbol{\eta}) = \pi(\boldsymbol{\eta})K(\boldsymbol{\eta}, \boldsymbol{\theta}).$$

At each step the Metropolis-Hastings algorithm we have two cases: either we accept the generated move, or reject it. If we reject the move, then $\boldsymbol{\theta} = \boldsymbol{\eta}$, so clearly we have detailed balance. Otherwise, $\boldsymbol{\eta}$ is the accepted move from $\boldsymbol{\theta}$, and so we have

$$\begin{aligned}
\pi(\boldsymbol{\theta})K(\boldsymbol{\theta}, \boldsymbol{\eta}) &= \pi(\boldsymbol{\theta})q(\boldsymbol{\eta}|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}, \boldsymbol{\eta}), \\
&= \pi(\boldsymbol{\theta})q(\boldsymbol{\eta}|\boldsymbol{\theta}) \min\left\{1, \frac{\pi(\boldsymbol{\eta})q(\boldsymbol{\theta}|\boldsymbol{\eta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\eta}|\boldsymbol{\theta})}\right\}, \\
&= \min\{\pi(\boldsymbol{\theta})q(\boldsymbol{\eta}|\boldsymbol{\theta}), \pi(\boldsymbol{\theta})q(\boldsymbol{\eta}|\boldsymbol{\theta})\}, \\
&= \pi(\boldsymbol{\eta})K(\boldsymbol{\eta}, \boldsymbol{\theta}),
\end{aligned}$$

where the final line follows from the symmetry of taking the minimum of two numbers. This means the Markov chain in the Metropolis-Hastings algorithm satisfies detailed balance, and so has $\pi(\boldsymbol{\theta})$ as its stationary distribution as required.

3.3 The Gibbs Sampler

The other MCMC method we investigate is the Gibbs sampler, as developed in Geman and Geman (1984). In the Gibbs sampler, we partition our parameters of interest $\boldsymbol{\theta}$ into blocks $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)})$ and calculate the conditional distribution of each block $\pi(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(\text{not } j)})$, such that it is possible to sample from each conditional distribution (even though we cannot sample from $\pi(\boldsymbol{\theta})$). By repeatedly sampling from the conditional distributions (dependent on the most recently sampled values from the other blocks) we generate an approximate sample of $\boldsymbol{\theta}$.

Algorithm 4 General Gibbs Sampler

```
Require: Number of iterations $M$, initial value $\boldsymbol{\theta}_0$.
Partition $\boldsymbol{\theta}$ into blocks such that $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)})$.
for $k \in \{1, \ldots, M\}$ do
    for $j \in \{1, \ldots, m\}$ do
        Generate $\boldsymbol{\theta}^{(j)*}$ from $\pi(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}_k^{(1)}, \ldots, \boldsymbol{\theta}_k^{(j-1)}, \boldsymbol{\theta}_{k-1}^{(j+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(m)})$.
        Let $\boldsymbol{\theta}_k^{(j)} = \boldsymbol{\theta}^{(j)*}$.
    end for
end for
```

The Gibbs sampler can be considered a special case of the Metropolis-Hastings algorithm. When updating block j, we have proposal distribution $q(\boldsymbol{\theta}^{(j)*}|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}^{(j)*}|\boldsymbol{\theta}^{(\text{not } j)})$, and so

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{\pi(\boldsymbol{\theta}^{(j)*}|\boldsymbol{\theta}^{(\text{not } j)})q(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{*})}{\pi(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(\text{not } j)})q(\boldsymbol{\theta}^{(j)*}|\boldsymbol{\theta})}\right\}, \\
&= \min\left\{1, \frac{\pi(\boldsymbol{\theta}^{(j)*}|\boldsymbol{\theta}^{(\text{not } j)})\pi(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(\text{not } j)})}{\pi(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(\text{not } j)})\pi(\boldsymbol{\theta}^{(j)*}|\boldsymbol{\theta}^{(\text{not } j)})}\right\}, \\
&= 1.
\end{aligned}$$

This means we always accept the value generated from the proposal distribution, as in the Gibbs sampler presented above. Since we have shown that the Metropolis-Hastings algorithm has stationary distribution $\pi(\boldsymbol{\theta})$, this means the Gibbs sampler also has stationary distribution $\pi(\boldsymbol{\theta})$.

4 Bayesian Inference on the SIR Model

4.1 MCMC Algorithm

We now use these MCMC methods to construct an algorithm for inference on the SIR model (as seen in Chapter 9 of Held et al. (2019)). Based on the joint posterior distribution calculated earlier, the conditional posterior distributions of β, γ, and $\mathbf{i}$ are:

$$\begin{aligned}
\pi(\beta|\mathbf{i}, \mathbf{r}, \gamma) &\propto \beta^{n+a_\beta-2} \exp\left(-\beta\left(\frac{1}{N}\sigma(\mathbf{i}, \mathbf{r}) + b_\beta\right)\right), \\
\pi(\gamma|\mathbf{i}, \mathbf{r}, \beta) &\propto \gamma^{n+a_\gamma-1} \exp\left(-\gamma\left(\sum_{j=1}^{n}(r_j - i_j) + b_\gamma\right)\right), \\
\pi(\mathbf{i}|\mathbf{r}, \beta, \gamma) &\propto \left(\prod_{j=2}^{n}(I(i_j) - 1)\right) \exp\left(-\frac{\beta}{N}\sigma(\mathbf{i}, \mathbf{r}) - \gamma\sum_{j=1}^{n}(r_j - i_j)\right).
\end{aligned}$$

Clearly the conditional distributions for β and γ have the shape of a gamma distribution, so we have

$$\begin{aligned}
\beta|\mathbf{i}, \mathbf{r}, \gamma &\sim \text{Gamma}\left(n + a_\beta - 1, \frac{1}{N}\sigma(\mathbf{i}, \mathbf{r}) + b_\beta\right), \\
\gamma|\mathbf{i}, \mathbf{r}, \beta &\sim \text{Gamma}\left(n + a_\gamma, \sum_{j=1}^{n}(r_j - i_j) + b_\gamma\right).
\end{aligned}$$

Unfortunately, it is still not possible to sample from the distribution of $\mathbf{i}$. Instead we will use a Metropolis-Hastings step. We know that each infection period is exponentially distributed with rate γ, and ends at a known recovery time. For each individual j, we can generate a new length of the infective period x from $X \sim \text{Exp}(\gamma)$, and let our new infection time be $i_j^* = r_j - x$. Let $\mathbf{i}^*$ be equal to the current value of $\mathbf{i}$, but replacing i_j with i_j^*. We then accept i_j^* to replace the current i_j with probability α, where

$$\alpha = \min\left\{1, \frac{\pi(\mathbf{i}^*|\mathbf{r}, \beta, \gamma)f_X(r_j - i_j)}{\pi(\mathbf{i}|\mathbf{r}, \beta, \gamma)f_X(r_j - i_j^*)}\right\}.$$

This means our overall algorithm uses a Gibbs sampler, with one part of each iteration using a Metropolis-Hastings step:

Algorithm 5 Inference for the SIR model using MCMC

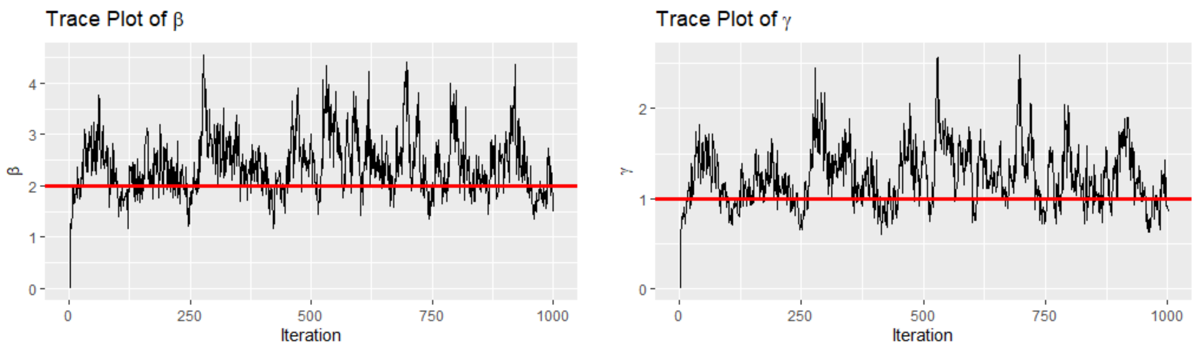
```
Require: Number of iterations $M$, population size $N$, recovery times $\mathbf{r}$, initial value $\beta_0$, initial value $\gamma_0$, initial values $\mathbf{i}_0$.
for $k \in \{1, \dots, M\}$ do
    Generate $\beta_k$ directly from $\pi(\beta|\mathbf{i}_{k-1}, \mathbf{r}, \gamma_{k-1})$.
    Generate $\gamma_k$ directly from $\pi(\gamma|\mathbf{i}_{k-1}, \mathbf{r}, \beta_k)$.
    for $j = 1$ to $N$ do
        Generate $x$ from $X \sim \text{Exp}(\gamma_k)$.
        Let $\mathbf{i}^* = \mathbf{i}_{k-1}$, and then let $i_j^* = r_j - x$.
        Let
            $\alpha = \min\left\{1, \frac{\pi(\mathbf{i}^*|\mathbf{r}, \beta_k, \gamma_k) f_X(r_j - i_j)}{\pi(\mathbf{i}_{k-1}|\mathbf{r}, \beta_k, \gamma_k) f_X(r_j - i_j^*)}\right\}.$
        With probability $\alpha$, let $i_{j,k} = i_j^*$. Otherwise let $i_{j,k} = i_{j,k-1}$.
    end for
end for
```

4.2 Simulated Example

We now use Algorithm 5 to estimate β, γ, and $\mathbf{i}$, given the recovery times from the data simulated in Section 2 (which was simulated with parameters $\beta = 2$ and $\gamma = 1$). The population contains 100 individuals, 76 of whom have finite recovery times. This means only 76 individuals ever become infected, so $N = 100$ and $n = 76$. We assume we are working in a scenario where we have the recovery time data $\mathbf{r}$, but no knowledge of β, γ, or $\mathbf{i}$.

Without any knowledge of β or γ, we just let $\beta_0 = \gamma_0 = 0$. We let the initial infection time for individual j be halfway between the times 0 and r_j, so we have $\mathbf{i}_0 = \mathbf{r}/2$. We also choose flat, high variance priors for β and γ – we let $a_\beta = a_\gamma = 1$ and $b_\beta = b_\gamma = 10^{-4}$. These are the same priors used in Chapter 9 of Held et al. (2019).

We run the Markov chain for 1000 iterations, producing the following trace plots in Figure 2. Obviously we want to investigate the trace plots of β and γ, but it is not reasonable to produce plots of all 76 infection times. Instead we have traces plots of one infection time (arbitrarily chosen to be i_{51}) and the sum of all infection times $\sum_{j=1}^{76} i_j$. The true values of all parameters are shown in red. We see that the four chains below all mix well, although they seem to spend more time above the true value than below it.



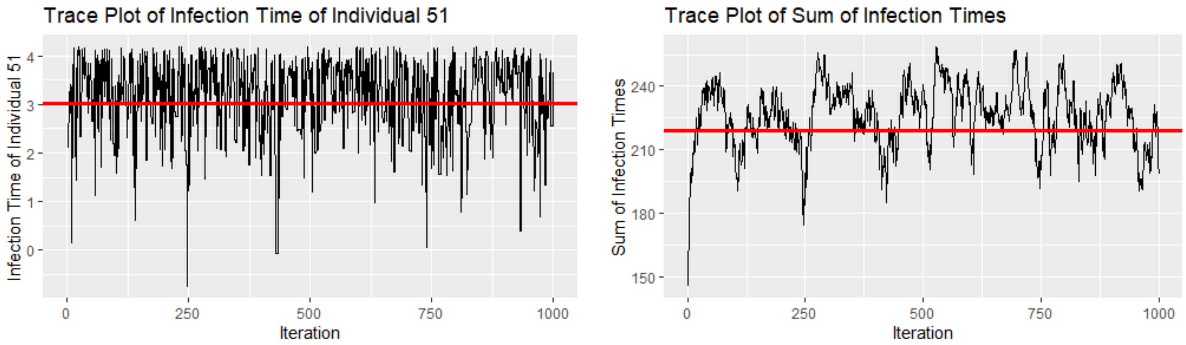
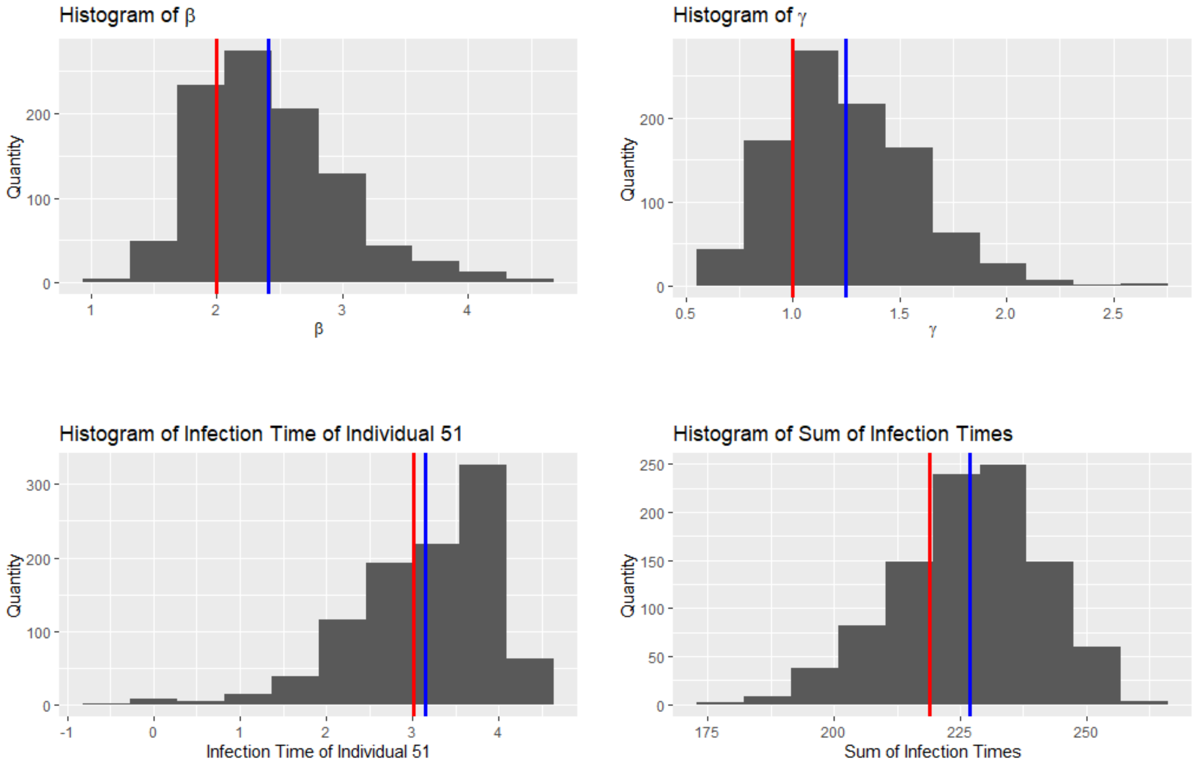


Figure 2: Trace plots for the Markov chains to sample for β, γ, i_{51}, and $\sum_{j=1}^{76} i_j$, with true values in red.

We also produce the following histograms for β, γ, i_{51}, and $\sum_{j=1}^{76} i_j$ in Figure 3. Since the basic reproduction number R_0 is equal to β/γ, we can easily calculate a vector of values for R_0 by $R_{0,k} = \beta_k/\gamma_k$ for k from 1 to 1000. This means we can also plot a histogram for R_0 in Figure 3. The first 20 iterations of all parameters are removed from the data, since it takes several iterations for the Markov chain to converge (this removed data is called *burn-in*, and the point when burn-in ends is generally chosen by eye).

The mean values for β and γ were 2.41 and 1.25 respectively (to 3 s.f.), compared with the true values of 2 and 1. The mean value for R_0 was 1.97 (to 3s.f.), compared with the true value of 2. This means the estimates for β and γ over-estimate by $20 - 25\%$, but that the estimate for R_0 is correct within 5%. Again the true values of all parameters are shown in red, with the means of the values generated by the Markov chain shown in blue.



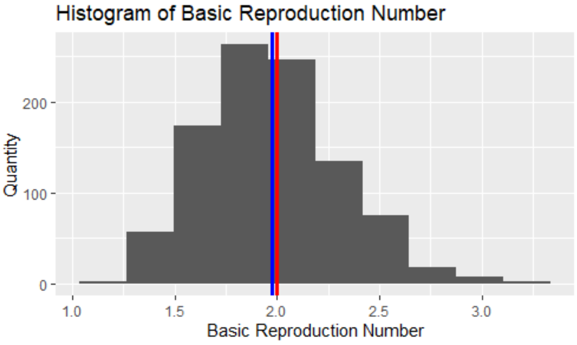


Figure 3: Histograms for the samples of β, γ, i_{51}, $\sum_{j=1}^{76} i_j$, and R_0, with true values in red and average values in blue.

5 Model Extensions

5.1 Different Infectious Period Distributions

We now look to improve the SIR model by removing previous assumptions and generalising the model for a greater number of situations. First, we investigate infection period distributions other than the exponential distribution.

Our previous assumption that the infectious period is exponentially distributed was somewhat simplistic; the memorylessness property means that the probability of recovery in the same at any given time after infection, and only having one parameter means that the mean and variance are not independent of each other. A natural extension of the gamma distribution or the Weibull distribution are suggested in Held et al. (2019) – both have more than one parameter, and are part of the exponential family of distributions. Having multiple parameters means that the mean and the variance of the infectious period can be determined separately.

This means the recovery times part of the likelihood of **i** and **r** becomes

$$\prod_{j=1}^{n} L(r_j|\beta, \boldsymbol{\theta}) = \prod_{j=1}^{n} f_X(r_j - i_j|\boldsymbol{\theta}),$$

where X is the distribution of each infectious period (with parameters $\boldsymbol{\theta}$). During Bayesian inference we now also sample for all parameters $\boldsymbol{\theta}$.

5.2 Variable Infection Rate

Previously we have assumed that each individual is infected at constant rate β for each currently infective individual. This is often not a reasonable assumption: each individual in the population will have more contact with certain individuals (family, friends, etc.) than others, and not everyone will pass on the pathogen at the same rate (for example, those with weaker immune systems may be more venerable). We will now relax this assumption, and define a different rate $\beta_{i,j}$ for each pair of individuals i and j.

If the infectious period has distribution X with parameters $\boldsymbol{\theta}$, then from Chapter 9 of Held et al. (2019)

the likelihood of $\mathbf{i}$ and $\mathbf{r}$ is

$$\begin{aligned}L(\mathbf{i},\mathbf{r}|\beta,\boldsymbol{\theta}) &= \prod_{j=1}^{n} L(i_j|\beta,\boldsymbol{\theta})L(r_j|\beta,\boldsymbol{\theta}),\\ &= \left(\prod_{j=2}^{n}\rho(j,\beta,\mathbf{i},\mathbf{r})\right)\exp\left(-\sigma_*(\beta,\mathbf{i},\mathbf{r})\right)\prod_{j=1}^{n} f_X(r_j-i_j|\boldsymbol{\theta}),\end{aligned}$$

where

$$\rho(j,\beta,\mathbf{i},\mathbf{r}) = \sum_{k:i_k<i_j<r_k}\beta_{k_j},$$

the total rate of infection towards individual j just before being infected, and where

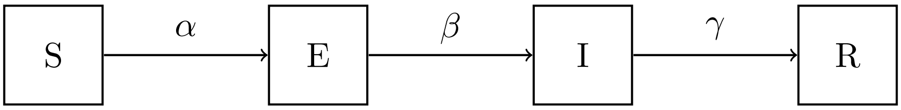
$$\sigma_*(\beta,\mathbf{i},\mathbf{r}) = \sum_{j=1}^{n}\sum_{k=1}^{N}\beta_{j,k}\left(\min\{r_j,i_k\}-\min\{i_j,i_k\}\right),$$

the total rate of infection over time towards all individuals not yet infected.

Generally we are interested in $\beta_{i,j}$ when it has some known form based on i and j. For example, in Chapman et al. (2018), the transmission rate is scaled based on the distance between the households of individuals i and j (both exponential decay and Cauchy-style decay are tested). We can take this approach further by also letting the transmission rate change through time, such as in Kypraios et al. (2010), although we will not cover the details in this report.

5.3 Exposed Period (SEIR Model)

Another extension we can make to the model is the addition of an 'exposed' (E) category, for individuals who have been infected with the pathogen, but cannot yet transmit it to others. This is a common property of real-world pathogens. After adding this category we now have the *SEIR model.*



Instead of a transmission rate β from S to I, we now have a constant exposure rate α from S to E, and a constant transmission rate β from E to I. This means we have the following differential equations:

$$\begin{aligned}\frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\frac{\alpha S(t)I(t)}{N},\\ \frac{\mathrm{d}E(t)}{\mathrm{d}t} &= \frac{\alpha S(t)I(t)}{N} - \beta E(t),\\ \frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \beta E(t) - \gamma I(t),\\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma I(t).\end{aligned}$$

Let $\mathbf{e}$ be the set of times when individuals first become exposed to the pathogen, $\mathbf{i}$ be the set of times when individuals first become infectious, and $\mathbf{r}$ be the set of times when individuals recover. Instead of exponential distribution rates β and γ, let X be the distribution of each infectious period (with parameters $\boldsymbol{\theta}$) and Y be the distribution of each exposure period (with parameters $\boldsymbol{\eta}$). Then the joint likelihood of $\mathbf{e}$, $\mathbf{i}$, and $\mathbf{r}$ is

$$\begin{aligned} L(\mathbf{e},\mathbf{i},\mathbf{r}|\alpha,\boldsymbol{\theta},\boldsymbol{\eta}) &= \prod_{j=1}^{n} L(e_j|\alpha,\boldsymbol{\theta},\boldsymbol{\eta})L(i_j|\alpha,\boldsymbol{\theta},\boldsymbol{\eta})L(r_j|\alpha,\boldsymbol{\theta},\boldsymbol{\eta}), \\ &= \frac{\beta^{n-1}}{N^{n-1}}\left(\prod_{j=2}^{n}(I(e_j)-1)\right)\exp\left(-\frac{\beta}{N}\sigma_E(\mathbf{e},\mathbf{i},\mathbf{r})\right)\prod_{j=1}^{n} g_Y(i_j-e_j|\boldsymbol{\eta})f_X(r_j-i_j|\boldsymbol{\theta}), \end{aligned}$$

where

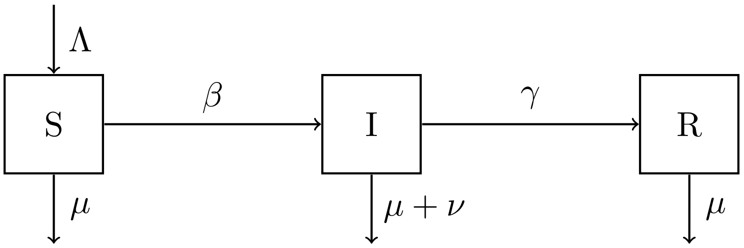
$$\sigma_E(\mathbf{e},\mathbf{i},\mathbf{r}) = \sum_{j=1}^{n}\sum_{k=1}^{N}\left(\min\{r_j,e_k\}-\min\{i_j,e_k\}\right),$$

the total exposure-time of all individuals before being exposed themselves.

From this likelihood we can calculate the joint posterior distribution of $\mathbf{e}$, $\mathbf{i}$, α, $\boldsymbol{\theta}$, and $\boldsymbol{\eta}$ given $\mathbf{r}$, and then use a similar MCMC algorithm to Algorithm 5 to approximate values from this distribution. However, this means that for each recovery time r_j we are estimating both an exposure time e_j and infection time i_j – the model is over-parameterised, and so will lead to correlation between parameters. To deal with this issue with our MCMC algorithm we will need significant prior information about the exposure period and/or the infectious period (so that the model can determine the difference between them). Alternatively, we could use particle MCMC methods that can account for the correlation between parameters, such as in Rosato et al. (2022).

5.4 Non-Fixed Population

We now return to the SIR model. We have previously assumed a fixed population of size N, with no changes in population due to immigration, emigration, births, or deaths (other than deaths from the pathogen, which are counted as recoveries). This is an acceptable assumption if we are modelling a short time frame and/or an isolated community, but breaks down over longer periods of time. We now add to the model the possibility of births (with rate λ), natural deaths (with rate μ), and deaths due to the pathogen (with rate ν, no longer counted as recoveries). We assume the size of the population fluctuates around N.



This means we have the following system of differential equations:

$$\begin{aligned} \frac{\mathrm{d}S(t)}{\mathrm{d}t} &= \lambda N - \frac{\beta S(t)I(t)}{N} - \mu S(t), \\ \frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \frac{\beta S(t)I(t)}{N} - \gamma I(t) - (\mu+\nu)I(t), \\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma I(t) - \mu R(t). \end{aligned}$$

Again we have we have a significantly greater number of parameters, so we may need methods that can account for the correlation between parameters, such as particle MCMC.

5.5 Other Extensions

There are many other extensions we can make to the SIR model. We can add the possibility of relapsing (returning to the infected category after recovering), as in Chapman et al. (2018), or add the possibility of losing immunity to the pathogen sometime after recovering (and returning to the susceptible category, also known as the SIRS model). Another possible extension is to add changepoints to the model, in order to accommodate sudden changes in model parameters (such as the transmission rate of the pathogen). This was used to account for new variants of COVID-19 in Gu and Yin (2022). We can also relax the assumption of a homogeneously mixing population and consider a network model, such as in Altarelli et al. (2014) and Britton and O'Neill (2002).

6 Open Problems

We now briefly explore open problems in the area of stochastic epidemic modelling. A good summary of these open challenges can be found on Britton et al. (2014), which is what we use in this section. Pellis et al. (2015) is also a good resource for open problems specifically for stochastic epidemic modelling using network models.

A key area for further study is the emergence of endemic behaviour – what is the probability the pathogen survives after the initial epidemic stage and becomes endemic. Approximate solutions are provided in van Herwaarden (1997) and Meerson and Sasorov (2009), but still more progress needs to made (specifically for situations more complex than the standard SIR model).

We have previously assumed that population is constant (or approximately constant over time). However, this assumption is not accurate if the overall birth and death rates are different; in this case the average size of the population will be changing over time. This model has been studied in Britton and Trapman (2014).

Another important area for further study is how to account for the mutation of pathogens in stochastic epidemic models. One method that has been used is the application of changepoints to the transmission rate in Gu and Yin (2022).

7 Conclusion

In this report, we have introduced the SIR model, a compartmental model for the spread of a pathogen through a population. The Gillespie algorithm was used to construct an algorithm for simulating from the model (which we then did). We also introduced two Markov chain Monte Carlo methods for generating an approximate sample from a given distribution: the Metropolis-Hastings algorithm, and the Gibbs sampler. We then applied these MCMC methods to the SIR model (with missing data) to construct an algorithm to perform inference for our model parameters, and used this algorithm on our simulated dataset. After running the Markov chain for 1000 iterations we generated reasonable estimates of the model parameters. We also investigated a variety of improvements that can be made to the standard SIR model, and explored open research areas in stochastic epidemic modelling.

Code Availability

The code used to produce the results in Sections 2 and 4 can be found at the following GitHub page: `https://github.com/neilljn/RT2`

References

Allen, L. J. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(2):128–142.

Altarelli, F., Braunstein, A., Dall'Asta, L., Lage-Castellanos, A., and Zecchina, R. (2014). Bayesian inference of epidemics on networks via belief propagation. *Physical review letters*, 112(11):118701.

Britton, T., House, T., Lloyd, A. L., Mollison, D., Riley, S., and Trapman, P. (2014). Eight challenges for stochastic epidemic models involving global transmission.

Britton, T. and O'Neill, P. D. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29(3):375–390.

Britton, T. and Trapman, P. (2014). Stochastic epidemics in growing populations. *Bulletin of mathematical biology*, 76:985–996.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)*, 47(1):69–100.

Chapman, L. A., Jewell, C. P., Spencer, S. E., Pellis, L., Datta, S., Chowdhury, R., Bern, C., Medley, G. F., and Hollingsworth, T. D. (2018). The role of case proximity in transmission of visceral leishmaniasis in a highly endemic village in Bangladesh. *PLoS neglected tropical diseases*, 12(10):e0006453.

Cook, A., Marion, G., Butler, A., and Gibson, G. (2007). Bayesian inference for the spatio-temporal invasion of alien species. *Bulletin of mathematical biology*, 69:2005–2025.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434.

Gu, J. and Yin, G. (2022). Bayesian SIR model with change points with application to the omicron wave in Singapore. *Scientific Reports*, 12(1):20864.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Held, L., Hens, N., D O'Neill, P., and Wallinga, J. (2019). *Handbook of Infectious Disease Data Analysis.* CRC Press.

Kypraios, T., O'Neill, P. D., Huang, S. S., Rifas-Shiman, S. L., and Cooper, B. S. (2010). Assessing the role of undetected colonization and isolation precautions in reducing methicillin-resistant Staphylococcus aureus transmission in intensive care units. *BMC infectious diseases*, 10:1–10.

Mbuvha, R. and Marwala, T. (2020). Bayesian inference of COVID-19 spreading rates in South Africa. *PloS one*, 15(8):e0237126.

Meerson, B. and Sasorov, P. V. (2009). WKB theory of epidemic fade-out in stochastic populations. *Physical Review E*, 80(4):041130.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D., and Del Valle, S. Y. (2017). Forecasting seasonal influenza with a state-space SIR model. *The annals of applied statistics*, 11(1):202.

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

Pellis, L., Ball, F., Bansal, S., Eames, K., House, T., Isham, V., and Trapman, P. (2015). Eight challenges for network epidemic models. *Epidemics*, 10:58–62.

Rosato, C., Harris, J., Panovska-Griffiths, J., and Maskell, S. (2022). Inference of stochastic disease transmission models using particle-MCMC and a gradient based proposal. In *2022 25th International Conference on Information Fusion*, pages 1–8. IEEE.

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

van Herwaarden, O. A. (1997). Stochastic epidemics: the probability of extinction of an infectious disease at the end of a major outbreak. *Journal of mathematical biology*, 35:793–813.