Fast and efficient Bayesian inference for stochastic epidemic models

Samuel Whitaker

Thesis submitted for the degree of Doctor of Philosophy

School of Mathematics, Statistics & Physics

Newcastle University

Newcastle upon Tyne

United Kingdom

March 26, 2025

Acknowledgements

This PhD thesis marks the culmination of a journey that would not have been possible without the support, guidance, and encouragement of many individuals. I would like to express my deepest gratitude to everyone who has been part of this journey.

First and foremost, I would like to thank my supervisors, Andrew Golightly and Colin Gillespie. Your unwavering support, insightful feedback, and constant encouragement have been instrumental in shaping this thesis. A special thanks must go to Andy, without whom this work would not exist. Your (seemingly) endless patience has undoubtedly been tested multiple times throughout the last five years, from staring at R code to find a bug to replying to my (sometimes relentless) emails. You have always found time to help.

I am profoundly grateful to my family for their unconditional love and support. To my mum, thank you for instilling in me the value of education and hard work, without which I would not have been able to undertake this research. Your constant encouragement has been my driving force. To my brother, Jacob, thank you for always believing in me and cheering me on, for being the voice of reason when I needed one, and for being unapologetically honest when the truth was difficult to hear. To my devoted partner, Shaun, thank you for putting up with me through the latest of nights and the most stressful of days. The extent of your support is immeasurable, and for that, I am truly grateful. Finally, thanks must go to Laura and Peter for taking me in as one of your own and for being two of my biggest cheerleaders.

A special thank you to my friends and colleagues in the School of Mathematics, Statistics, and Physics at Newcastle University. Your camaraderie, stimulating discussions, and shared experiences have made this journey enjoyable and memorable.

To my friends outside of academia, thank you for providing balance in my life. From the ridiculous Zoom quizzes during the lockdowns to the countless hours spent watching football together, you have provided me with a much-needed escape from my research.

This thesis is dedicated to all those who have supported and believed in me. Thank you for being part of my journey.

Abstract

Epidemics are inherently stochastic in nature, and stochastic kinetic models (SKMs) provide an appropriate way to describe and analyse such phenomena. Given temporal data consisting of, for example, the number of new infections or removals in a given time window, a continuous-time discrete-valued Markov process provides a natural description of the dynamics of each model component, typically taken to be the number of susceptible, exposed, infected or removed individuals. Fitting the resulting SEIR model to time-course data is a challenging task due to the problem of partial observations and, consequently, the intractability of the observed data likelihood. Whilst sampling based inference schemes such as Markov chain Monte Carlo are routinely applied, their computational cost typically restricts analysis to data sets of no more than a few thousand infected cases. Moreover, upon receipt of new data, these schemes typically need to be restarted from scratch.

This thesis addresses these issues via two complementary approaches. First, we develop a sequential inference scheme that makes use of a computationally cheap approximation of the most natural Markov process model. Crucially, the resulting model allows a tractable conditional parameter posterior which can be summarised in terms of a set of low dimensional statistics. This is used to rejuvenate parameter samples in conjunction with a novel bridge construct for propagating state trajectories conditional on the next observation of cumulative incidence. The resulting inference framework also allows for stochastic infection and reporting rates. Second, we tackle the intractability of the observed data likelihood in a batch inference

setting. We adopt a stochastic differential equation (SDE) representation of the underlying epidemic dynamics by matching the infinitesimal mean and variance to the drift and diffusion coefficients of an Itô SDE. We then approximate the SDE to give a tractable Gaussian process, that is, the linear noise approximation (LNA). Unless the observation model linking the LNA to the data is both linear and Gaussian, the observed data likelihood remains intractable. To circumvent this, we marginalise over the latent process by enforcing a Gaussian approximation of the observation model and use a forward filter to efficiently calculate the resulting approximation of the observed data likelihood. The proposed inference methodology is illustrated using both real and synthetic data sets. Where possible, we compare against competing approaches.

Declaration

Parts of this thesis have been submitted for publication:

- Whitaker, S.A., Golightly, A., Gillespie, C.S., and Kypraios, T. 'Sequential Bayesian inference for stochastic epidemic models of cumulative incidence', arXiv preprint 2405.13537.
- Golightly, A., Wadkin, L.E., Whitaker, S.A., Baggaley, A.W., Parker, N.G., and Kypraios, T. 'Accelerating Bayesian inference for stochastic epidemic models using incidence data', *Statistics and Computing*, 2023. doi: 10.1007/s11222-023-10311-6.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of the spread of infectious diseases is important from both a social and economic perspective. Consequently, statistical epidemiology is a large and ever-growing field that combines statistical models with state-of-the-art inference techniques to analyse complex data arising from observational studies. Epidemics are inherently random, and capturing this aspect necessitates the use of stochastic models, which typically take the form of a continuous-time, discrete-valued Markov jump process (MJP), describing transitions between different states such as susceptible, exposed, infected, removed (SEIR, see e.g. Allen, 2017). Inference for the parameters governing such models is complicated by the availability of incomplete and imperfect observation regimes, as arising, for example, when only a subset of states or transitions are recorded at discrete times. This problem can be circumvented via the use of Markov chain Monte Carlo and data augmentation (see e.g. O'Neill and Roberts, 1999; Jewell et al., 2009). This involves updating parameters conditional on a complete representation of the times and types of events (e.g. infection, removal etc.), and then updating the times and types conditional on parameters

and observations. This requires a mechanism for sampling event times and types between observations; this is necessarily computationally expensive. Consequently, these schemes appear limited to applications where the total population size is of the order of a few thousand individuals (Stockdale et al., 2021). Therefore, a key challenge is the development of fast and reliable inference techniques to allow real-time decision-making (Swallow et al., 2022).

Computationally efficient approaches to inference for stochastic epidemic models include the use of approximate inference schemes (see e.g. Kypraios et al., 2017; McKinley et al., 2018; Minter and Retkute, 2019, for approximate Bayesian computation schemes) or direct approximation of the most natural MJP (see e.g. Cauchemez and Ferguson, 2008; Fuchs, 2013; Fintzi et al., 2022, for approximations based on stochastic differential equations) and direct approximation of the observed data likelihood (see e.g. Whiteley and Rimella, 2021; Golightly et al., 2023).

1.1 Thesis aims

This thesis makes two main contributions to existing literature on inference for stochastic epidemic models. The first contribution is to replace the MJP model of cumulative incidence with a model in which the number of transition events over a time interval whose length is chosen by the practitioner is assumed to be Poisson distributed. The resulting time discretised SEIR model (dSEIR) has a number of advantages over other approximate models, including recognition of the discrete stochastic nature of epidemic spread, easy incorporation of time varying parameters, specification of a time step that trades off accuracy and computational efficiency and, crucially, tractability of the conditional posterior for rate parameters, given a

particular choice of prior.

We fit the dSEIR model within a sequential Bayesian framework. That is, we update our inferences about the static parameters and latent dynamic process as each observation becomes available. The observations are noisy measurements of cumulative incidence, specifically, the number of new infections or removals in fixed-length time windows. The inference procedure uses sequential Monte Carlo (see e.g. Kantas et al., 2009; Chopin and Papaspiliopoulos, 2020) via recursive application of a series of propagate, weight, resample and rejuvenate steps. The latter is used to alleviate sample impoverishment of the static parameter particle set and leverages the tractability of the parameter posterior conditional on the latent dynamic process, summarised through a collection of summary statistics. This approach was first described in Storvik (2002) (see also Fearnhead, 2002) and is a key ingredient of particle learning (Carvalho et al., 2010), which has been used by Dukic et al. (2012) and Lin and Ludkovski (2014), among others, in the context of stochastic epidemic models. Our approach is most related to the latter, which considers a particle learning algorithm for the most natural MJP representation of an epidemic compartment model of prevalence data, coupled with a latent seasonal component. However, unlike their approach, our modelling framework allows for additional flexibility via the use of a stochastic differential equation (SDE) to describe time varying parameters (e.g. infection and/or reporting rates) and the specification of an observation model to allow for imperfect observations on cumulative incidence. Moreover, we adapt the bridge construct of Golightly and Wilkinson (2015) (see also Golightly and Kypraios, 2018, or in the context of batch methods see Pooley et al., 2015) to the context of cumulative incidence, so that particle trajectories are propagated conditional on the following observation. We illustrate the resulting inference scheme in three scenarios: the first uses synthetic data in order to assess the accuracy of

our approach for different numbers of particles (bench-marked against the output of a pseudo-marginal Metropolis-Hastings scheme (PMMH, e.g. Andrieu et al., 2010) while the two final scenarios consider the spread of Ebola in West Africa (Fintzi et al., 2022) and COVID-19 in New York (Spannaus et al., 2022).

The second contribution is an extension of the methodology found in Fintzi et al. (2022) by adopting a linear noise approximation of the latent cumulative incidence process. Whereas Fintzi et al. (2022) integrate over the uncertainty in the latent process via a sampling approach, we introduce a further Gaussian approximation of the observation model, allowing analytic integration of the latent incidence process. Our framework additionally allows for a time-varying infection rate, as is typically required for accounting for seasonality and/or interventions. The infection rate is modelled stochastically as an additional component in the system of stochastic differential equations, which the LNA approximates. Hence, our contribution is a fast and efficient sampling-based framework for inferring the parameters of a general class of stochastic epidemic models. We benchmark the performance of our approach against state-of-the-art correlated pseudo-marginal methods (Dahlin et al., 2015; Deligiannidis et al., 2018) in terms of both accuracy and efficiency, using a susceptible-infectious-removed (SIR) model. Finally, we consider the application of the methodology to the infestation of the oak processionary moth (OPM), *Thaumetopoea processionea*, in Richmond Park, London. Using time course data consisting of the yearly removal incidence of infested trees between the years 2013 and 2020, we compare and contrast an SIR model with a model in which infected trees can re-enter the susceptible class. The assumed initial population size is some 40,000 oak trees, with the number of susceptible trees reducing to around 35,000 over the time frame of the data set. The size of the epidemic necessitates analytic integration of the latent incidence process; discrete stochastic models combined with exact

(simulation-based) inference methods such as data augmentation (see e.g. Jewell et al., 2009) are practically infeasible here; see Stockdale et al. (2021) for a recent discussion.

1.2 Thesis outline

The remainder of this thesis is organised as follows. In Chapter 2 we review Monte Carlo methods for inference of a general target $\pi(\boldsymbol{\theta}|y)$, building up from stochastic simulation for estimation of expectations, to Markov chain Monte Carlo and (correlated) pseudo marginal Metropolis Hastings. We then consider the sequential use of Bayes theorem to introduce some sequential inference schemes. First, we consider the problem of state filtering where, given $y_{1:t} = (y_1, y_2, \ldots, y_t)$, we target the density $\pi(x_t|y_{1:t})$ using sequential importance (re)sampling and the bootstrap particle filter (Gordon et al., 1993). We then introduce schemes for state and parameter filtering, namely the Liu and West algorithm (Liu and West, 2001) and the Storvik particle filter (Storvik, 2002).

In Chapter 3 we introduce stochastic kinetic models, which we use to model the spread of disease through a population of fixed size. We begin by partitioning the population into a number of disjoint subsets, and describe the interactions between these subsets as a chemical reaction network. The S(E)IR model is then defined, and an algorithm for exact simulation from this model is given. Since one focus of this thesis is speed, we introduce two approximations to the Markov jump process, namely the Poisson leap and the linear noise approximation (LNA). We end this chapter by extending the models to allow for dynamic contact (or infection) rates.

In Chapter 4 we consider batch methods for Bayesian inference of S(E)IR model

parameters (and unobserved dynamic processes) given incomplete observations on the incidence process, which we assume to be subject to measurement error. We take two distinct approaches to inference; in the first, we make use of the tractability of the LNA to evaluate the observed data likelihood analytically. In the second approach we use the Poisson leap as the inferential model, leading to an intractable observed data likelihood. Progress is made by obtaining estimates of the observed data likelihood through use of a particle filter, which allows for exact, simulation-based, inference.

Chapter 5 focuses on sequential methods for Bayesian inference of the parameters (and unobserved dynamic processes). We develop a particle filter, following the approach of Storvik (2002) to alleviate particle degeneracy (or sample impoverishment) issues. A novel bridge construct is also derived, allowing us to generate end-point trajectories when propagating particles forward by conditioning on the next observation. We end this chapter with a simulated data example, to demonstrate the effectiveness of the bridge construct.

In Chapter 6 we consider a total of five applications of the methodology; two simulation studies and three real data examples. We begin by considering three applications of the particle filter. In the first application, we assess the accuracy of the particle filter through use of synthetic data, using output from a pseudo-marginal scheme as a benchmark. In the two subsequent applications, we apply the particle filter methodology to real data. Firstly, we consider its application to the Ebola data analysed in Fintzi et al. (2022), before considering the COVID-19 data taken from Spannaus et al. (2022). In each case, we compare the particle filter performance with competing methods from the relevant literature. Following this, we consider two applications of the batch methods; one simulation study for assessing

the performance of the schemes, and a real data application to the infestation of oak processionary moth (OPM) in Richmond Park, London.

Chapter 7 summarises the contributions of this thesis to the existing literature. We also briefly discuss some limitations of the work, as well as consider some potential extensions to the methodology.

Chapter 2

Monte Carlo methods

This chapter provides a detailed discussion of Monte Carlo methods, which we use to formulate various Bayesian inference techniques. These techniques will become the basis for our analysis in later chapters. We begin this chapter by reviewing some standard, well studied, Monte Carlo methods in a general setting, before discussing Markov chain Monte Carlo. We end the chapter by introducing some sequential Bayesian inference techniques, followed by some numerical examples.

2.1 Bayesian computation

To begin, suppose that we have a model for some data x (taken to be a realisation of the random variable X), which is parameterised by the continuous vector θ. Suppose further that the two are linked via the density function $f(x|\theta)$. When provided with the observed data x, we can view $f(x|\theta)$ as a function of θ and, in this case, we refer to $f(x|\theta)$ as the *likelihood function*. If we summarise our prior beliefs about the parameter vector θ as a density, denoted $\pi(\theta)$ and referred to as the *prior*

distribution, then the full joint density over data and parameters is determined by $\pi(\theta)$ and $f(x|\theta)$ as

$$f(\theta, x) = \pi(\theta) f(x|\theta).$$

Given the joint density, we are then able to compute its marginals and conditionals as

$$f(x) = \int f(\theta, x) d\theta = \int \pi(\theta) f(x|\theta) d\theta$$

and

$$f(\theta|x) = \frac{f(\theta, x)}{f(x)} = \frac{\pi(\theta) f(x|\theta)}{\int \pi(\theta) f(x|\theta) d\theta}. \tag{2.1}$$

The density function, $f(\theta|x)$, in Equation (2.1) is referred to as the *posterior density* and is typically denoted by $\pi(\theta|x)$. This leads us to the continuous form of Bayes theorem,

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int \pi(\theta) f(x|\theta) d\theta}, \tag{2.2}$$

which provides us with a method of updating our beliefs about the parameter vector θ after observing some data x through the prior density $\pi(\theta)$ and the likelihood function $f(x|\theta)$. Furthermore, since the denominator is not a function of θ, we can write

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta) \tag{2.3}$$

where the constant of proportionality is chosen to ensure that the density integrates to one.

In principle, this covers the basics about Bayesian inference, specifying the posterior distribution of the parameters given the data. However, the posterior distribution may be far from trivial to work with. Typically, we are able to write down in closed

form the numerator of Bayes theorem:

$$\text{numerator} = \pi(\theta) f(x|\theta).$$

The numerator is *not* itself a density function, i.e. it does not integrate to 1 with respect to θ. A function like this, i.e. one that is an unknown multiple of a true density function, is called a *kernel*. So while we can usually write down a kernel of the posterior distribution (i.e. we can write down the numerator in closed form), working out the denominator can be impossible analytically. The denominator has the form

$$\text{denominator} = \int \pi(\theta) f(x|\theta) d\theta$$

and so is an integral (or summation in the discrete case) over the space of parameters. We can try to evaluate this numerically, but even that may be very hard to do because

- the integral may be high dimensional, i.e. there might be many parameters, and
- the support of θ may be complicated.

The support of a random variable is the set of values on which its density is non-zero. Similar problems arise when we want to marginalize or work out an expectation:

$$E[\theta|x] = \int_\theta \theta \pi(\theta|x) d\theta.$$

Both types of integral – the denominator in Bayes theorem and the expectation – can effectively be evaluated via stochastic simulation.

2.2 Monte Carlo integration

The rationale for stochastic simulation can be summarised easily: to understand a statistical model, simulate many realisations from it and study them. For example, consider a bivariate density $f_{X,Y}(x, y)$ and suppose that the marginals $f_X(x)$ and $f_Y(y)$ are difficult to compute. If we can simulate lots of realisations of X and Y then we can look at histograms of the X and Y values, to get an idea of the marginals. We can also look at the sample mean and variance of the X values (for example) to find out the mean and variance of the marginal for X.

Suppose we want to evaluate an integral of the form

$$\mu_h = \int_{\mathcal{X}} h(x) f(x) dx$$

for some function $h(\cdot)$ and some probability density function (pdf) $f(\cdot)$ with support $\mathcal{X}$. It should be clear that the integral is an expectation, i.e. $\mu_h = E[h(X)]$. By the law of large numbers, we can estimate μ_h by taking a sample from the distribution of X and using the sample mean as an estimate of the theoretical mean. In particular, if we let $X_1, X_2, \ldots, X_N$ be independent and identically distributed (iid) draws from $f(\cdot)$, then

$$\hat{\mu}_h = \frac{1}{N} \sum_{i=1}^{N} h(X_i)$$

is an unbiased *Monte Carlo estimator* of μ_h. Furthermore, provided that the variance of $h(X)$ is finite, the estimator is also consistent. Unbiasedness can be seen by taking expectations of the estimator

$$E[\hat{\mu}_h] = \frac{1}{N} \sum_{i=1}^{N} E[h(X_i)] = E[h(X)] = \mu_h$$

and, provided $\sigma_h^2 = \text{Var}[h(X)] < \infty$ exists, consistency of the estimator can be seen by noting

$$\text{Var}(\hat{\mu}_h) = \frac{1}{N^2}\text{Var}\left[\sum_{i=1}^{N} h(X_i)\right] = \frac{\sigma_h^2}{N}$$

which vanishes as N tends to infinity. Hence, via the *Central Limit Theorem*

$$\frac{\sqrt{N}(\hat{\mu}_h - \mu_h)}{\sigma_h} \xrightarrow{D} N(0,1)$$

where the $\xrightarrow{D}$ notation denotes a convergence in distribution. Thus,

$$\hat{\mu}_h \approx N\left(\mu_h, \sigma_h^2/N\right).$$

The above properties hold provided the $\{X_i\}$ are independent. Note that the size of the error in $\hat{\mu}_h$ is proportional to the standard deviation of the estimator. This error (or equivalently, speed of convergence of the estimator) is $\mathcal{O}(N^{-1/2})$[1], rather than $\mathcal{O}(N^{-1/d})$, as would be obtained using simple numerical integration (and note that d is the dimension of X).

2.3 Importance sampling

Consider $\mu_h = E[h(X)]$ as in Section 2.2 above but suppose that $f(x)$ is difficult to sample from. Suppose we can find a *proposal density* $g(x)$ such that $g(x) > 0$ for all x with $f(x) \geq 0$, and that is easy to sample from. Then consider

$$\mu_h = \int_{\mathcal{X}} h(x)f(x)dx = \int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)}g(x)dx$$

[1] A sequence x_n is $\mathcal{O}(g_n)$ as $n \to \infty$ if for all sufficiently large n, $|x_n| \leq bg_n$ for some $b < \infty$.

which we recognise as $E_g[h(X)f(X)/g(X)]$ where the subscript makes clear that the expectation is with respect to $g(x)$. It is convenient to write this expectation as $E_g[h(X)w(X)]$ where $w(X) = f(X)/g(X)$ is known as the *weight function*. Hence, given iid draws $X_1, \ldots, X_N$ from $g(\cdot)$, an unbiased and consistent *importance sampling estimator* of μ_h can be constructed as

$$\hat{\mu}_h = \frac{1}{N} \sum_{i=1}^{N} h(X_i)w(X_i).$$

Constructing an estimator in such a way is known as *importance sampling*, and was first introduced by Kloek and van Dijk (1978). Note that in general, the *importance weights* $w(X_1), \ldots, w(X_N)$ do not add up to one. We may therefore wish to consider the *self-normalised estimator*

$$\tilde{\mu}_h = \frac{1}{\sum_{i=1}^{N} w(X_i)} \sum_{i=1}^{N} h(X_i)w(X_i).$$

In this case, the estimator becomes *biased* but *consistent*. To see consistency, note that

$$\begin{aligned}
\tilde{\mu}_h &= \frac{\frac{1}{N}\sum_{i=1}^{N} h(X_i)w(X_i)}{\frac{1}{N}\sum_{i=1}^{N} w(X_i)} \\
&\to \frac{E_g[h(X)f(X)/g(X)]}{E_g[f(X)/g(X)]} \\
&= \frac{\int_{\mathcal{X}} h(x)f(x)dx}{\int_{\mathcal{X}} f(x)dx} \\
&= E_f[h(X)]
\end{aligned}$$

where the second line follows from the strong law of large numbers, in the limit as $N \to \infty$. An advantage of using the self-normalised estimator is that it can be applied when $f(\cdot)$ is only known up to a multiplicative constant (as is often the case

for a Bayesian posterior). To see this, replace $f(x)$ with $k\pi(x)$ in the expression for $\tilde{\mu}_h$ and note that the unknown normalising constant k cancels.

2.4 Weighted resampling

Since the choice of $h(\cdot)$ used in Section 2.3 was arbitrary, we can use the sample and normalised weights to approximate any (suitable) expectation with respect to $f(x)$. If we let $\tilde{w}(x_i) = w(x_i)/\sum_{j=1}^{N} w(x_j)$ denote the ith normalised weight, then we can then view $\{x_i, \tilde{w}(x_i)\}_{i=1}^{N}$ as an *empirical approximation* of $f(x)$ via

$$\hat{f}(x) = \sum_{i=1}^{N} \tilde{w}(x_i)\delta(x - x_i)$$

where δ denotes the Dirac mass function, and is a convenient way of representing a discrete distribution as a pdf. The important point to note here is that we have approximated a continuous distribution $f(x)$ via a discrete distribution taking values $\{x_1, \ldots, x_N\}$ with associated probabilities $\{\tilde{w}(x_1), \ldots, \tilde{w}(x_N)\}$.

Often, we wish to use the weighted sample to construct summaries such as the mean, variance etc. This is most easily achieved by resampling (with replacement) from the discrete distribution on $\{x_1, \ldots, x_N\}$, generated using the proposal density $g(\cdot)$, with associated probabilities $\{\tilde{w}(x_1), \ldots, \tilde{w}(x_N)\}$. The result is an equally weighted sample, approximately distributed according to $f(\cdot)$. This technique is known as *weighted resampling*, and the general algorithm is outlined in Algorithm 1. Note that the discrete approximation to the density f, obtained via weighted resampling, becomes exact in the limit as $N \to \infty$. In other words, the weighted resampling algorithm becomes an exact method when the number of samples generated in the first step tends to infinity.

2.5 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a technique used for simulating from analytically intractable distributions, which is particularly useful for Bayesian inference. As we have seen, given a large sample from some target $\pi(\theta)$, we can estimate almost any aspect of the distribution. Often, the challenge comes in the form of obtaining the sample in the first place. To circumvent this, MCMC generates samples from a continuous Markov chain whose stationary distribution is the target density, $\pi(\theta)$. By generating samples from such a Markov chain, we can take the realisations as (dependent) samples from $\pi(\theta)$ which in turn can be used to evaluate integrals or perform inference, given that we give the chain time to reach stationarity. Note that although interest here will typically be in sampling a posterior $\pi(\theta|x)$, we will present the algorithm for a general target $\pi(\theta)$. Later, we will consider specific Bayesian problems for which interest will be in the posterior $\pi(\theta|x)$.

Algorithm 1 Weighted resampling

1. Generate N realisations $x_1, x_2, \ldots, x_n$ from the proposal density $g(\cdot)$;
2. Calculate the normalised weights $\tilde{w}(x_i)$ for each x_i via

$$\tilde{w}(x_i) = \frac{f(x_i)/g(x_i)}{\sum_{j=1}^{N} f(x_j)/g(x_j)}, \quad i = 1, 2, \ldots, N;$$

3. Resample M times with replacement from $\{x_1, x_2, \ldots, x_N\}$, using the normalised weights as probabilities.

2.5.1 Continuous state-space Markov chains

We begin by considering a univariate continuous state-space Markov chain, $\{\Theta_n\}$, over the state-space $\mathcal{X}$. Extensions of the following to the multivariate case are straightforward. The *conditional cumulative distribution function* of the chain is given by

$$P(\phi|\theta) = \mathbb{P}\left(\Theta_{n+1} \leq \phi | \Theta_n = \theta\right).$$

It is often easier for us to work with the transition density, particularly when the state-space $\mathcal{X}$ is multidimensional. This is given as

$$p(\phi|\theta) = \frac{\partial}{\partial \phi} P(\phi|\theta).$$

Now, for a density $\pi(\cdot)$ to be the stationary distribution of the Markov chain, it must satisfy

$$\pi(\theta) = \int_{\mathcal{X}} \pi(\phi) p(\theta|\phi) d\phi. \tag{2.4}$$

If we wish to check whether a density $\pi(\cdot)$ is a stationary distribution, we typically refer to the detailed balance equation, which is given as

$$\pi(\theta)p(\phi|\theta) = \pi(\phi)p(\theta|\phi) \quad \forall \theta, \phi \in \mathcal{X} \tag{2.5}$$

and note that any density $\pi(\cdot)$ satisfying Equation (2.5) is a stationary distribution of the chain. To see this, take the integral of both sides over $\mathcal{X}$ with respect to ϕ as

follows

$$\begin{aligned}\int_{\mathcal{X}} \pi(\theta)p(\phi|\theta)d\phi &= \int_{\mathcal{X}} \pi(\phi)p(\theta|\phi)d\phi \\ \pi(\theta)\int_{\mathcal{X}} p(\phi|\theta)d\phi &= \int_{\mathcal{X}} \pi(\phi)p(\theta|\phi)d\phi \\ \pi(\theta) &= \int_{\mathcal{X}} \pi(\phi)p(\theta|\phi)d\phi.\end{aligned}$$

The final line is exactly Equation (2.4), so the result holds.

A Markov chain, $\{\Theta_n\}$, with a state-space of $\mathcal{X}$ and a stationary distribution $\pi(\cdot)$ is *aperiodic* if there does not exist $d \geq 2$ disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d$ of the state-space $\mathcal{X}$ with $\mathbb{P}(\Theta_{n+1} \in \mathcal{X}_{i+1}|\Theta_n \in \mathcal{X}_i) = 1$ $(i = 1, \ldots, d-1)$ and $\mathbb{P}(\Theta_1 \in \mathcal{X}_{n+1}|\Theta_n \in \mathcal{X}_d) = 1$, and such that $\mathbb{P}_\pi(\Theta \in \mathcal{X}_1) > 0$, where $\mathbb{P}_\pi(\cdot)$ denotes the probability under the distribution $\pi(\cdot)$. Otherwise, the chain is periodic, with period d. In other words, a Markov chain is *periodic* if there are subsets of the state-space which can only be visited at regularly spaced times. Furthermore, the Markov chain, $\{\Theta_n\}$, is *μ-irreducible* if there exists a distribution, $\mu(\cdot)$, on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\mathbb{P}_\mu(\Theta \in A) > 0$, i.e. the probability of $\Theta \in A$ under the distribution $\mu(\cdot)$ greater than 0, and for all $\theta \in \Theta$ there exists a positive integer $n = n(\theta, A)$ such that $\mathbb{P}(\Theta_n \in A|\Theta_0 = \theta) > 0$. (NB: $\mu(\cdot)$ could simply be taken to be $\pi(\cdot)$). This is to say that a Markov chain is μ-irreducible if, for any initial state Θ_0, there is a distribution $\mu(\cdot)$ such that the probability of the chain visiting any subset A of the state-space (which itself holds a non-zero probability under $\mu(\cdot)$) is non-zero.

It can be shown (see Roberts and Rosenthal, 2004) that, if a Markov chain on $\mathbb{R}^d$ or an open or closed subset of $\mathbb{R}^d$ is μ-irreducible and aperiodic, and has a proper

stationary distribution, $\pi(\cdot)$, then for all $\theta \in \mathcal{X}$,

$$\mathbb{P}(\Theta_n \in A | \Theta_0 \in \theta) \to \mathbb{P}_\pi(\Theta \in A)$$

for all measurable $A \subseteq \mathcal{X}$. Here, it is understood that a 'proper' stationary distribution is one for which the integral over the entire space is 1.

2.5.2 The Metropolis-Hastings algorithm

The concept of the Metropolis-Hastings algorithm was first introduced by Metropolis et al. (1953), and was later generalised by Hastings (1970) as a statistical simulation tool that could overcome the curse of dimensionality met by regular Monte Carlo methods. The algorithm, outlined in Algorithm 2, utilises a transition kernel $q(\cdot|\theta)$, referred to as the *proposal distribution*, to explore the parameter space. This distribution typically has the same support as the target density $\pi(\cdot)$ and will be easy to simulate from, but it need not have $\pi(\cdot)$ as its stationary distribution. New parameter values, θ^*, are nominated from the proposal distribution and then

Algorithm 2 The Metropolis-Hastings algorithm

1. Initialise the chain at $\theta^{(0)}$ somewhere in the support of $\pi(\theta)$. Set the iteration counter $j = 1$;
2. Generate a proposed value θ^* using the transition kernel $q\left(\theta^*|\theta^{(j-1)}\right)$;
3. Evaluate the acceptance probability $\alpha\left(\theta^*|\theta^{(j-1)}\right)$ of the proposed move, defined by
$$\alpha\left(\theta^*|\theta\right) = \frac{\pi(\theta^*)q(\theta|\theta^*)}{\pi(\theta)q(\theta^*|\theta)};$$
4. Put $\theta^{(j)} = \theta^*$ with probability $\alpha\left(\theta^*|\theta^{(j-1)}\right)$. Otherwise put $\theta^{(j)} = \theta^{(j-1)}$;
5. Set $j := j + 1$ and return to step 2.

either accepted or rejected according to an acceptance probability $\alpha(\theta^*|\theta)$. This probability takes the form

$$\alpha(\theta^*|\theta) = \min\left\{1, \frac{\pi(\theta^*)q(\theta|\theta^*)}{\pi(\theta)q(\theta^*|\theta)}\right\}. \tag{2.6}$$

Note that the target density $\pi(\cdot)$ only appears in the acceptance probability as a ratio, meaning we need only know the target up to a constant of proportionality. Hence, if the target is a Bayesian posterior $\pi(\theta|x)$, the acceptance probability can be written as

$$\alpha(\theta^*|\theta) = \min\left\{1, \frac{\pi(\theta^*)f(x|\theta^*)q(\theta|\theta^*)}{\pi(\theta)f(x|\theta)q(\theta^*|\theta)}\right\}. \tag{2.7}$$

In what is described above, the proposal density $q(\cdot|\theta)$ is generic, meaning we have the freedom of choice. A good choice of proposal density is one which leads to a chain that converges quickly to its stationary distribution and mixes well; meaning it should explore the parameter space efficiently by moving often and well around the support of $\pi(\cdot)$.

A visualisation of a good mixing chain is presented in Figure 2.1 (top right panel) along with two examples of poorly mixing chains. In this example, we attempt to target a standard normal density. In both the top left panel and the bottom panel we see clear examples of poor mixing. In the top left, the chain takes small steps around the target space, exploring the space extremely slowly and inefficiently. In the bottom panel, we can see another example of poor mixing. This time, when the chain moves, it makes large jumps around the space, and remains stationary for large numbers of iterations, evidenced by the numerous horizontal lines seen in the trace plot. In contrast, the top right trace plot shows good mixing, where the chain is moving frequently and exploring the target space well.

To visualise convergence to the target distribution, Figure 2.2 shows three chains, again targeting the standard normal density, each with different initial states. The chain represented by the blue trace is initiated at $x_0 = 20$, the green trace at $x_0 = 0$ and the red trace at $x_0 = -20$. We can seed that, after approximately 200 iterations, the three chains are indistinguishable from each other. This shows that the three chains all converge to the correct stationary distribution in this case.

In what follows, we discuss some typical forms of $q(\cdot|\theta)$.

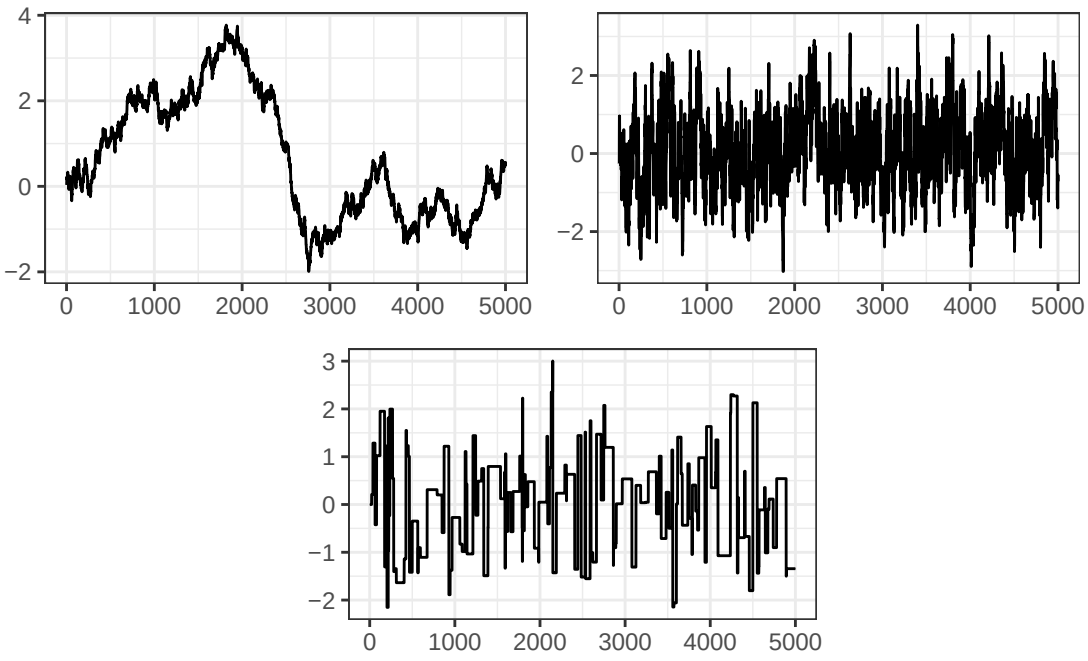

Figure 2.1: Trace plots obtained via the Metropolis-Hastings algorithm targeting the standard normal density exhibiting varying levels of mixing.

Independence proposal / sampler

In this case, the proposed transition is completely independent of the current position of the chain, and so $q(\theta^*|\theta) = q(\theta^*)$. The acceptance probability becomes

$$\alpha(\theta^*|\theta) = \min\left\{1, \frac{\pi(\theta^*)}{q(\theta^*)} \times \frac{q(\theta)}{\pi(\theta)}\right\}.$$

For the independence sampler, choosing a proposal that is close to the prior density $\pi(\cdot)$ will yield the best acceptance probability and hence the most efficient space exploration.

Symmetric proposal

A symmetric proposal is one for which $q(\theta^*|\theta) = q(\theta|\theta^*)$ for all values of θ and θ^* in the state-space. In this case, the acceptance probability simplifies significantly to give

$$\alpha(\theta^*|\theta) = \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta)}\right\}. \tag{2.8}$$

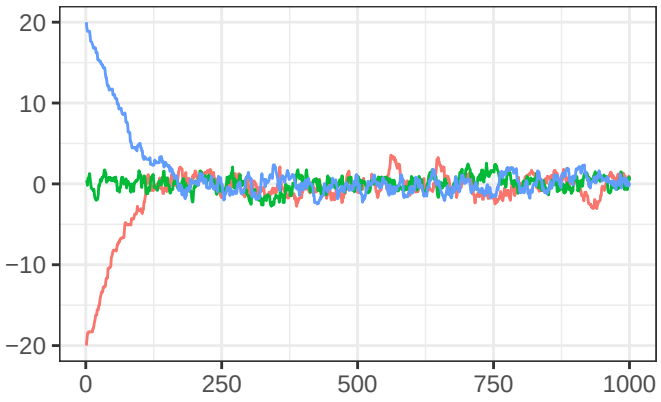

Figure 2.2: Trace plots of three runs of the Metropolis-Hastings algorithm targeting the standard normal density, inititated at $x_0 = 20$ (blue), $x_0 = 0$ (green) and $x_0 = -20$ (red).

Note that the proposal density does not feature in the acceptance probability for a symmetric proposal.

Random walk Metropolis

We can define the proposed move at iteration j to be $\theta^* = \theta^{(j-1)} + \epsilon^{(j)}$ where $\epsilon^{(j)}$ is a $d \times 1$ random vector which is completely independent of the state of the chain. Suppose that the $\epsilon^{(j)}$ have density $g(\cdot)$ which is easy to simulate from. The proposal kernel is then $q(\theta^*|\theta) = g(\theta^* - \theta)$ and this can be used to calculate the acceptance probability. Of course, if $g(\theta^* - \theta) = g(\theta - \theta^*)$, then we have a symmetric random walk chain, and the acceptance probability reduces to that given in Equation (2.8) and hence does not involve the proposal density at all.

A common choice is to let $\epsilon^{(j)} \sim N_d(0_d, V)$ where 0_d is the length-d vector of zeroes and the variance V is a tuning parameter to be chosen by the practitioner, allowing us to alter the innovation variance in order to maximise efficiency. Choosing large values in V corresponds to a chain which makes large jumps around the parameter space. This, in turn, leads to a low acceptance rate due to many proposed values being rejected. Conversely, choosing small values in V results in a chain which only makes small jumps around the space, meaning that, whilst the acceptance rate may be high, coverage of the parameter space will be quite poor. Under certain constraints on the target distribution, when d is large it can be shown that the optimal choice for the variance is

$$V = \frac{2.38^2}{d}\text{Var}(\theta) \tag{2.9}$$

leading to an optimal acceptance rate of 0.234 (see Roberts and Rosenthal, 2001,

for more details). In practice, we don't know the posterior variance $\text{Var}(\theta)$ and so we typically substitute this for the variance obtained from a short pilot run using an arbitrary choice of innovation variance. It should be noted that the variance given in Equation (2.9) is simply a guide and in practice, particularly in the case when d is small, the optimal acceptance rate could fall anywhere in the region of 0.1 to 0.4.

A visualisation of the effect of the innovation variance on the performance of the Metropolis-Hastings algorithm is provided in Figure 2.3. In this example, the target distribution is a mixture of bivariate normal densities, and is represented by the blue contours in each panel. Shown as black dots are the accepted values of the chain, which are nominated using a random walk with Gaussian innovations. The variance of the innovations is taken to be of the form $c\widehat{\text{Var}}(\theta)$, where c is a constant, taking values 1/100 (top left), 1 (top right) and 100 (bottom), and $\widehat{\text{Var}}(\theta)$ is an estimate of the posterior variance obtained from a short pilot run. The acceptance probabilities for $c = 1/100$, 1 and 100 were 92.4%, 43.5% and 1.5% respectively based on a 5000 iteration run of the scheme. The behaviour we describe above is clearly visible here; when $c = 1/100$, the chain accepts almost all the proposed values, but coverage of the target is poor. Conversely, when $c = 100$, we see that very few proposed values are accepted, and so the chain remains in one place for large periods of time. Finally, when we choose $c = 1$, the chain moves efficiently around the space, providing a clear balance between coverage and acceptance. Given in Figure 2.4 are trace plots for this experiment setup, which clearly highlight the effect that the innovation variance has on the mixing of the chain.

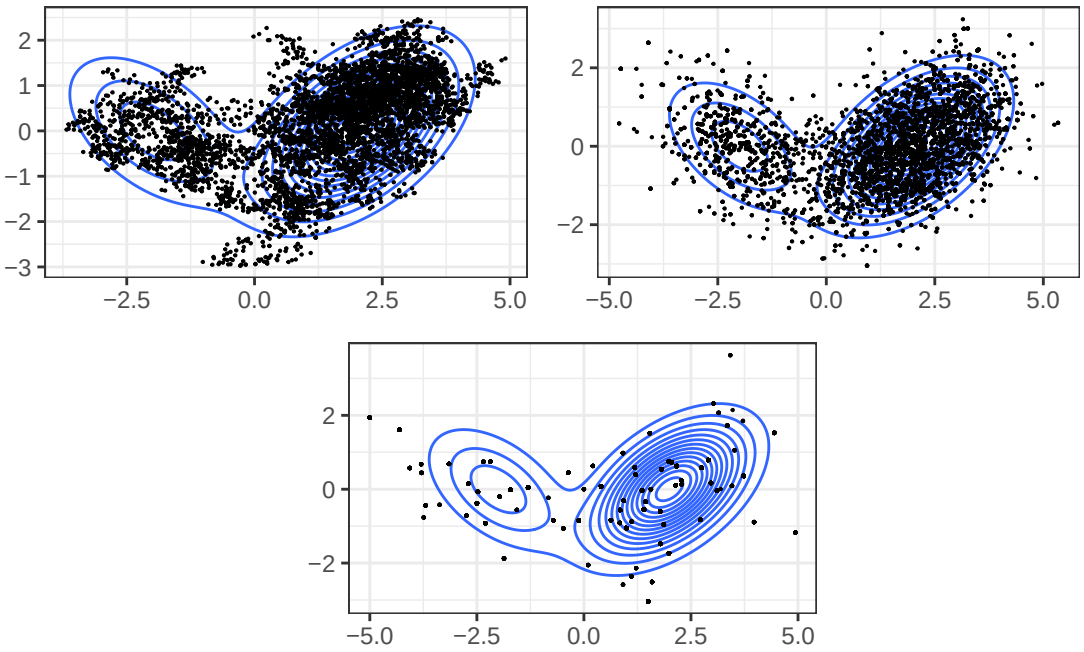

Figure 2.3: Target density (blue contours) overlaid with sample obtained using random walk Metropolis with Gaussian innovations (black dots) with variances $c\widehat{\mathrm{Var}(\theta)}$, for $c = 0.01$ (top left), $c = 1$ (top right) and $c = 100$ (bottom).

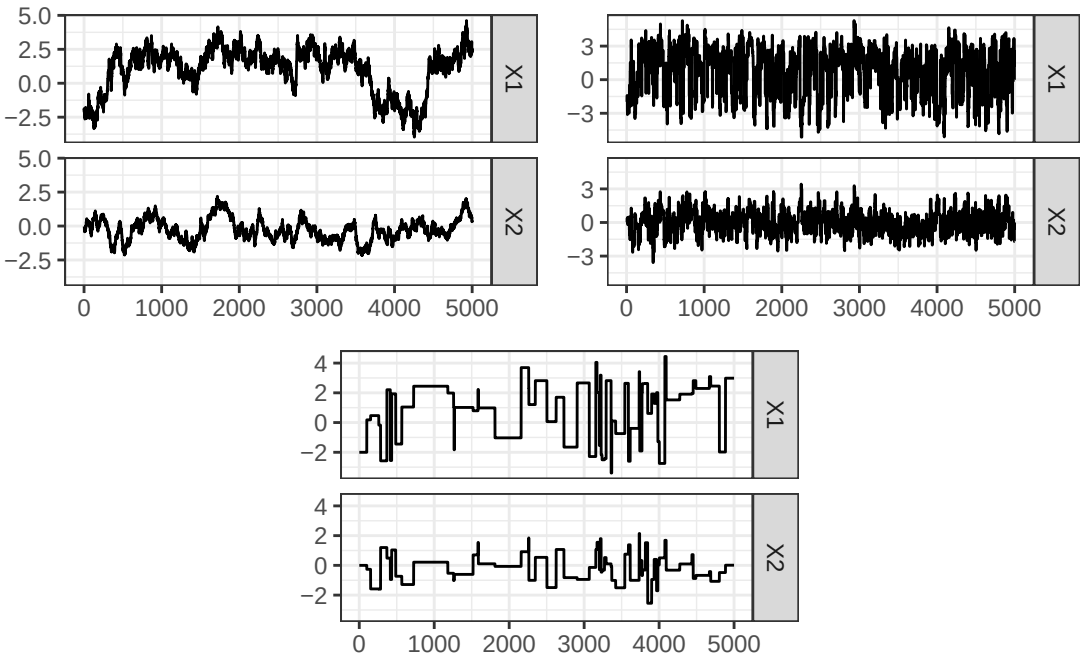

Figure 2.4: Trace plot pairs obtained using random walk Metropolis with Gaussian innovations with variances $c\widehat{\mathrm{Var}(\theta)}$, for $c = 0.01$ (top left), $c = 1$ (top right) and $c = 100$ (bottom).

2.5.3 Validity of Metropolis-Hastings

For the Metropolis-Hastings algorithm to be valid, it must target the correct distribution. In particular, the chain generated via Metropolis-Hastings must have $\pi(\theta)$ as its stationary distribution, and it must converge to this distribution. We can check whether $\pi(\theta)$ is a stationary distribution by checking whether the detailed balance equation in Equation (2.5) is satisfied. To show convergence, we need to show that the chain is μ-irreducible and aperiodic. Below we show that detailed balance is satisfied, but refer the reader to Brooks et al. (2011) for an in-depth discussion regarding convergence.

Consider the kernel for the Metropolis-Hastings algorithm. Assuming the chain moves, this is given by

$$p(\phi|\theta) = \alpha(\phi|\theta)q(\phi|\theta) \quad \text{when } \theta \neq \phi.$$

There is also a non-zero probability that the chain does not move, given by

$$1 - \int_{\mathcal{X}} \alpha(\phi|\theta)q(\phi|\theta)d\theta.$$

We can then check whether detailed balance holds for the non-trivial case corresponding to when the chain moves:

$$\begin{aligned}\pi(\theta)p(\phi|\theta) &= \pi(\theta)q(\phi|\theta)\min\left\{1, \frac{\pi(\phi)\,q(\theta|\phi)}{\pi(\theta)\,q(\phi|\theta)}\right\} \\ &= \min\left\{\pi(\theta)\,q(\phi|\theta), \pi(\phi)\,q(\theta|\phi)\right\}.\end{aligned}$$

This expression is clearly symmetric in θ and ϕ. Hence, detailed balance holds, and we may conclude that the Metropolis-Hastings algorithm defines a Markov chain

with stationary distribution $\pi(\theta)$.

2.6 Pseudo-marginal Metropolis-Hastings

In some cases, the target density $\pi(\theta)$ might not have an analytic form, even up to proportionality. In this case, progress may be made using pseudo-marginal Metropolis-Hastings (PMMH, see Andrieu and Roberts, 2009), where instead we make use of a non-zero unbiased estimator of $\pi(\theta)$, which we denote by $\hat{\pi}_U(\theta)$. Note here the dependence of the estimator on the U, which is the collection of random variables needed to generate the estimator of the target such that a single realisation of $U \sim g(u)$, for some density $g(\cdot)$, gives an estimate $\hat{\pi}_U(\theta)$.

As suggested by the name, the PMMH scheme is a Metropolis-Hastings scheme. Unlike the Metropolis-Hastings algorithm presented in Section 2.5.2, the PMMH scheme targets a joint density, which is then marginalised to give exact draws from the desired distribution. In particular, PMMH targets

$$\pi(\theta, u) \propto \hat{\pi}_u(\theta) g(u) \tag{2.10}$$

by first simulating values from a proposal density of the form $q(\theta^*|\theta)g(u^*)$ and then either accepting or rejecting the proposed move according to the acceptance probability, which is given as

$$\begin{aligned}
\alpha(\theta^*, u^*|\theta, u) &= \min\left\{1, \frac{\pi(\theta^*, u^*)}{\pi(\theta, u)} \times \frac{q(\theta|\theta^*)g(u)}{q(\theta^*|\theta)g(u^*)}\right\} \\
&= \min\left\{1, \frac{\hat{\pi}_{u^*}(\theta^*)g(u^*)}{\hat{\pi}_u(\theta)g(u)} \times \frac{q(\theta|\theta^*)g(u)}{q(\theta^*|\theta)g(u^*)}\right\} \\
&= \min\left\{1, \frac{\hat{\pi}_{u^*}(\theta^*)}{\hat{\pi}_u(\theta)} \times \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right\},
\end{aligned}$$

where the second line follows from the first via Equation (2.10). Note that the density $g(\cdot)$ does not appear in the acceptance probability, meaning we only need to know how to sample from $g(\cdot)$, without ever needing to evaluate its density function.

Now, to show that the desired target distribution is a marginal distribution of the joint density targeted by PMMH, we consider the integral

$$\begin{aligned}\int \pi(\theta, u)du &= \int \hat{\pi}_u(\theta)g(u)du \\ &= E_U[\hat{\pi}_U(\theta)] \\ &= \pi(\theta)\end{aligned}$$

which is true under the assumption that the estimator $\hat{\pi}_U(\theta)$ is a non-zero unbiased estimator of the target $\pi(\theta)$.

As with random walk Metropolis, there is a level of tuning required in order to maximise efficiency of PMMH. Here, the variance of the estimator $\hat{\pi}_U(\theta)$ acts as the tuning parameter, which we are able to increase or decrease in order to balance mixing of the chain and computational efficiency. In some scenarios, finding the optimal variance of the estimator can make the scheme computationally infeasible, for example when generation of one such estimate takes a considerable amount of time. When this situation arises, we can employ correlated PMMH (CPMMH) which reduces the variance of the estimator by inducing correlation between successive estimates. CPMMH is discussed in more depth in Section 2.7. We return to the topic of tuning the PMMH scheme in Chapter 4, after discussing methods for generating estimates $\hat{\pi}_u(\theta)$ in the context of a stochastic epidemic model.

2.6.1 Pseudo-marginal toy example

To illustrate the pseudo-marginal methodology, we consider the following 'toy' example. Suppose we are interested in generating draws from a $N(0,1)$ distribution and, for the purposes of this example, we aren't able to use standard functions (such as `rnorm` in R) to help us do this. Our desired target distribution is therefore given by

$$\pi(\theta) \propto \exp\{-\theta^2/2\}$$

up to proportionality. Suppose instead that we are only able to obtain estimates $\hat{\pi}_u(\theta)$, realisations of the estimator $\hat{\pi}_U(\theta)$, from the density

$$\hat{\pi}_U = \pi(\theta)U$$

where U has a density function $g(u)$ and an expectation of 1. Clearly, to generate an estimate from this density we first generate a sample u from $g(u)$, and then multiply this by the density function of our target $\pi(\theta)$. It is easy to see that this setup leads to an unbiased estimator for $\pi(\theta)$ by considering

$$\begin{aligned} E_U[\hat{\pi}_U(\theta)] &= E_U[\pi(\theta)U] \\ &= \pi(\theta)E[U] \\ &= \pi(\theta). \end{aligned}$$

Therefore, any $g(u)$ with an expectation of 1 will work. To highlight the impact of the variance of the estimator on the mixing of the PMMH scheme, we choose to use a LogNormal$(-v^2/2, v^2)$ distribution for U such that

$$g(u) \propto \theta^{-1} \exp\left\{-\frac{(\log(\theta) + v^2/2)^2}{2v^2}\right\}, \quad E[U] = 1, \quad \text{Var}(U) = \exp\{v^2\} - 1$$

and so by increasing the value of v, we increase the estimator's variance.

Presented in Figure 2.5 are histograms of the samples obtained by running the PMMH scheme as described above, using random walk innovations i.e. $\theta^* = \theta + \epsilon$ and $\epsilon \sim N(0, 1)$, with $v^2 = \log(2)$ (top left), $v^2 = \log(101)$ (top centre) and $v^2 = \log(1001)$ (top right). Overlaid on each histogram is the target density, a $N(0, 1)$. We can see that for the scheme using $v^2 = \log(2)$, and hence a variance of 1, the samples match the target density very well. This indicates that the scheme is well mixing and is targeting the correct distribution. As we increase v^2 to $\log(101)$ and then $\log(1001)$, consequently increasing the variance from 1 to 100 and 1000, we can see that the histograms begin to deviate from the target density more and more. As discussed in Sherlock et al. (2015), this is due to the high variance providing overestimates of the target, which in turn results in a chain that moves infrequently. This 'sticky' behaviour means the chain remains in one place for multiple iterations of the scheme and is best seen by examining the trace plots, which are given in the bottom row of plots in Figure 2.5. On the left, when the variance is at its smallest, the trace plot indicates a well mixed chain and doesn't display any sign of the sticking behaviour. However, as we progress through the plots to the right, we see a deterioration in the traces, with sticking points clearly visible; appearing as horizontal lines in many places.

2.7 Correlated PMMH

As previously discussed, correlated pseudo-marginal Metropolis-Hastings (CPMMH) (see Deligiannidis et al., 2018; Dahlin et al., 2015; Golightly et al., 2019) is employed when the variance of the estimator is large, or when obtaining an estimator with

optimal variance is computationally infeasible. In order to reduce the variance of the estimator, CPMMH introduces correlation between successive estimates through a specific choice of the proposal density. Recall that, for the basic PMMH scheme, the proposal density is a joint density of the form $q(\theta^*|\theta)g(u^*)$. For CPMMH, we replace the density $g(u^*)$ with a kernel $K(u^*|u)$, which is chosen such that it satisfies the detailed balance equation

$$g(u)K(u^*|u) = g(u^*)K(u|u^*). \tag{2.11}$$

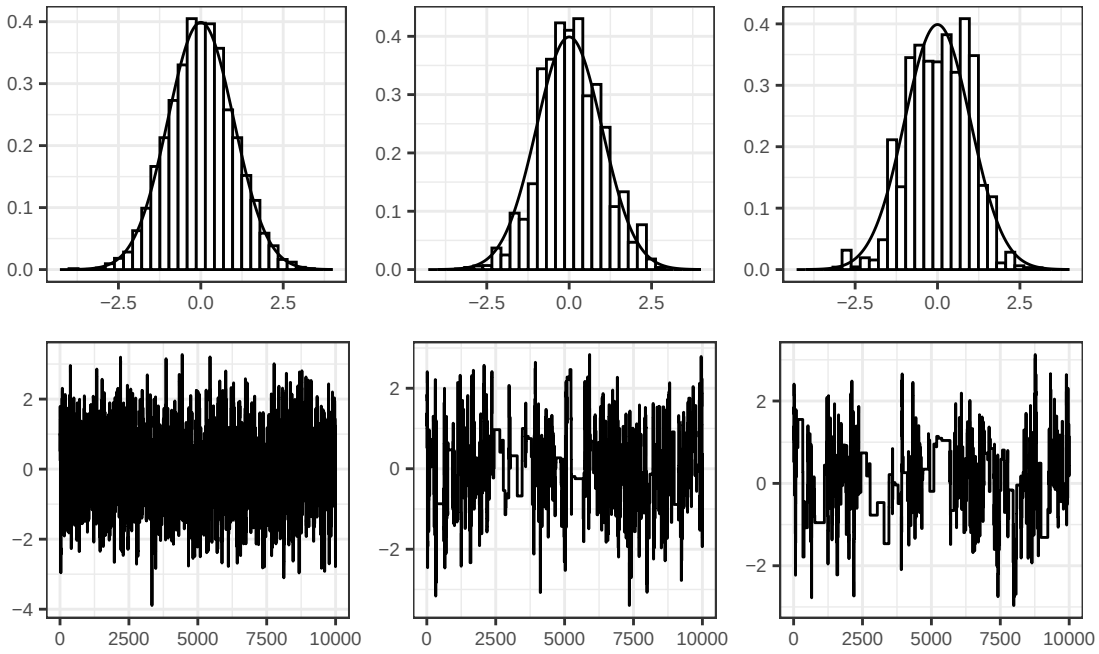

Figure 2.5: Histograms (top row) and trace plots (bottom row) of samples of a $N(0,1)$ distribution obtained via a random-walk PMMH scheme with estimator $\hat{\pi}_U(\theta) = \pi(\theta)U$ and $U \sim \text{LogNormal}(-v^2/2, v^2)$ for $v = \log(2)^{1/2}$ (left), $v = \log(101)^{1/2}$ (centre) and $v = \log(1001)^{1/2}$ (right). Overlaid on each histogram is the target density.

In particular, the choices of density $g(\cdot)$ and kernel $K(\cdot|\cdot)$ are a standard multivariate normal density and a Crank-Nicolson kernel, respectively. That is, we choose

$$g(u) = N(u; 0, I_d) \qquad \text{and} \qquad K(u^*|u) = N\left(u^*; \zeta u, (1-\zeta^2)I_d\right),$$

where I_d is the $d \times d$ identity matrix and ζ is a parameter taking values in $(-1, 1)$, which is to be chosen by the practitioner and controls the correlation between u and u^*. Typically, we choose ζ to be close to one, to induce a high level of positive correlation and therefore reduce the variance of the estimator.

It is straightforward to show that this setup satisfies the detailed balance equation by substituting the above Gaussian densities into one side of Equation (2.11). For simplicity, consider the left-hand side of Equation (2.11) in the univariate case, i.e. when $d = 1$:

$$\begin{aligned}
g(u)K(u^*|u) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} \times \frac{1}{\sqrt{2\pi(1-\zeta^2)}} \exp\left\{-\frac{(u^* - \zeta u)^2}{2(1-\zeta^2)}\right\} \\
&= \frac{1}{2\pi\sqrt{(1-\zeta^2)}} \exp\left\{-\frac{u^2(1-\zeta^2) + (u^* - \zeta u)^2}{2(1-\zeta^2)}\right\} \\
&= \frac{1}{2\pi\sqrt{(1-\zeta^2)}} \exp\left\{\frac{u^2 - \zeta^2 u^2 + u^{*2} + \zeta^2 u^2 - 2\zeta u^* u}{2(1-\zeta^2)}\right\} \\
&= \frac{1}{2\pi\sqrt{(1-\zeta^2)}} \exp\left\{\frac{u^2 + u^{*2} - 2\zeta u^* u}{2(1-\zeta^2)}\right\} \\
&= \frac{1}{2\pi\sqrt{(1-\zeta^2)}} \exp\left\{\frac{u^{*2} + (u - \zeta u^*)^2 - \zeta^2 u^{*2}}{2(1-\zeta^2)}\right\} \\
&= \frac{1}{2\pi\sqrt{(1-\zeta^2)}} \exp\left\{\frac{u^{*2}(1-\zeta^2) + (u - \zeta u^*)^2}{2(1-\zeta^2)}\right\} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^{*2}}{2}\right\} \times \frac{1}{\sqrt{2\pi(1-\zeta^2)}} \exp\left\{-\frac{(u - \zeta u^*)^2}{2(1-\zeta^2)}\right\} \\
&= g(u^*)K(u|u^*)
\end{aligned}$$

and so the left-hand side is equivalent to the right-hand side of Equation (2.11), and therefore detailed balance is satisfied. In the above, the fifth line follows from the fourth line by recognising that

$$u^2 - 2\zeta u^* u = (u - \zeta u^*)^2 - \zeta^2 u^{*2}.$$

It is shown in Deligiannidis et al. (2018) that, in the context of likelihood estimates in state-space models, under certain assumptions regarding the differentiability and integrability of the normalised importance weight, the variance of the log-likelihood ratio of estimators for the CPMMH scheme is smaller than that of the PMMH scheme. Moreover, it is shown that, under these same conditions, the CPMMH scheme is less prone to sticking behaviour than PMMH at stationarity.

2.7.1 Normal example continued

To demonstrate the effectiveness of CPMMH, we return to the toy example of Section 2.6.1, where we attempt to (marginally) target a standard normal density. Recall that, in this toy example, we obtain estimates $\hat{\pi}_u(\theta)$, which are realisations of the unbiased estimator $\hat{\pi}_U(\theta)$, by first generating a value u from a density $g(u)$, where $g(u)$ is taken to be the density function of a LogNormal$(-v^2/2, v^2)$ random variable, and then multiplying this by the density function of the target, $\pi(\theta)$. For CPMMH, we require a standard normal density to use the Crank-Nicolson kernel. We can make progress by applying the inverse cdf method, where we first obtain a realisation $\tilde{u}$ from the standard normal random variable $\tilde{U}$ with pdf and cdf given by $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. This can then be transformed into a realisation of U via

$$U = G^{-1}\left(\Phi(\tilde{U})\right)$$

where $G^{-1}(\cdot)$ is the inverse cdf of the random variable U. It should be clear that this setup jointly targets the density

$$\begin{aligned}\pi(\theta, \tilde{u}) &\propto \hat{\pi}_{\tilde{u}}(\theta)\phi(\tilde{u}) \\ &\propto \pi(\theta)G^{-1}(\Phi(\tilde{u}))\phi(\tilde{u}).\end{aligned}$$

To show that the estimator $\hat{\pi}_{\tilde{U}}(\theta)$ is unbiased, we can consider the expectation

$$\begin{aligned}E_{\tilde{U}}[\hat{\pi}_{\tilde{U}}(\theta)] &= E_{\tilde{U}}[\pi(\theta)G^{-1}(\Phi(\tilde{U}))] \\ &= \pi(\theta)E_{\tilde{U}}[G^{-1}(\Phi(\tilde{U}))] \\ &= \pi(\theta)E_U[U] \\ &= \pi(\theta).\end{aligned}$$

Since the new estimator is also unbiased as an estimator of the target $\pi(\theta)$, we can conclude that the CPMMH scheme constructed in this way generates samples from the correct marginal density of interest, $\pi(\theta)$.

Presented in Figure 2.6 are histograms and the corresponding trace plots of three runs of the CPMMH scheme, each for 10k iterations using $\zeta = 0.999$. To allow for comparison to Figure 2.5 from Section 2.6.1, the parameter values remain the same in this example. The improvement is clear; in each case the histogram is consistent with the target density and there is no evidence of the sticky behaviour in the trace plots. Increasing the variance of U doesn't have as profound an effect on the samples as in the case of the vanilla PMMH scheme from Section 2.6.

2.8 Sequential Monte Carlo

Often, observations arrive sequentially in time and one is interested in performing inference *on-line*. From a Bayesian perspective, updating the posterior distribution as data becomes available is necessary. Examples include tracking aircraft using radar measurements, estimating a communications signal using noisy measurements or estimating the volatility of financial instruments using stock market data.

The MCMC applications that we have looked at so far can be thought of as *batch analyses*, that is, we run the sampler for all available data we are in receipt of. A naive approach to implementing MCMC in a sequential fashion involves re-running the sampler (from scratch) as each observation arrives. This is wasteful in terms of

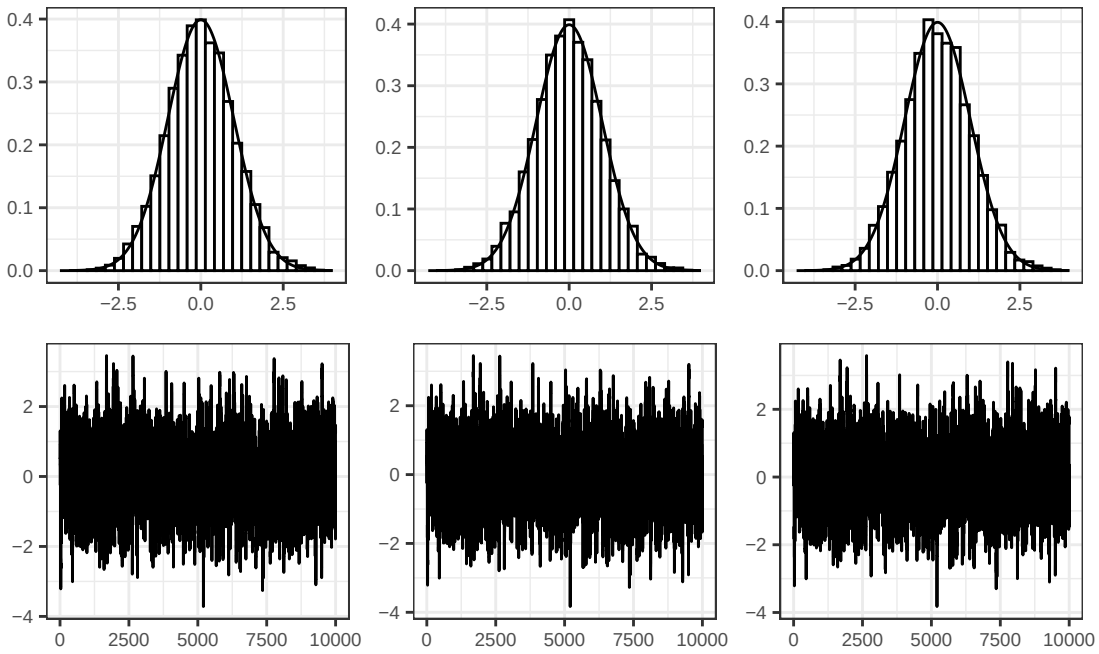

Figure 2.6: Histograms (top row) and trace plots (bottom row) of samples of a $N(0,1)$ distribution obtained via a random-walk CPMMH scheme with estimator $\hat{\pi}_{\tilde{u}}(\theta) = \pi(\theta)G^{-1}(\Phi(\tilde{u}))$ where $\tilde{U} \sim N(0,1)$ and $U \sim \text{LogNormal}(-v^2/2, v^2)$ for $v = \log(2)^{1/2}$ (left), $v = \log(101)^{1/2}$ (centre) and $v = \log(1001)^{1/2}$ (right). Overlaid on each histogram is the target density.

storage and computational cost, so a different approach is required.

2.8.1 Hidden Markov models

A state-space model (SSM) is a time series model that consists of two discrete-time processes $\{X_t, t \geq 0\}$ and $\{Y_t, t \geq 0\}$ taking values respectively in spaces $\mathcal{X}$ and $\mathcal{Y}$. The first process, $\{X_t, t \geq 0\}$, referred to as the *latent process*, is not observed; instead, only a noisy function of the state is observed, and only at discrete times. These observations form the *observed process*, given by $\{Y_t, t \geq 0\}$, which is linked to the latent process via some density $\pi(y_t|x_t, y_{0:t-1}, \theta)$ parameterised by θ. The model is completed by the specification of $\pi(x_0|\theta)$ and $\pi(x_t|x_{0:t-1}, \theta)$ and is typically presented as a series of equations such as

$$\begin{aligned} X_t|X_{0:t-1}, \theta &\sim \pi(x_t|x_{0:t-1}, \theta), \\ Y_t|X_t, Y_{0:t-1}, \theta &\sim \pi(y_t|x_t, y_{0:t-1}, \theta), \\ X_0|\theta &\sim \pi(x_0|\theta). \end{aligned}$$

When the underlying state process is Markovian, i.e. when the transition densities depend only on the state of the process at the current time, we refer to these models as *hidden Markov models* (HMMs). The densities specifying a hidden Markov model therefore reduce to $\pi(x_t|x_{t-1}, \theta)$, $\pi(y_t|x_t, \theta)$ and $\pi(x_0|\theta)$ leading to the set of equations given by

$$\begin{aligned} X_t|X_{t-1}, \theta &\sim \pi(x_t|x_{t-1}, \theta), \\ Y_t|X_t, \theta &\sim \pi(y_t|x_t, \theta), \\ X_0|\theta &\sim \pi(x_0|\theta) \end{aligned}$$

with associated joint density given by

$$\pi(x_{0:T}, y_{0,T}|\theta) = \pi(x_0|\theta) \prod_{t=0}^{T} \pi(y_t|x_t, \theta) \prod_{t=1}^{T} \pi(x_t|x_{t-1}).$$

This describes a generative probabilistic model, where X_0 is drawn from the initial density $\pi(x_0|\theta)$, and then each X_t is drawn conditionally on the previous draw $X_{t-1} = x_{t-1}$ according to the density $\pi(x_t|x_{t-1}, \theta)$, and each Y_t conditionally on the most recent $X_t = x_t$ from $\pi(y_t|x_t, \theta)$.

The model can also be represented graphically by having variables as nodes and an edge between two variables that are related by one of the kernels in the above definition. An example of such a graphical representation is given in Figure 2.7.

When analysing HMMs, there are typically 3 key aims:

- **Filtering:** after a new observation arrives, we wish to learn the corresponding hidden state via the filtering density $\pi(x_t|y_{0:t})$;
- **Smoothing:** given all of the data, learn the hidden states via the smoothing density $\pi(x_{0:t}|y_{0:T})$;
- **Parameter inference:** learn the parameters via the (marginal) posterior

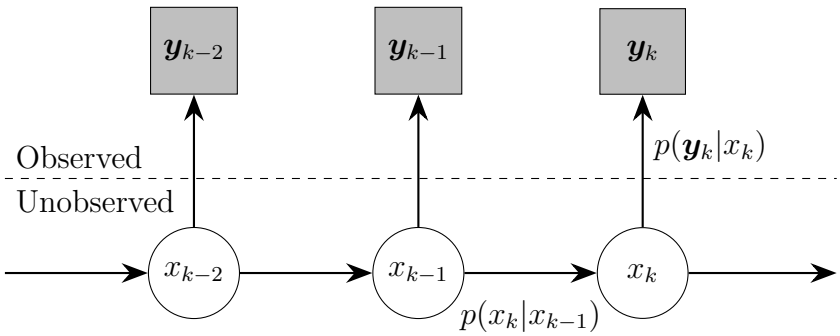

Figure 2.7: Partially observed Markov process as a directed graphical model.

density $\pi(\theta|y_{0:T})$.

The first two aims are tractable given the parameters θ, that is to say that the filtering and smoothing distributions are available in closed form when:

1. the unobserved Markov chain has a finite number of states or
2. the unobserved Markov chain is a Gaussian autoregressive process.

Except in these two cases, sequential Monte Carlo methods provide the best generic methods for filtering and smoothing. These methods are known as *particle filtering* and *particle smoothing* (see Chopin and Papaspiliopoulos, 2020, for an introduction). The methods presented in this thesis will focus on filtering problems, therefore the remainder of this chapter will focus on particle filtering, which is based on the idea of approximating distributions by weighted samples. For the methods presented in Sections 2.8.2 and 2.8.3, we assume that the parameters θ are fixed and known, and therefore we drop them from the notation where possible.

2.8.2 Sequential importance sampling

Sequential importance sampling recursively targets the joint filtering distribution $\pi(x_{0:t}|y_{1:t})$ at time t, which can be factorised as

$$\begin{aligned}
\pi(x_{0:t}|y_{1:t}) &= \pi(x_{0:t}|y_t, y_{1:t-1}) \\
&= \frac{\pi(x_{0:t}, y_t|y_{1:t-1})}{\pi(y_t|y_{1:t-1})} \\
&= \frac{\pi(y_t|y_{1:t-1}, x_{0:t})\pi(x_{0:t}|y_{1:t-1})}{\pi(y_t|y_{1:t-1})} \\
&= \frac{\pi(y_t|y_{1:t-1}, x_{0:t})\pi(x_t, x_{0:t-1}|y_{1:t-1})}{\pi(y_t|y_{1:t-1})} \\
&= \frac{\pi(y_t|x_t)\pi(x_t|x_{0:t-1}, y_{1:t-1})\pi(x_{0:t-1}|y_{1:t-1})}{\pi(y_t|y_{1:t-1})} \\
&\propto \pi(y_t|x_t)\pi(x_t|x_{t-1})\pi(x_{0:t-1}|y_{1:t-1})
\end{aligned}$$

using the definition of conditional probability and the Markov property. This takes the form of a typical Bayes posterior, in the sense that the posterior is proportional to a prior multiplied by a likelihood term.

Now, assume that at time t we have a weighted sample from the target at time $t-1$ given by $\left\{x_{0:t-1}^{(k)}, w_{t-1}^{(k)}\right\}_{k=1}^{N}$, which we call *particles*. Using this sample of particles and letting $\delta(x_{0:t-1}^{(k)})$ be the Dirac mass function, an approximation of the target at time t can then be obtained via

$$\hat{\pi}(x_{0:t}|y_{1:t}) \propto \pi(y_t|x_t)\pi(x_t|x_{t-1})\hat{\pi}(x_{0:t-1}|y_{1:t-1})$$

where the empirical approximation of $\pi(x_{0:t-1}|y_{1:t-1})$, denoted by $\hat{\pi}(x_{0:t-1}|y_{1:t-1})$, is of the form

$$\hat{\pi}(x_{0:t-1}|y_{1:t-1}) = \sum_{k=1}^{N} w_{t-1}^{(k)}\delta(x_{0:t-1}^{(k)}).$$

The key to sequential importance sampling is the choice of importance density $g(x_{0:t}|y_{1:t})$. Suppose we take an importance density that can be factorised as

$$g(x_{0:t}|y_{1:t}) = g(x_t|x_{t-1}, y_t)g(x_{0:t-1}|y_{1:t-1})$$

then it should be clear that we can obtain a sample from the importance density at time t via the sample from the importance density at time $t-1$ and so on. We obtain an expression for the weight by considering the ratio of the target density to the importance density at each time. For a general time t, this is

$$\begin{aligned} w_t &= \frac{\pi(x_{0:t}|y_{1:t})}{g(x_{0:t}|y_{1:t})} \\ &\propto \frac{\pi(y_t|x_t)\pi(x_t|x_{t-1})\pi(x_{0:t-1}|y_{1:t-1})}{g(x_t|x_{t-1}, y_t)g(x_{0:t-1}|y_{1:t-1})} \\ &\propto \frac{\pi(y_t|x_t)\pi(x_t|x_{t-1})}{g(x_t|x_{t-1}, y_t)} w_{t-1} \\ &\propto \frac{\pi(y_t, x_t|x_{t-1})}{g(x_t|x_{t-1}, y_t)} w_{t-1}. \end{aligned} \tag{2.12}$$

Repeating these propagate and re-weight steps through time forms the basis of sequential importance sampling. A full algorithm is provided in Algorithm 3.

A typical sequential importance sampling algorithm pitfall is *sample degeneracy*. Sample degeneracy, also known as sample impoverishment, occurs when only a small number of the particles in the sample yield a significant importance weight. This typically occurs when the target density is of high dimension, or when the time horizon over which we apply Algorithm 3 is large. Consequently, this small number of particles dominate the empirical approximation of $\pi(x_{0:t-1}|y_{1:t-1})$, resulting in a poor approximation of the target density. In the worst case, all but one particle are given a weight of zero and the target density is represented only by a single point mass.

One method of circumventing this particle degeneracy problem is to add a re-sample step at the end of Algorithm 3. The effect of this is to essentially prune out particles with negligible weights, resulting in an equally weighted sample from the target distribution. This method is referred to as the sequential importance re-sampling method, and the full algorithm is given in Algorithm 4. As presented, an equally weighted sample, approximately distributed according to the marginal density $\pi(x_t|y_{1:t})$, is available after step 3.

2.8.3 Bootstrap particle filter

The bootstrap particle filter is a sequential importance re-sampling algorithm first introduced in Gordon et al. (1993). Specifically, sequential importance re-sampling becomes the bootstrap particle filter when we take the importance density $g(x_t|x_{t-1}, y_t)$ to be exactly the density specifying the latent process of the HMM, $\pi(x_t|x_{t-1})$.

In this case, the importance weight at time t reduces to $w_t = \pi(y_t|x_t)w_{t-1}$. To see

Algorithm 3 Sequential importance sampling

Initialise with a sample $\left\{x_0^{(k)}, w_0^{(k)}\right\}_{k=1}^N$ by sampling $x_0^{(k)} \sim \pi(x_0)$ and setting $w_0^{(k)} = 1$ for $k = 1, 2, \ldots, N$.

For $t = 1, 2, \ldots, T$:

1. Propagate each $x_{t-1}^{(k)}$ forward via $g(x_t^{(k)}|x_{t-1}^{(k)})$ for $k = 1, 2, \ldots, N$;
2. Evaluate the weights

$$w_t^{(k)} = \frac{\pi(y_t, x_t^{(k)}|x_{t-1}^{(k)})}{g(x_t^{(k)}|x_{t-1}^{(k)}, y_t)} w_{t-1}^{(k)}, \quad \text{for } k = 1, 2, \ldots, N$$

and normalise using $\tilde{w}_t^{(k)} = w_t^{(k)} / \sum_{k=1}^N w_t^{(k)}$.

this, consider the expansion of the weight on the third line of Equation (2.12), which is given as

$$w_t \propto \frac{\pi(y_t|x_t)\pi(x_t|x_{t-1})}{g(x_t|x_{t-1}, y_t)} w_{t-1}.$$

Clearly, upon making the substitution $g(x_t|x_{t-1}, y_t) = \pi(x_t|x_{t-1})$, the fractional part of the weight reduces to just $\pi(y_t|x_t)$. The full algorithm is given below in Algorithm 5.

2.8.4 Liu and West algorithm

The sequential algorithms considered up to this point have dealt solely with the problem of state filtering. Thus, so far, we have made the assumption that the parameters θ are fixed and known. In practice, this is often not the case, and so we typically want to include the parameter vector θ in our inferential model.

Algorithm 4 Sequential importance re-sampling

Initialise with a sample $\left\{x_0^{(k)}, \tilde{w}_0^{(k)}\right\}_{k=1}^{N}$ by sampling $x_0^{(k)} \sim \pi(x_0)$ and setting $w_0^{(k)} = 1$ for $k = 1, 2, \ldots, N$.

For $t = 1, 2, \ldots, T$:

1. Propagate each $x_{t-1}^{(k)}$ forward via $g(x_t^{(k)}|x_{t-1}^{(k)})$ for $k = 1, 2, \ldots, N$;
2. Evaluate the weights

$$w_t^{(k)} = \frac{\pi(y_t, x_t^{(k)}|x_{t-1}^{(k)})}{g(x_t^{(k)}|x_{t-1}^{(k)}, y_t)} w_{t-1}^{(k)}, \quad \text{for } k = 1, 2, \ldots, N$$

   and normalise using $\tilde{w}_t^{(k)} = w_t^{(k)} / \sum_{k=1}^{N} w_t^{(k)}$;
3. Resample N times with replacement from $\left\{x_t^{(k)}, w_t^{(k)}\right\}$ using the normalised weights $\tilde{w}_t^{(k)}$ as probabilities. Set each $w_t^{(k)} = 1$.

Consequently, the target distribution at time t becomes $\pi(x_{0:t}, \theta|y_{1:t})$. Expanding this target using Bayes theorem gives

$$\pi(x_{0:t}, \theta|y_{1:t}) \propto \pi(x_{0:t-1}, \theta|y_{1:t-1})\pi(x_t|x_{t-1}, \theta)\pi(y_t|x_t, \theta)$$

which, upon taking the particle approximation

$$\hat{\pi}(x_{0:t-1}, \theta|y_{1:t-1}) = \sum_{k=1}^{N} w_{t-1}^{(k)} \delta(x_{0:t-1}^{(k)}, \theta^{(k)}),$$

allows for sequential Bayesian inference by propagating states forward via $\pi(x_t|x_{t-1}, \theta)$ and weighting according to $\pi(y_t|x_t, \theta)$.

The Liu and West algorithm (see Liu and West, 2001) is the first state and parameter filtering algorithm we consider. The idea is to treat each $\theta^{(k)}$ as a dynamic process which is propagated through time according to a Gaussian kernel density estimate, similar to the jittering approach for state inference found in Gordon et al. (1993).

Algorithm 5 Bootstrap particle filter

Initialise with a sample $\left\{x_0^{(k)}, \tilde{w}_0^{(k)}\right\}_{k=1}^{N}$ by sampling $x_0^{(k)} \sim \pi(x_0)$ and setting $\tilde{w}_0^{(k)} = 1/N$ for $k = 1, 2, \ldots, N$.

For $t = 1, 2, \ldots, T$:

1. Propagate each $x_{t-1}^{(k)}$ forward via $\pi(x_t^{(k)}|x_{t-1}^{(k)})$ for $k = 1, 2, \ldots, N$;
2. Evaluate the weights
$$w_t^{(k)} = \pi(y_t|x_t^{(k)})w_{t-1}^{(k)}, \quad \text{for } k = 1, 2, \ldots, N$$
and normalise using $\tilde{w}_t^{(k)} = w_t^{(k)} / \sum_{k=1}^{N} w_t^{(k)}$;
3. Resample N times with replacement from $\left\{x_t^{(k)}, \tilde{w}_t^{(k)}\right\}$ using the normalised weights $\tilde{w}_t^{(k)}$ as probabilities. Set each $w_t^{(k)} = 1/N$.

The Liu and West algorithm circumvents issues arising from errors in the kernel density estimate by employing a shrinkage modification of kernel smoothing. In particular, at each step, the sample is propagated forward via

$$\theta_{t+1}^{(k)} \sim N(m_t^{(k)}, s^2 V_t)$$

where $m_t^{(k)} = a\theta_t^{(k)} + (1-a)\bar{\theta}_t$, $V_t = \sum_{k=1}^{N}(\theta_t^{(k)} - \bar{\theta}_t)(\theta_t^{(k)} - \bar{\theta}_t)'/N$ and $\bar{\theta}_t = \sum_{k=1}^{N}\theta_t^{(k)}/N$. The constant scaling factor s in the variance can be interpreted as a tuning parameter to be chosen by the practitioner, and its value determines the value of a via the relationship $a = \sqrt{1-s^2}$. It is recommended in Liu and West (2001) to determine the values of s and a using a discount factor δ, where $s^2 = 1 - ((3\delta - 1)/2\delta)^2$, with a suggestion that values of δ close to 0.99 are typically more relevant. For clarity, it should be noted that the t subscript on the parameter vector θ_t is to indicate that this sample is from the time t posterior and does not imply that the parameters are assumed time-varying. The full Liu and West algorithm is given in Algorithm 6.

2.8.5 Storvik particle filter

One clear issue with the Liu and West algorithm is the arbitrary selection of the tuning parameter s or the discount factor δ equivalently. Whilst guidance is available (see Liu and West, 2001), it's not clear what value s or δ should take to optimise the inference scheme.

To remove this ambiguity, Storvik (2002) proposed a new algorithm which marginalises the parameters out of the posterior distribution via further factorisation of the target density. Specifically, assuming that the conditional distribution $\pi(\theta|x_{0:t}, y_{1:t})$ is

analytically tractable and can be summarised by some low dimensional sufficient statistics $T_t := T_t(x_{0:t}, y_{1:t})$, the target can be factorised as follows

$$\pi(x_{0:t}, \theta | y_{1:t}) \propto \pi(x_{0:t-1} | y_{1:t-1}) \pi(\theta | T_{t-1}) \pi(x_t | x_{t-1}, \theta) \pi(y_t | x_t, \theta)$$

by noticing that $\pi(x_{0:t-1}, \theta | y_{1:t-1}) = \pi(x_{0:t-1} | y_{1:t-1}) \pi(\theta | x_{0:t-1}, y_{1:t})$. The inference scheme then follows by propagating the state vector forward via $\pi(x_t | x_{t-1}, \theta)$, weighting according to $\pi(y_t | x_t, \theta)$ and finally rejuvenating the parameter vector via $\pi(\theta | T_t)$. The full algorithm is given in Algorithm 7.

2.8.6 Discussion

This section has considered several algorithms for performing parameter and state inference sequentially. For mathematical convenience, the target takes the form

Algorithm 6 Liu and West filter

Initialise with a sample $\left\{ x_0^{(k)}, \theta_0^{(k)}, \tilde{w}_0^{(k)} \right\}_{k=1}^{N}$ by sampling $x_0^{(k)} \sim \pi(x_0)$, $\theta_0^{(k)} \sim \pi(\theta)$, and setting $w_0^{(k)} = 1$ for $k = 1, 2, \ldots, N$.

For $t = 1, 2, \ldots, T$:

1. Sample $\theta_t^{(k)} \sim N(m_{t-1}^{(k)}, s^2 V_{t-1})$ for $k = 1, 2, \ldots, N$;
2. Propagate each $x_{t-1}^{(k)}$ forward via $\pi(x_t^{(k)} | x_{t-1}^{(k)}, \theta_t^{(k)})$ for $k = 1, 2, \ldots, N$;
3. Evaluate the weights
$$w_t^{(k)} = \pi(y_t | x_t^{(k)}, \theta_t^{(k)}) w_{t-1}^{(k)}, \quad \text{for } k = 1, 2, \ldots, N$$
and normalise using $\tilde{w}_t^{(k)} = w_t^{(k)} / \sum_{k=1}^{N} w_t^{(k)}$;
4. Resample N times with replacement from $\left\{ x_t^{(k)}, \theta_t^{(k)}, w_t^{(k)} \right\}$ using the normalised weights $\tilde{w}_t^{(k)}$ as probabilities.

$\pi(x_{0:t}, \theta|y_{1:t})$; in practice, the algorithms give draws from $\pi(x_t, \theta|y_{1:t})$. Although it is straightforward to resample the entire state trajectory, such paths will necessarily coalesce (see e.g. Karppinen et al., 2024). Hence, when full state inference is required, it is natural to perform additional backward smoothing steps (see e.g. Chopin and Papaspiliopoulos, 2020).

We also note that resampling need not be performed at every time point. It is possible to monitor some degeneracy criteria e.g. effective sample size (ESS) (see Elvira et al., 2022),

$$\text{ESS} = 1/\left(\sum_{k=1}^{N} \tilde{w}_t^{(k)}\right)^2$$

and then perform resampling if ESS is less than some fraction of N. When resampling takes place, we use multinomial resampling; for other resampling strategies, we refer the reader to Murray et al. (2016).

Algorithm 7 Storvik particle filter

Initialise with a sample $\left\{x_0^{(k)}, \theta^{(k)}, w_0^{(k)}\right\}_{k=1}^{N}$ by sampling $x_0^{(k)} \sim \pi(x_0)$, $\theta^{(k)} \sim \pi(\theta)$, and setting $w_0^{(k)} = 1$ for $k = 1, 2, \ldots, N$.

For $t = 1, 2, \ldots, T$:

1. Propagate each $x_{t-1}^{(k)}$ forward via $\pi(x_t^{(k)}|x_{t-1}^{(k)}, \theta^{(k)})$ for $k = 1, 2, \ldots, N$
2. Evaluate the weights
$$w_t^{(k)} = \pi(y_t|x_t^{(k)}, \theta^{(k)}) w_{t-1}^{(k)}$$
and normalise using $\tilde{w}_t^{(k)} = w_t^{(k)} / \sum_{k=1}^{N} w_t^{(k)}$;
3. Resample N times with replacement from $\left\{x_t^{(k)}, \theta_t^{(k)}, w_t^{(k)}\right\}$ using the normalised weights $\tilde{w}_t^{(k)}$ as probabilities;
4. Update the sufficient statistic $T_t^{(k)} := T_t(T_{t-1}^{(k)}, x_t^{(k)})$ for $k = 1, 2, \ldots, N$;
5. Rejuvenate the parameter sample by sampling $\theta^{(k)} \sim \pi(\theta|T_t^{(k)})$ for $k = 1, 2, \ldots, N$.

Chapter 3

Stochastic kinetic models of epidemics

In this chapter, we will introduce the well-studied class of statistical models known as stochastic kinetic models (SKMs), which are frequently used in epidemics to model disease transmission. The models reviewed in this chapter will form the principal focus of the analyses in later chapters, where we attempt to infer model parameters (as well as the unobserved cumulative incidence process). We introduce the Markov jump process (MJP) as a natural description of the underlying dynamics of an SKM, before introducing the SEIR and SIR models as specific examples. We then consider an approximation to the MJP, known as the linear noise approximation (LNA), for which we present the derivation and solution. Finally, we present some simulation examples, comparing an exact simulation method to approximate simulation methods.

3.1 Markov jump process

Suppose we are interested in a system in which there are u species $\mathcal{X}_1, \ldots, \mathcal{X}_u$ and v distinct interactions between them, denoted $\mathcal{R}_1, \ldots, \mathcal{R}_v$. As is standard across the literature (see e.g. Wilkinson, 2018), we write this system as a series of chemical reactions such as

$$\begin{aligned}
\mathcal{R}_1 &: \; a_{1,1}\mathcal{X}_1 + \cdots + a_{1,u}\mathcal{X}_u \xrightarrow{c_1} b_{1,1}\mathcal{X}_1 + \cdots + b_{1,u}\mathcal{X}_u \\
\mathcal{R}_2 &: \; a_{2,1}\mathcal{X}_1 + \cdots + a_{2,u}\mathcal{X}_u \xrightarrow{c_2} b_{2,1}\mathcal{X}_1 + \cdots + b_{2,u}\mathcal{X}_u \\
&\qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\qquad\quad \vdots \\
\mathcal{R}_v &: \; a_{v,1}\mathcal{X}_1 + \cdots + a_{v,u}\mathcal{X}_u \xrightarrow{c_v} b_{v,1}\mathcal{X}_1 + \cdots + b_{v,u}\mathcal{X}_u
\end{aligned}$$

where the $a_{i,j}$ and the $b_{i,j}$ are the stoichiometries of the system and the c_i are the reaction rates. Often, it is useful for us to work with the *net effect matrix* A, whose (i,j)th element is given by $b_{i,j} - a_{i,j}$. When written in this form, the system is referred to as a *reaction network*, and these can be used to describe the interactions between all species in the system to an arbitrary level of detail.

At any time t the population of the jth species $\mathcal{X}_j$ is denoted $X_{j,t}$ and so the state of the system at time t is the u-vector $X_t = (X_{1,t}, \ldots, X_{u,t})'$. When a reaction of type $\mathcal{R}_i$ occurs, the state vector X_t changes according to the ith row of the net effect matrix, which we denote by A_i. Equivalently, it's possible to use the *stoichiometry matrix*, S, which is given as the matrix transpose of the net effect matrix, i.e. $S = A^T$. Then, the state vector changes according to the ith column of S.

The dynamics of $\{X_t, t \geq 0\}$ are most naturally described by a Markov jump process (MJP), which is a continuous time, discrete valued Markov process. Similarly, if we denote by $N_{i,t}$ the number of occurrences of the reaction $\mathcal{R}_i$ up to time t, then the

counting process $\{N_t, t \geq 0\}$, where $N_t = (N_{1,t}, N_{2,t}, \ldots, N_{v,t})'$, is also an MJP. It follows that the two processes are linked via the equation

$$X_t = x_0 + \sum_i A_i' N_{i,t} \tag{3.1}$$

where x_0 denotes the initial state of the system. In the context of epidemics, the processes $\{X_t, t \geq 0\}$ and $\{N_t, t \geq 0\}$ are referred to as the latent prevalence and cumulative incidence processes, respectively; this thesis will focus on scenarios where observations are noisy measurements of the latter, i.e. the incidence process.

As discussed in Golightly and Gillespie (2013), under fairly weak assumptions, we can take the rate of any given reaction to be constant in a small enough time interval. We denote the rate constant associated with reaction type $\mathcal{R}_i$ by c_i and let the hazard of this reaction occurring be $h_i(X_t, c_i)$. Under the assumption of mass-action stochastic kinetics, and assuming a homogeneously mixing population, we can express these hazard functions, up to a constant of proportionality, in the following way

$$h_i(X_t, c_i) = c_i \prod_{j=1}^{n} \binom{X_{j,t}}{a_{i,j}}.$$

This is the rate constant multiplied by a product of binomial coefficients expressing the number of ways the reaction can occur. We also assume that the epidemic of interest takes place in a fixed area, so that the rate law is equivalent to both a frequency and density dependent approach (Begon et al., 2002).

3.2 Example: The SIR and SEIR models

Of particular relevance to this thesis will be the SIR (Andersson and Britton, 2012; Kermack and McKendrick, 1927) and SEIR models (Hethcote, 2000), within which a population of fixed size N_{pop} is classified into compartments consisting of susceptible (S), exposed (E), infectious (I) and removed (R) individuals. In what follows, we describe the stochastic SEIR model as the most general case, and make clear how the SIR model can be obtained as a simplification thereof. The SEIR compartment

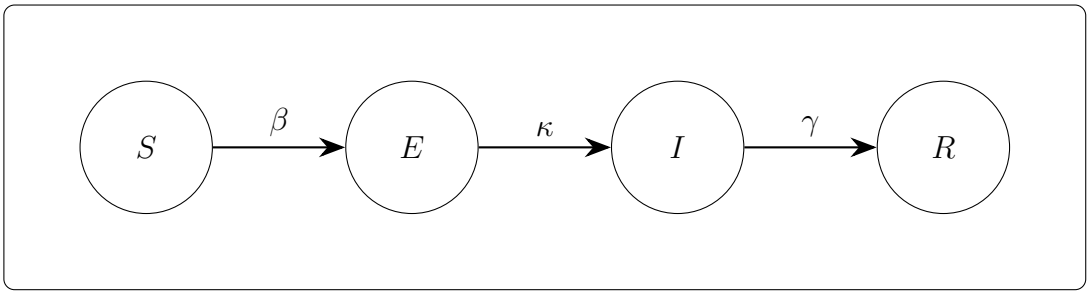

Figure 3.1: SEIR compartment model.

model is shown in Figure 3.1. Transitions between compartments can be described by the set of pseudo-reactions given by

$$\mathcal{R}_1 : \; S + I \xrightarrow{\beta} E + I, \quad \mathcal{R}_2 : \; E \xrightarrow{\kappa} I, \quad \mathcal{R}_3 : \; I \xrightarrow{\gamma} R.$$

The first transition describes contact of an infective individual with a susceptible and with the net effect resulting in an exposed individual and one fewer susceptible. The second transition moves an exposed individual to the infected class, and the final transition accounts for an infected individual's removal (recovered with immunity, quarantined or dead). The components of $\theta = (\beta, \kappa, \gamma)'$ denote contact, infection and removal rates. Note that setting $\kappa = 0$ and substituting E for I gives the set of

pseudo-reactions governing the SIR model as

$$\mathcal{R}_1 : \ S + I \xrightarrow{\beta} 2I, \quad \mathcal{R}_2 : \ I \xrightarrow{\gamma} R.$$

The resulting compartment model is given in Figure 3.2.

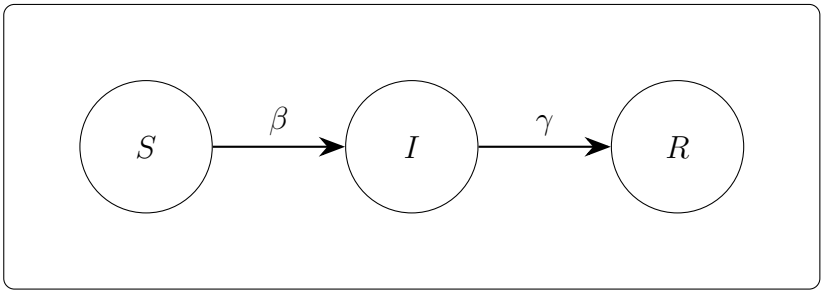

Figure 3.2: SIR compartment model.

Often of interest is a quantity known as the *basic reproduction number*, typically denoted by R_0. In the context of the SEIR and, subsequently, the SIR model, the basic reproduction number is the expected number of individuals to whom a single infected individual will transmit the modelled disease. Thus, we can define $R_0 = N_{\text{pop}}\beta/\gamma$.

Let $X_t = (S_t, E_t, I_t)'$ denote the numbers in each state at time $t \geq 0$ and note that $R_t = N_{\text{pop}} - S_t - E_t - I_t$ for all $t \geq 0$. Assuming that, for small Δt, at most one event can occur over a time interval $(t, t + \Delta t]$ and that the state of the system at time t is $x_t = (s_t, e_t, i_t)'$, the MJP $\{X_t, t \geq 0\}$ is characterised by transition probabilities

of the form

$$\begin{aligned}
\mathbb{P}(X_{t+\Delta t} = (s_t - 1, e_t + 1, i_t)' | x_t, \theta) &= \beta s_t i_t \, \Delta t + o(\Delta t), \\
\mathbb{P}(X_{t+\Delta t} = (s_t, e_t - 1, i_t + 1)' | x_t, \theta) &= \kappa e_t \, \Delta t + o(\Delta t), \\
\mathbb{P}(X_{t+\Delta t} = (s_t, e_t, i_t - 1)' | x_t, \theta) &= \gamma i_t \, \Delta t + o(\Delta t), \\
\mathbb{P}(X_{t+\Delta t} = (s_t, e_t, i_t)' | x_t, \theta) &= 1 - (\beta s_t i_t + \kappa e_t + \gamma i_t) \, \Delta t + o(\Delta t),
\end{aligned}$$

and $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$. Similarly, the cumulative incidence of contact, infection and removal events $\{N_t, t \geq 0\}$ is an MJP governed by the transition probabilities

$$\begin{aligned}
\mathbb{P}(N_{t+\Delta t} = (n_{se} + 1, n_{ei}, n_{ir})' | n_t, x_t, \theta) &= \beta s_t i_t \, \Delta t + o(\Delta t), \\
\mathbb{P}(N_{t+\Delta t} = (n_{se}, n_{ei} + 1, n_{ir})' | n_t, x_t, \theta) &= \kappa e_t \, \Delta t + o(\Delta t), \\
\mathbb{P}(N_{t+\Delta t} = (n_{se}, n_{ei}, n_{ir} + 1)' | n_t, x_t, \theta) &= \gamma i_t \, \Delta t + o(\Delta t), \\
\mathbb{P}(N_{t+\Delta t} = (n_{se}, n_{ei}, n_{ir})' | n_t, x_t, \theta) &= 1 - (\beta s_t i_t + \kappa e_t + \gamma i_t) \, \Delta t + o(\Delta t),
\end{aligned}$$

where $n_t = (n_{se}, n_{ei}, n_{ir})'$ denotes the contact, infection and removal events.

The instantaneous rate or hazard function $h(x_t) = (h_1(x_t), h_2(x_t), h_3(x_t))'$ for the SEIR model is given by

$$\begin{aligned}
h(x_t) &= \lim_{\Delta t \to 0} \mathbb{P}(X_{t+\Delta t} | x_t, \theta)/\Delta t \\
&= (\beta s_t i_t, \kappa e_t, \gamma i_t)'
\end{aligned}$$

and we suppress the dependence of the hazard function on θ to simplify the notation.

Finally, the net effect matrix for this model is given by

$$A = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

where it is understood that each row describes the effect on each component of $X_t = (S_t, E_t, I_t)'$ by the respective occurrence of a contact, infection or removal event. Using this and recalling the relationship between $\{X_t, t \geq 0\}$ and $\{N_t, t \geq 0\}$ given in Equation (3.1), we can re-write the hazard function explicitly in terms of the cumulative incidence process. The resulting hazard function for the SEIR model is therefore given by

$$\tilde{h}(n_t) = (\beta[s_0 - n_{se}][i_0 + n_{ei} - n_{ir}], \kappa[e_0 + n_{se} - n_{ei}], \gamma[i_0 + n_{ei} - n_{ir}])'$$

again suppressing dependence on θ for brevity. It should be clear that the hazard function for the SIR model can be obtained by taking the first and last components of the above expressions and swapping any occurrences of n_{se} or n_{ei} with n_{si}, resulting in

$$h(x_t) = (\beta s_t i_t, \gamma i_t)'$$

and

$$\tilde{h}(n_t) = (\beta[s_0 - n_{si}][i_0 + n_{si} - n_{ir}], \gamma[i_0 + n_{si} - n_{ir}])'.$$

Given x_0 and θ, generating exact realisations of the incidence process, and therefore, the stochastic SEIR model is straightforward and can be achieved by using well-known simulation algorithms from the stochastic kinetic models literature (see, e.g. Wilkinson, 2018). The simplest approach is Gillespie's direct method (Gillespie,

1977), which simulates the time to the next event as an exponential random variable with rate $h_0(x_t) = \sum_{i=1}^{3} h_i(x_t)$. The event that occurs will be of type i (with 1 = exposure, 2 = infection, 3 = removal) with probability proportional to $h_i(x_t)$. The full algorithm is presented in Algorithm 8.

3.3 Time discretisation

One aim of this thesis is to produce efficient algorithms for Bayesian inference in large scale epidemic models. Although the exactness of the direct method is extremely desirable, in certain scenarios it can become very inefficient to simulate every reaction event. This is particularly evident when working with large data sets spanning long periods of time, as is often the case in statistical epidemiology. Significant savings can be made by approximating the MJP in a way that still maintains the overall kinetics of the model. To this end, we consider two approximations of the

Algorithm 8 Gillespie's direct method

Set $t = 0$ and let the initial state of the system, i.e. the state at time $t = 0$, be $x = (x_{1,0}, \ldots, x_{u,0})'$. Initialise the vector of rate constants θ.

1. Calculate the hazard function $h_i(x, \theta_i)$ of each reaction $\mathcal{R}_i$ for $i = 1, \ldots, v$;
2. Calculate the total hazard $h_0(x, \theta) = \sum_{i=1}^{v} h_i(x, \theta_i)$;
3. Simulate the inter-event time $t' \sim \text{Exp}(h_0(x, \theta))$ and set $t := t + t'$;
4. Simulate the reaction index $j \in \{1, \ldots, v\}$ where each j has probability given by $h_j(x, \theta_j)/h_0(x, \theta)$;
5. Update the state vector x according to the j-th column of the stoichiometry matrix S. That is, set $x := x + S^j$;
6. Output the state vector x and the current time t;
7. If $t < T_{\max}$, return to step 1;

MJP in this section, namely the Poisson leap and the chemical Langevin equation (CLE) (see Golightly and Gillespie, 2013, for an in-depth discussion).

3.3.1 The Poisson leap

To begin, consider a time interval $(t, t+\Delta t]$ and denote by $\Delta N_t = (\Delta N_{1,t}, \Delta N_{2,t}, \Delta N_{3,t})'$ the length-3 vector containing the number of events of each type (exposure, infection, removal) over this interval. We make the assumption that Δt is small enough to reasonably assume that $X_s \approx x_t$ for $s \in (t, t+\Delta t)$ and consequently, the reaction hazards $h_i(x_t, \theta)$ remain constant almost surely. The individual components of ΔN_t therefore follow independent Poisson processes, with rates given by $h_i(x_t, \theta)\Delta t$. It should then be clear that the ith component of ΔN_t follows a Poisson distribution with rate $h_i(x_t, \theta)\Delta t$. Explicitly,

$$\Delta N_{1,t} \sim \text{Po}(\beta\, s_t i_t \Delta t), \quad \Delta N_{2,t} \sim \text{Po}(\kappa\, e_t \Delta t), \quad \Delta N_{3,t} \sim \text{Po}(\gamma\, i_t \Delta t).$$

Therefore, given an initial state x_0 and a parameter vector θ, we can simulate approximate realisations from the model by repeatedly simulating ΔN_t, updating the prevalence process according to Equation (3.1), then incrementing time $t := t + \Delta t$. The full algorithm is presented in Algorithm 9.

Note that we can choose the time step, Δt. This essentially controls the balance between the approximation's accuracy and the simulation's relative speed up. Choosing small values of Δt increases the accuracy of the approximation at the expense of an increased computational cost, whereas choosing large values of Δt will result in a scheme favouring computational efficiency over the accuracy of the approximation. Figure 3.3 provides a visual representation of the effect of changing Δt on the

accuracy of the approximation.

The top left panel shows the mean and 95% credible interval from 10^4 simulations of the cumulative incidence of infection events, $N_{1,t}$, from the SIR model via Gillespie's direct method. For this illustration, we use initial values and parameter values consistent with the boarding school epidemic from BMJ News and Notes (1978). That is, we take $x_0 = (762, 5)'$ and $\theta = (\exp(-6), 0.5)'$. The remaining three panels (top right, bottom left and bottom right) are the same summaries arising from 10^4 simulations from the SIR model using the Poisson leap with varying time steps (0.01, 0.1 and 1, respectively). The mean incidence obtained from the exact simulation method is overlaid on the three Poisson leap plots in red for ease of comparison. Table 3.1 summarises the relative speed up of using the Poisson leap for each value of Δt compared to Gillespie's direct method.

We can clearly see the effect of increasing the time step on the accuracy of the approximation. Starting with the top right plot, which was generated using the Poisson leap with $\Delta t = 0.01$, we can see that the approximation is indistinguishable from the exact simulation method. This choice of time step was made such that the computational cost of simulating from the model 10^4 times using the Poisson leap

Algorithm 9 Simulation via the Poisson leap

Set $t = 0$ and let the initial state of the system, i.e. the state at time $t = 0$, be $x_0 = (x_{1,0}, \ldots, x_{u,0})'$. Initialise the vector of rate constants θ.

1. Calculate the hazard function $h_i(x_t, \theta)$ of each reaction $\mathcal{R}_i$ for $i = 1, \ldots, v$;
2. Simulate the incidence increment $\Delta N_{t+\Delta t}$ by drawing $\Delta N_{i,t+\Delta t} \sim Po(h_i(x_t, \theta)\Delta t)$, for $i = 1, 2, \ldots, v$;
3. Update the state vector x_t via $x_t = x_t + A'\Delta N_{t+\Delta t}$ and set $t := t + \Delta t$;
4. Output the state vector x_t and the current time t;
5. If $t < T_{max}$, return to step 1;

was approximately equal to simulating the same number of realisations using the direct method. The bottom left plot, generated using a time step of $\Delta t = 0.1$, still does a good job at capturing the dynamics of the cumulative incidence process, with summaries lying very close together throughout. Using this time step meant that the simulations finished approximately one-tenth of the time, indicating that $\Delta t = 0.1$ provides a good balance between accuracy and efficiency. Finally, the bottom right plot was produced using $\Delta t = 1$. Although the computational saving is very large, taking about one-hundredth of the time of the exact method simulations, it is clear that the approximation is not capturing the true process's dynamics very well.

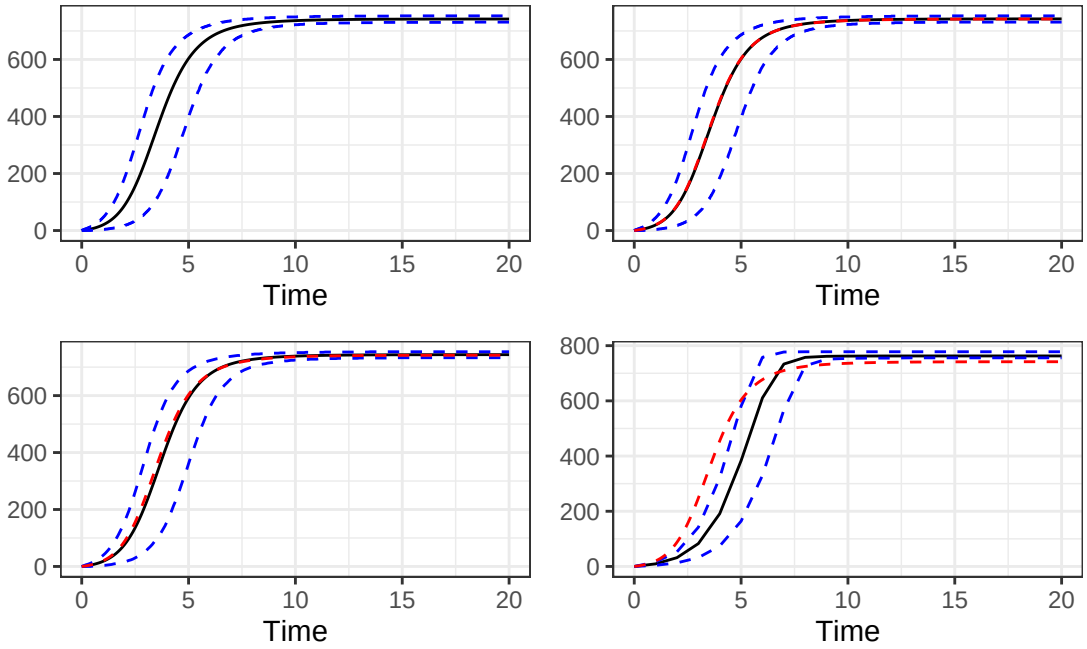

Figure 3.3: Mean (black) and 95% credible interval (blue) of 10^4 realisations of the $N_{t,1}$ process from the SIR model using the direct method (top left) and the Poisson leap with $\Delta t = 0.01$ (top right), $\Delta t = 0.1$ (bottom left) and $\Delta t = 1$ (bottom right). Overlaid on each Poisson leap plot is the direct method's mean line (red). All simulations assume $x_0 = (762, 5)'$ and $\theta = (\exp(-6), 0.5)'$.

3.3.2 The chemical Langevin equation

The chemical Langevin equation is a continuous state space approximation to the MJP, which can be derived in a number of more or less formal ways (see Gillespie, 2000, for an example of the former).

Consider the case where the interval in Section 3.3.1 becomes infinitesimal, i.e. let Δt approach zero such that the interval becomes $(t, t + dt]$. In this limit, the individual components of the cumulative incidence process, denoted by $dN_{i,t}$ for $i = 1, 2, \ldots, v$, are Poisson distributed with the rate given by $h_i(x_t, \theta)dt$. Stacking these quantities in the vector dN_t, it should be clear that

$$\mathrm{E}(dN_t) = h(x_t, \theta)dt, \qquad \mathrm{Var}(dN_t) = \mathrm{diag}\{h(x_t, \theta)\}dt.$$

An Itô stochastic differential equation (SDE), satisfied by a diffusion process $\{X_t, t \geq 0\}$, takes the form,

$$dX_t = a(X_t, t)dt + \sqrt{b(X_t, t)}dW_t,$$

where W_t is the standard Brownian motion process. The terms $a(X_t, t)$ and $b(X_t, t)$ are known as the *drift* and *diffusion* coefficients, respectively, and each correspond to the expected infinitesimal change in X_t and the infinitesimal variance, respectively.

Δt	0.01	0.1	1
$T_{DM}/T_{\Delta t}$	1	10	100

Table 3.1: Ratio of CPU time required to simulate 10^4 realisations of the cumulative incidence process from the SIR model using Gillespie's direct method (T_{DM}) to the CPU time required when using the Poisson leap for $\Delta t = 0.01, 0.1$ and 1 ($T_{\Delta t}$).

When viewed in integral form, as

$$X_t = X_0 + \int_0^t a(X_s, s)dt + \int_0^t \sqrt{b(X_t, t)}dW_t$$

it becomes clear that this SDE is made up of a deterministic part, and a stochastic (or random) part; the latter of which is referred to as the Itô stochastic integral. For a more comprehensive discussion of SDEs, we refer the reader to Oksendal (2013); Särkkä and Solin (2019).

Now, the Itô SDE that best matches the MJP representation of the incidence process is given by

$$dN_t = h(x_t, \theta)dt + \text{diag}\left\{\sqrt{h(x_t, \theta)}\right\} dW_t, \tag{3.2}$$

where W_t is a length-v vector of uncorrelated standard Brownian motion processes and $\text{diag}\left\{\sqrt{h(x_t, \theta)}\right\}$ is a $v \times v$ diagonal matrix with non-zero entries given by $\sqrt{h_i(x_t, \theta)}$ for $i = 1, 2, \ldots, v$. Note that the RHS of Equation (3.2) can be written explicitly in terms of n_t via Equation (3.1) which gives $x_t = x_0 + A'n_t$. We then further define the hazard function written in terms of n_t as $\tilde{h}(n_t) = h(x_0 + A'n_t, \theta)$ for which Equation (3.2) becomes

$$dN_t = \tilde{h}(n_t)dt + \text{diag}\{\sqrt{\tilde{h}(n_t)}\}\, dW_t. \tag{3.3}$$

Typically, the CLE is presented as an SDE approximation of the prevalence process $\{X_t, t \geq 0\}$, which can be obtained by applying the link between prevalence and incidence, given in Equation (3.1) as $X_t = x_0 + \sum_i A_i' N_{i,t}$, and noting that $dX_t = A'dN_t$. Consequently, the SDE for the infinitesimal dXt is

$$dX_t = A'h(X_t, \theta)dt + \sqrt{A' \,\text{diag}\{h(X_t, \theta)\} A}\, dW_t \tag{3.4}$$

Unfortunately, the CLE rarely has a tractable solution in practice. In the cases where the CLE does not permit an analytic solution, progress can be made via an approximate numerical solution such as the Euler-Maruyama method (see e.g. Kloeden and Platen, 1992). For this, we approximate the solution to the SDE over intervals of fixed length Δt via

$$N_{t+\Delta t} = N_t + h(x_t, \theta)\Delta t + \text{diag}\left\{\sqrt{h(x_t, \theta)}\right\} \Delta W_t \tag{3.5}$$

where $\Delta W_t := (W_{t+\Delta t} - W_t) \sim N(0, I_v \Delta t)$ and I_v is the $v \times v$ identity matrix.

3.4 Linear noise approximation (LNA)

The linear noise approximation (LNA) is most commonly presented as a Gaussian process approximation of the MJP description of the prevalence process $\{X_t, t \geq 0\}$ (see e.g. Ross et al. (2009); Fearnhead et al. (2014); Fuchs (2013) in the epidemic context and Ferm et al. (2008); Komorowski et al. (2009); Stathopoulos and Girolami (2013) in a wider systems biology context). As in Fintzi et al. (2022), we additionally require an approximation of the cumulative incidence process $\{N_t, t \geq 0\}$. We follow Fearnhead et al. (2014) by first deriving the LNA for a generic SDE, before applying the results to the CLE for the incidence process given in Equation (3.3) above.

3.4.1 Derivation of the LNA

Consider a generic SDE for a length-d vector X of the form

$$dX_t = a(X_t)dt + b(X_t)dW_t, \qquad X_0 = x_0 \tag{3.6}$$

where it is assumed that the stochastic term $b(X_t)$ is small compared to the drift. The SDE in (3.6) can be linearised by considering a partition of X_t of the form

$$X_t = \eta_t + R_t \tag{3.7}$$

where η_t is a deterministic process satisfying the ordinary differential equation (ODE)

$$\frac{d\eta_t}{dt} = a(\eta_t) \tag{3.8}$$

and $R_t = X_t - \eta_t$ is a residual stochastic process. Assuming that R_t is small over the time interval of interest and substituting Equation (3.7) into Equation (3.6), we obtain the typically intractable SDE

$$d(\eta_t + R_t) = \{a(\eta_t + R_t)\}\, dt + b(\eta_t + R_t)\, dW_t. \tag{3.9}$$

An approximate, tractable $\hat{R}_t$ can be obtained by Taylor expanding $a(X_t)$ and $b(X_t)$ about η_t. Retaining the first two terms in the expansion of the former and the first term in the expansion of the latter gives

$$d\hat{R}_t = F_t \hat{R}_t\, dt + b(\eta_t)\, dW_t \tag{3.10}$$

where F_t is the Jacobian matrix with (i,j)th element given by the partial derivative of the ith component of $a(\eta_t)$ with respect to the jth component of η_t, i.e.

$$(F_t)_{i,j} = \frac{\partial a_i(\eta_t)}{\partial \eta_{j,t}}.$$

3.4.2 Solution of the LNA

Given that the initial condition on $\hat{R}_t$ is either fixed or Gaussian, then Equation (3.10) is a linear combination of Gaussian distributions and is therefore itself Gaussian. Assuming that $\hat{R}_0 \sim N(\hat{r}_0, \hat{V}_0)$, this linear Gaussian structure allows an analytic solution of the form

$$\hat{R}_t | \hat{R}_0 = \hat{r}_0 \sim N(G_t \hat{r}_0, G_t \psi_t G_t')$$

where G_t is the fundamental matrix satisfying

$$\frac{dG_t}{dt} = F_t G_t, \quad G_0 = I_d \tag{3.11}$$

and ψ_t satisfies

$$\frac{d\psi}{dt} = G_t^{-1} b(\eta_t)^2 (G_t^{-1})', \quad \psi_0 = \hat{V}_0.$$

Recalling that $X_t = R_t + \eta_t$ leads to

$$(X_t | X_0 = \eta_0 + r_0) \sim N(\eta_t + G_t r_0, V_t) \tag{3.12}$$

where $V_t = G_t \psi_t G_t'$ satisfies

$$\frac{dV_t}{dt} = V_t F_t' + b(\eta_t)^2 + F_t V_t, \quad V_0 = 0_d. \tag{3.13}$$

Note that I_d and 0_d are the $d \times d$ identity and zero matrices respectively. The LNA for the process X_t is then summarised by Equation (3.12) and Equations (3.8), (3.11) and (3.13).

To apply this result to the CLE for the incidence process of an SKM given in

Equation (3.3), we first note that the drift term is given by

$$a(N_t) = \tilde{h}(n_t)$$

and the diffusion term is given by

$$b(N_t) = \text{diag}\left\{\sqrt{\tilde{h}(n_t)}\right\}.$$

Thus, the distribution of N_t under the LNA is

$$(N_t|N_0 = \eta_0 + r_0) \sim N(\eta_t + G_t r_0, V_t) \tag{3.14}$$

where η_t, G_t and V_t are solutions to the coupled ODE system given by

$$\frac{d\eta_t}{dt} = \tilde{h}(\eta_t), \tag{3.15}$$
$$\frac{dG_t}{dt} = F_t G_t, \tag{3.16}$$
$$\frac{dV_t}{dt} = V_t F_t' + \text{diag}\left\{\tilde{h}(\eta_t)\right\} + F_t V_t. \tag{3.17}$$

Thus, the LNA for the cumulative incidence process is then summarised by Equation (3.14) and Equations (3.15) to (3.17). The algorithm combining these steps to simulate a realisation of N_t is given in Algorithm 10.

3.4.3 Restarting the LNA

As noted by Fearnhead et al. (2014); Golightly and Gillespie (2013) among others, the solution of the LNA over large time horizons can become disjointed from the true MJP, leading to a poor approximation that worsens over time. Fearnhead

et al. (2014) propose a solution to this issue; at each observation time t_i, re-initialise the ODE for η_t by setting $\eta_{t_i} = n_{t_i}$. By restarting the ODE in this way, we are recentring the point about which the Taylor expansion is made, aligning it with the most up-to-date estimate of n_t and therefore minimising the impact of the higher order terms, which we are discarding at each step. This reinitialisation, or 'restart', has the effect of setting $\hat{r}_t = 0$ for all t, meaning the ODE for G_t need never be solved, reducing the dimensionality of the problem, and therefore increasing the computational efficiency. The algorithm for simulating a realisation of N_t using the LNA with restart is given in Algorithm 11 below.

3.4.4 Performance comparison

In Figure 3.4 we compare the competing simulation methods. The top row of plots show the summaries (mean and 95% credible interval) obtained from 10^4 realisations of the $N_{t,1}$ process from the SIR model using the direct method (left) and the Poisson leap with $\Delta t = 0.01$ (right). The bottom row of plots shows the same summaries from 10^4 realisations obtained using the LNA (left) and the LNA with restart (right). We can immediately see that both versions of the LNA do a good

Algorithm 10 Simulation via the LNA

1. Initialise with a vector of rate constants, θ, and initial conditions η_0, r_0, $G_0 = I_d$ and $V_0 = 0_d$. Set $t = 0$;
2. Solve the system of ODEs satisfied by η_t, G_t and V_t over the interval $(t, t+\Delta t]$;
3. Simulate $N_{t+\Delta t}$ from its conditional distribution $(N_{t+\Delta t}|N_0 = \eta_0 + r_0) \sim N(\eta_{t+\Delta t} + G_{t+\Delta t} r_0, V_{t+\Delta t})$;
4. Set $t := t + \Delta t$ and
5. Output t and N_t. If $t < T_{\max}$, return to step 2.

job of recovering the mean behaviour of the MJP, highlighted by the clear overlap between the black and red lines on these plots. This is a good indicator that the LNA provides an accurate approximation of the MJP for the incidence process, which we will leverage in later chapters to access a tractable form of the observed data likelihood. Furthermore, there is a clear improvement in accuracy from LNA without restart to LNA with restart, highlighting the impact of recentring the Taylor expansion at each observation time.

3.5 Time varying contact rate

In practice, assuming that the contact rate in the SEIR model (or infection rate in the SIR model) remains constant throughout the epidemic may be unreasonable. We, therefore, follow Dureau et al. (2013); Spannaus et al. (2022); Wadkin et al. (2022) among others and describe the contact rate via an Itô stochastic differential equation (SDE).

Let $\{\beta_t, t \geq 0\}$ denote the infection process and consider $\tilde{\beta}_t = \log(\beta_t)$, assumed to

Algorithm 11 Simulation via the LNA (with restart)

1. Initialise with a vector of rate constants, θ, and initial conditions $\eta_0 = n_0$ and $V_0 = 0_d$. Set $t = 0$;
2. Solve the system of ODEs satisfied by η_t and V_t over the interval $(t, t + \Delta t]$;
3. Simulate $N_{t+\Delta t}$ from its conditional distribution $(N_{t+\Delta t}|N_0 = \eta_0 + r_0) \sim N(\eta_{t+\Delta t}, V_{t+\Delta t})$;
4. Set $t := t + \Delta t$, $\eta_t = n_t$ and $V_t = 0$;
5. Output t and N_t. If $t < T_{\max}$, return to step 2.

satisfy a time-homogeneous SDE, parameterised by λ and of the form

$$d\tilde{\beta}_t = a(\tilde{\beta}_t, \lambda)dt + b(\tilde{\beta}_t, \lambda)dW_t \tag{3.18}$$

where $\{W_t, t \geq 0\}$ is a standard Brownian motion process. Choice of the drift and diffusion functions $a(\cdot)$ and $b(\cdot)$ in Equation (3.18) determine the properties of $\tilde{\beta}_t$. For example, $a(\tilde{\beta}_t, \lambda) = 0$ and $b(\tilde{\beta}_t, \lambda) = \lambda^{-1/2}$ gives a generalised Brownian motion process which admits an analytic solution over a time interval $(t, t + \Delta t]$ as

$$\tilde{\beta}_{t+\Delta t} = \tilde{\beta}_t + \lambda^{-1/2}\Delta W_t \tag{3.19}$$

where $\Delta W_t := (W_{t+\Delta t} - W_t) \sim \mathrm{N}(0, \Delta t)$ and λ is a precision parameter. In cases

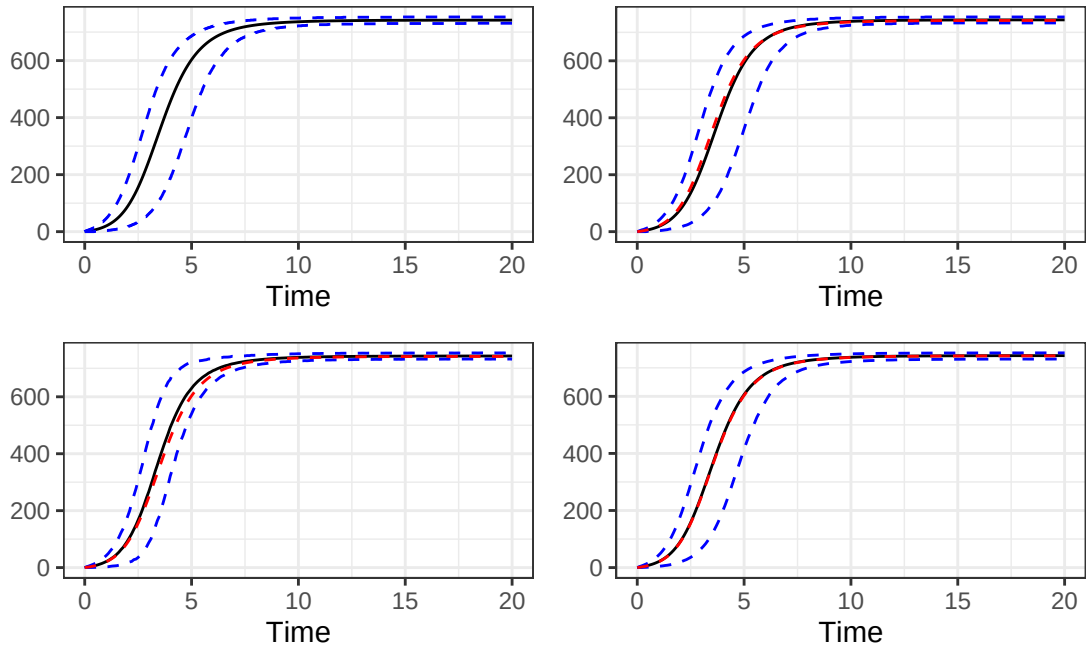

Figure 3.4: Mean (black) and 95% credible interval (blue) of 10^4 realisations of the $N_{t,1}$ process from the SIR model using the direct method (top left), the Poisson leap with $\Delta t = 0.1$ (top right), the LNA (bottom left) and the LNA with restart (bottom right). Overlaid on each of the Poisson leap and LNA plots is the mean line from the direct method (red). All simulations assume $x_0 = (762, 5)'$ and $\theta = (\exp(-6), 0.5)'$.

where the SDE (3.18) can't be solved analytically, a numerical approximation can be sought. Recall that the simplest such approximation is given by the Euler-Maruyama discretisation

$$\tilde{\beta}_{t+\Delta t} = \tilde{\beta}_t + a(\tilde{\beta}_t, \lambda)\Delta t + b(\tilde{\beta}_t, \lambda)\Delta W_t. \tag{3.20}$$

Replacing β with a time-varying $\beta_t = \exp(\tilde{\beta}_t)$ dictates that the hazard h_1 of the contact reaction is no longer constant between event occurrences. Generating exact realisations from the resulting SEIR model is no longer straightforward unless β_t can be bounded above, permitting the use of Poisson thinning (Lewis and Shedler, 1979).

We consider a time discretised stochastic SEIR model, as in Section 3.3.1, and subsequently base our inferential approach on the resulting approximation. In particular, the practitioner can choose the discretisation level to balance accuracy and computational efficiency. The time discretised SEIR model (dSEIR) with time-varying contact rates is given as

$$\begin{cases} X_{t+\Delta t} = x_t + A'\Delta N_t, \\ \tilde{\beta}_{t+\Delta t} = \tilde{\beta}_t + a(\tilde{\beta}_t, \lambda)\Delta t + b(\tilde{\beta}_t, \lambda)\Delta W_t. \end{cases} \tag{3.21}$$

Simulating exact realisations from dSEIR is now straightforward through a modification of Algorithm 9 to include the evolution of $\tilde{\beta}_t$; see Algorithm 12 for further details.

Including a time-varying contact rate within the LNA framework can be achieved by appending $\tilde{\beta}_t$ onto the cumulative incidence process N_t. That is, we let $N_{4,t} = \tilde{\beta}_t$ (or $N_{3,t} = \tilde{\beta}_t$ if we are working with the SIR model) and so, combining Equation (3.18)

with Equation (3.2) gives a coupled SDE for $N_t = (N_{1,t}, N_{2,t}, N_{3,t}, N_{4,t})'$ of the form

$$\begin{aligned} dN_t = &\left\{\tilde{h}_1(n_t), \tilde{h}_2(n_t), \tilde{h}_3(n_t), a(n_{4,t})\right\}' dt \\ &+ \text{diag}\left\{\sqrt{\tilde{h}_1(n_t)}, \sqrt{\tilde{h}_2(n_t)}, \sqrt{\tilde{h}_3(n_t)}, b(n_{4,t})\right\} dW_t \end{aligned} \tag{3.22}$$

where W_t is a length-4 vector of uncorrelated standard Brownian motion processes. The LNA of Equation (3.22) follows in the same way as Section 3.4, albeit with the Jacobian matrix F_t redefined as

$$F_t = \begin{pmatrix} \exp(\eta_{3,t})(s_0 - i_0 - 2\eta_{t,1} + \eta_{t,2}) & \gamma & 0 \\ \exp(\eta_{3,t})(\eta_{t,1} - s_0) & -\gamma & 0 \\ \exp(\eta_{3,t})(s_0 - \eta_{t,1})(i_0 + \eta_{t,1} - \eta_{t,2}) & 0 & \frac{\partial a(\eta_{3,t})}{\partial \eta_{3,t}} \end{pmatrix}'$$

and the RHS of Equations (3.15) and (3.17) augmented to include $a(n_{4,t})$ and $b^2(n_{4,t})$.

Naturally, allowing additional flexibility via the introduction of a time-varying contact rate poses potential identifiability issues. Although we don't pursue it in this thesis, we envisage the need for strong prior knowledge to overcome such problems.

Algorithm 12 Simulation of dSEIR with time-varying contact rate

Input: parameters $\theta = (\kappa, \gamma)'$, infection process parameters λ, initial conditions $x_0 = (s_0, e_0, i_0)'$ and $\tilde{\beta}_0$, time step Δt and end time $T = m\Delta t$.
For $j = 0, \ldots, m-1$:

1. Set $t := j\Delta t$ and calculate the hazard function $h(x_t, \tilde{\beta}_t) = (\exp(\tilde{\beta}_t)\, s_t i_t, \kappa\, e_t, \gamma\, i_t)'$;
2. Simulate the incidence increment ΔN_t by drawing $\Delta N_{i,t} \sim \text{Po}(h_i \Delta t)$, $i = 1, 2, 3$. Set $x_{t+\Delta t} = x_t + A'\Delta N_t$;
3. Simulate $\Delta W_t \sim \text{N}(0, \Delta t)$ and set $\tilde{\beta}_{t+\Delta t} = \tilde{\beta}_t + a(\tilde{\beta}_t, \lambda)\Delta t + b(\tilde{\beta}_t, \lambda)\Delta W_t$.

Output: trajectory $\{(x_t, \tilde{\beta}_t), t = 0, \Delta t, \ldots, T\}$.

3.6 Discussion

This chapter considered several stochastic representations of epidemic models. The most natural Markov jump process representation of a stochastic epidemic can be derived from a pseudo-reaction system coupled with a rate law. The computational cost of simulating every reaction event motivates approximations which either discretise time (e.g. the Poisson leap and CLE) or seek a tractable approximation to the transition density (e.g. the LNA). The Poisson leap approach can be made arbitrarily accurate by choosing an arbitrarily small time-step. The accuracy of the LNA, on the other hand, can be understood through the validity of the CLE, which can be formally derived from the MJP representation (see e.g. Kurtz, 1970, 1972), as a large volume limit. Nevertheless, in specific applications, Fuchs (2013) (see also Fintzi et al., 2022) recommend judging validity of the LNA via direct comparison with the MJP (e.g. via simulation).

Chapter 4

Bayesian inference for SKMs via batch methods

In this chapter, we consider the problem of performing fully Bayesian inference for the parameters (and unobserved dynamic processes) governing the S(E)IR model based on incidence observations that we assume are incomplete and subject to measurement error. We describe the observation model before considering the inference task, which we split into two categories: inference which leverages the tractability of the LNA for analytic evaluation of the observed data likelihood and inference based on the Poisson leap in which we obtain estimates of the observed data likelihood in such a way as to allow for exact inference. For ease of exposition, we assume a constant infection rate, but note that extension of the methodology to a time-varying infection rate is straightforward, as considered in Chapter 3.

4.1 Observation model

Without loss of generality, consider data $y = (y_{t_1}, \ldots, y_{t_L})'$ at integer times, where y_{t_i} is a (partial) observation on the cumulative incidence $\Delta N_{t_i} = N_{t_i} - N_{t_{i-1}}$ over a time interval $(t_{i-1}, t_i]$. Commonly used models for incidence data include additive Gaussian noise (Dureau et al., 2013), the Binomial distribution (Cauchemez and Ferguson, 2008) and the Negative Binomial distribution (Lloyd-Smith, 2007; Fintzi et al., 2022; Spannaus et al., 2022). The latter two models are typically used under the assumption of underreporting and to capture overdispersion, e.g. over-reporting, respectively. In large population settings, they may be well approximated by a Gaussian distribution which may offer computational benefits when combined with a Gaussian description of the underlying epidemic dynamics, such as the LNA described in Section 3.4 (see also Finkenstädt et al., 2013, for a discussion on observation processes and the LNA). These models take the form

$$Y_{t_i} | \Delta N_{t_i} \sim \mathrm{N}\left(P' \Delta N_{t_i}, \sigma^2\right), \tag{4.1}$$

$$Y_{t_i} | \Delta N_{t_i} \sim \mathrm{Bin}\left(P' \Delta N_{t_i}, \rho\right), \tag{4.2}$$

$$Y_{t_i} | \Delta N_{t_i} \sim \mathrm{NegBin}\left(\mu_i = \rho P' \Delta N_{t_i}, \sigma_i^2 = \mu_i + \mu_i^2/\nu\right) \tag{4.3}$$

for $i = 1, \ldots, L$. Here, P is a constant matrix allowing for observation of a subset of components of ΔN_{t_i} and ρ controls the accuracy of the observation process. In practice, we take $P' = (0, 1, 0)$ in the SEIR model and $P' = (1, 0)$ in the case of the SIR model, so that observations are noisy counts of new infections in a given time window. We assume that the observations are independent (given the latent process) and we let $\pi(y_{t_i} | \Delta n_{t_i}, \phi)$ denote the probability mass function linking y_{t_i} and $\Delta n_{t_i} = n_{t_i} - n_{t_{i-1}}$, with ϕ denoting the parameters governing the observation

model. For example, in Equation (4.3), $\phi = (\rho, \nu)'$ with ρ controlling the average proportion of cases seen and ν is the overdispersion parameter.

4.2 Inference task

We assume that interest lies in the vector of all static parameters $\psi = (\theta', \phi')'$, the latent incidence process $\{N_t, t_0 \leq t \leq t_L\}$ and the initial state vector x_0. Note that the initial state vector and incidence process are sufficient to determine the prevalence process $\{X_t, t_0 \leq t \leq t_L\}$ deterministically, through recursive application of

$$X_{t+\Delta t} = x_t + A' \Delta N_t.$$

In what follows we drop explicit dependence of the incidence process on x_0 from the notation for ease of exposition.

Recall that y_{t_i} denotes a noisy and incomplete observation on the cumulative incidence ΔN_{t_i} over $(t_{i-1}, t_i]$. We assume that the inter-observation interval $\Delta t = t_i - t_{i-1}$ is too large, over which reaction hazards cannot be assumed plausibly constant. We therefore partition each such time interval as

$$t_{i-1} = \tau_{i,0} < \tau_{i,1} < \cdots < \tau_{i,m} = t_i$$

with $\tau_{i,j} - \tau_{i,j-1} = \Delta\tau = \Delta t/m$. This allows the practitioner to choose the discretisation level m to balance the inferential model's accuracy and computational efficiency.

We let $\Delta n = (\Delta n_{\tau_{1,1}}, \Delta n_{\tau_{1,2}}, \ldots, \Delta n_{\tau_{L,m}})'$ denote the collection of incidences over sub-intervals $(\tau_{i,j-1}, \tau_{i,j}]$, for $i = 1, \ldots, L$ and $j = 1, \ldots, m$. Note that $\Delta n_{t_i} =$

$\sum_{j=1}^{m} \Delta n_{\tau_{i,j}}$ then gives the cumulative incidence over $(t_{i-1}, t_i]$. Upon ascribing a prior density $\pi(\psi)$ to ψ, Bayesian inference proceeds via the joint posterior

$$\pi(\psi, \Delta n|y) \propto \pi(\psi)\pi(\Delta n|\theta)\pi(y|\Delta n, \phi) \tag{4.4}$$

where

$$\pi(\Delta n|\theta) = \prod_{i=1}^{L}\prod_{j=1}^{m} \pi(n_{\tau_{i,j}}|n_{\tau_{i,j-1}}, \theta).$$

The form of $\pi(n_{\tau_{i,j}}|n_{\tau_{i,j-1}}, \theta)$ is dependent on the inferential model being considered. Under the Poisson leap, this becomes a product of Poisson random variables, leading to

$$\pi(\Delta n|\theta) = \prod_{i=1}^{L}\prod_{j=1}^{m}\prod_{k=1}^{3} \text{Po}(\Delta n_{k,\tau_{i,j}}; h_k(x_{\tau_{i,j-1}})\Delta\tau)$$

for the SEIR model. When basing inference on the LNA, $\pi(n_{\tau_{i,j}}|n_{\tau_{i,j-1}}, \theta)$ takes the form of the Gaussian transition density obtained from Equation (3.14) with the ODEs (3.15) and (3.17) integrated over $[t_{i-1}, t_i]$ with $\eta_{t_{i-1}} = n_{t_{i-1}}$ and $V_{t_{i-1}} = 0_2$. Note that in this case the residual $r_{t_{i-1}} = n_{t_{i-1}} - \eta_{t_{i-1}} = 0$ and subsequently the ODE satisfied by G_t in Equation (3.16) need not be integrated. Finally,

$$\pi(y|\Delta n, \phi) = \prod_{i=1}^{L} \pi(y_{t_i}|\Delta n_{t_i}, \phi).$$

Since the joint posterior in (4.4) will be intractable, we resort to Monte Carlo methods for generating samples of the parameters and latent dynamic process. A Gibbs sampler provides a natural mechanism for sampling (4.4), whereby one alternates between draws of $\psi|\Delta n, y$ and $\Delta n|\psi, y$. However, dependence between Δn and ψ can lead to poor mixing. For this reason, Fintzi et al. (2022) use a non-centred parameterisation whereby standard Gaussian innovations driving the generative form of the LNA are used as the effective components to be conditioned on. In what

follows, we take a different approach by marginalising out the latent process, either by further approximating the observation model in the non-Gaussian case or via (correlated) pseudo-marginal methods as presented in Section 2.7.

4.3 Marginalisation of the incidence process

The joint posterior density in (4.4) can be factorised as

$$\pi(\psi, \Delta n|y) = \pi(\psi|y)\pi(\Delta n|\psi, y) \tag{4.5}$$

where

$$\pi(\psi|y) \propto \pi(\psi)\pi(y|\psi). \tag{4.6}$$

The form of (4.5) suggests a two-step approach to inference whereby samples are first drawn from the marginal parameter posterior $\pi(\theta|y)$ in step 1, and then conditioned on in a second step when drawing samples of the latent process from $\pi(n|\theta, y)$. However, unless the observation model takes the linear Gaussian form of (4.1), and the LNA is used as the transmission model, neither the observed data likelihood $\pi(y|\theta)$ in Equation (4.6) nor the constituent densities in Equation (4.5) will be tractable, regardless of the inferential model used. The main focus of this chapter is exactly this *intractable* scenario, and we now consider two approaches to address it.

4.3.1 Analytic method via LNA

The LNA, when combined with the linear Gaussian observation model (4.1), permits analytic calculation of the observed data likelihood $\pi(y|\theta)$ and the conditional pos-

terior $\pi(n|\theta, y)$. Evaluation of the former can be efficiently achieved via a forward filter and draws from the latter via backward sampling. We apply these methods to the Binomial and Negative Binomial observation models in (4.2) and (4.3) through suitable Gaussian approximations thereof. We note that these approximations arise from the central limit theorem, and are therefore only valid when the number of trials becomes large. In our applications in Chapter 6, this condition is generally satisfied, however we acknowledge that this approximation is likely to break down in smaller epidemic settings. Additionally, we note that the skew-normal distribution can give a better approximation to the Negative Binomial (see Chang, 2008, for more details), however this would sacrifice the tractability of the observed data likelihood, which we aim to exploit. For reasons of brevity, we focus on the Binomial case, but note that our approach is easily extended to the Negative Binomial case. Where appropriate, we suppress the parameter vector θ from the notation for simplicity.

To make clear the two approximations to be used in the filtering recursions, consider the LNA written in state-space format over a time interval $(t_i, t_{i+1}]$, using a Gaussian approximation to the Binomial observation model. We have that

$$N_{t_{i+1}}|(N_{t_i} = n_{t_i}) \sim \mathrm{N}\left(\eta_{t_{i+1}} + G_{t_{i+1}}(n_{t_i} - \eta_{t_i}), V_{t_{i+1}}\right), \tag{4.7}$$

$$Y_{t_{i+1}}|(N_{t_{i+1}} = n_{t_{i+1}}, N_{t_i} = n_{t_i}) \sim \mathrm{N}\left(\rho P' \Delta n_{t_{i+1}}, \rho(1-\rho)P' \Delta n_{t_{i+1}}\right), \tag{4.8}$$

where $\eta_{t_{i+1}}$, $G_{t_{i+1}}$ and $V_{t_{i+1}}$ are obtained by integrating Equations (3.15) to (3.17) over $(t_i, t_{i+1}]$ with initial conditions of η_{t_i} (itself integrated from time 0), I_2 and 0_2. Although Equation (4.7) is linear in n_{t_i}, as noted in Section 3.4, η_t should be initialised at n_t. 'Restarting' the LNA in this way can avoid issues arising from the ODE solution becoming poor over long time intervals (see Fearnhead et al., 2014;

Minas and Rand, 2017, for an in-depth discussion). However, we now have that both Equations (4.7) and (4.8) involve nonlinear expressions of the latent process. Therefore, to permit the use of standard Kalman-filtering recursions (Kalman, 1960), we make further linear approximations.

Suppose that the filtering distribution at time t_i is $N_{t_i}|(Y_{1:t_i} = y_{1:t_i}) \sim \mathrm{N}(a_{t_i}, C_{t_i})$. Firstly, we set $\eta_{t_i} = a_{t_i}$, $V_{t_i} = C_{t_i}$ and integrate Equations (3.15) and (3.17) over $(t_i, t_{i+1}]$ to obtain $\eta_{t_{i+1}}$ and $V_{t_{i+1}}$. Finally, we replace $\Delta n_{t_{i+1}}$ in the variance of Equation (4.8) with $\Delta\widehat{n}_{t_{i+1}} := \mathrm{E}(\Delta N_{t_{i+1}}) = \eta_{t_{i+1}} - a_{t_i}$. Note that explicit conditioning of the expectation on $y_{1:t_i}$ has been suppressed for simplicity. The resulting linear and Gaussian state-space model is

$$N_{t_{i+1}}|(N_{t_i} = n_{t_i}) \sim \mathrm{N}\left(\eta_{t_{i+1}}, V_{t_{i+1}}\right), \tag{4.9}$$

$$Y_{t_{i+1}}|(N_{t_{i+1}} = n_{t_{i+1}}, N_{t_i} = n_{t_i}) \sim \mathrm{N}\left(\rho P'\Delta n_{t_{i+1}}, \rho(1-\rho)P'\Delta\widehat{n}_{t_{i+1}}\right). \tag{4.10}$$

In the remainder of this section, we derive the filtering recursions based on Equations (4.9) and (4.10) to compute the observed data likelihood and conditional posterior of the latent process.

We construct the observed data likelihood contribution $\pi(y_{t_{i+1}}|y_{1:t_i}, \theta)$ as follows. Conditional on $y_{1:t_i}$, we have that

$$\mathrm{Var}(\Delta N_{t_{i+1}}) = V_{t_{i+1}} + C_{t_i} - C_{t_i}G'_{t_{i+1}} - G_{t_{i+1}}C_{t_i}$$

where we have used that $\mathrm{Cov}(N_{t_{i+1}}, N_{t_i}) = G_{t_{i+1}}\mathrm{Var}(N_{t_i})$. Hence, combining with Equation (4.10) gives

$$\pi(y_{t_{i+1}}|y_{1:t_i}, \theta) = \mathrm{N}\left(y_{t_{i+1}}; \rho P'\mathrm{E}(\Delta N_{t_{i+1}}), \rho^2 P'\mathrm{Var}(\Delta N_{t_{i+1}})P + \widehat{\sigma}^2\right) \tag{4.11}$$

where $\widehat{\sigma}^2 = \rho(1-\rho)P'\Delta\widehat{n}_{t_{i+1}}$ is the observation variance in Equation (4.10). To update the filtering distribution, we construct the joint density of $N_{t_{i+1}}$ and $Y_{t_{i+1}}$ conditional on $Y_{1:t_i} = y_{1:t_i}$ as

$$\begin{pmatrix} N_{t_{i+1}} \\ Y_{t_{i+1}} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \eta_{t_{i+1}} \\ \rho P'\mathrm{E}(\Delta N_{t_{i+1}}) \end{pmatrix}, \begin{pmatrix} V_{t_{i+1}} & \mathrm{Cov}(N_{t_{i+1}}, Y_{t_{i+1}}) \\ \mathrm{Cov}(Y_{t_{i+1}}, N_{t_{i+1}}) & \rho^2 P'\mathrm{Var}(\Delta N_{t_{i+1}})P + \widehat{\sigma}^2 \end{pmatrix} \right\}$$

where $\mathrm{Cov}(N_{t_{i+1}}, Y_{t_{i+1}}) = \rho(V_{t_{i+1}} - G_{t_{i+1}}C_{t_i})P$. Hence, conditioning on $Y_{t_{i+1}} = y_{t_{i+1}}$ gives $N_{t_{i+1}} | \left(Y_{1:t_{i+1}} = y_{1:t_{i+1}}\right) \sim \mathrm{N}(a_{t_{i+1}}, C_{t_{i+1}})$ with mean

$$a_{t_{i+1}} = \eta_{t_{i+1}} + \mathrm{Cov}(N_{t_{i+1}}, Y_{t_{i+1}})(\rho^2 P'\mathrm{Var}(\Delta N_{t_{i+1}})P + \widehat{\sigma}^2)^{-1}(y_{t_{i+1}} - \rho P'\mathrm{E}(\Delta N_{t_{i+1}})) \tag{4.12}$$

and variance

$$C_{t_{i+1}} = V_{t_{i+1}} - \mathrm{Cov}(N_{t_{i+1}}, Y_{t_{i+1}})(\rho^2 P'\mathrm{Var}(\Delta N_{t_{i+1}})P + \widehat{\sigma}^2)^{-1}\mathrm{Cov}(Y_{t_{i+1}}, N_{t_{i+1}}). \tag{4.13}$$

Calculation of (4.11), (4.12) and (4.13) constitutes a single step of the forward filter; see Algorithm 13, which can be iterated over t to give an evaluation of the observed data likelihood (under the LNA), $\pi(y|\theta)$. Hence, draws from the marginal parameter posterior $\pi(\theta|y)$, with the LNA as the inferential model, are obtained in a straightforward manner via Metropolis-Hastings e.g. random walk Metropolis (RWM).

We can use the LNA to generate draws from $\pi(n|\theta, y)$. Note the factorisation

$$\pi(n|\theta, y) = \prod_{i=1}^{L-1} \pi(n_{t_i}|n_{t_{i+1}}, y_{1:t_i}, \theta)$$

where each constituent term is a Gaussian density. Under the LNA, the joint density of N_{t_i} and $N_{t_{i+1}}$ conditional on $Y_{1:t_i} = y_{1:t_i}$ is

$$\begin{pmatrix} N_{t_i} \\ N_{t_{i+1}} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} a_{t_i} \\ \eta_{t_{i+1}} \end{pmatrix}, \begin{pmatrix} C_{t_i} & C_{t_i} G'_{t_{i+1}} \\ G_{t_{i+1}} C_{t_i} & V_{t_{i+1}} \end{pmatrix} \right\}.$$

Conditioning on $N_{t_{i+1}} = n_{t_{i+1}}$ gives $N_{t_i} | \left(N_{t_{i+1}} = n_{t_{i+1}}, Y_{1:t_i} = y_{1:t_i}, \theta\right) \sim \mathrm{N}(\tilde{a}_t, \tilde{C}_t)$ with mean and variance

$$\begin{aligned} \tilde{a}_{t_i} &= a_{t_i} + C_{t_i} G'_{t_{i+1}} V_{t_{i+1}}^{-1} \left(n_{t_{i+1}} - \eta_{t_{i+1}}\right), \\ \tilde{C}_{t_i} &= C_{t_i} - C_{t_i} G'_{t_{i+1}} V_{t_{i+1}}^{-1} G_{t_{i+1}} C_{t_i}. \end{aligned}$$

Hence, the components of the cumulative incidence n can be drawn via backward sampling for $t = t_L, t_{L-1}, \ldots, t_0$, given storage of the LNA ODE output and filtering mean/variance from the forward filter (see West and Harrison, 2006, for more de-

Algorithm 13 Step t_{i+1} of the LNA Forward Filter

Input: Parameter θ; a_{t_i} and C_{t_i}, the initial conditions of (3.15) and (3.17); $\pi(y_{1:t_i}|\theta)$, the current observed data likelihood; $y_{t_{i+1}}$, the next observation.

1. Prior at t_{i+1}. Initialise the LNA with $\eta_{t_i} = a_{t_i}$, $G_{t_i} = 1_2$ and $V_{t_i} = C_{t_i}$. Integrate (3.15), (3.16) and (3.17) forward to t_{i+1} to obtain $\eta_{t_{i+1}}$, $G_{t_{i+1}}$ and $V_{t_{i+1}}$. Thus
$$N_{t_{i+1}} | \left(Y_{1:t_i} = y_{1:t_i}\right) \sim \mathrm{N}(\eta_{t_{i+1}}, V_{t_{i+1}});$$
2. Likelihood update. Compute
$$\pi(y_{1:t_{i+1}}|\theta) = \pi(y_{1:t_i}|\theta)\pi(y_{t_{i+1}}|y_{1:t_i}, \theta)$$
where $\pi(y_{t_{i+1}}|y_{1:t_i}, \theta)$ is given by Equation (4.11);
3. Posterior at t_{i+1}. Combining the distributions of $N_{t_{i+1}}$ and $Y_{t_{i+1}}$ (given $y_{1:t_i}$) and then conditioning on $y_{t_{i+1}}$ gives $N_{t_{i+1}} | \left(Y_{1:t_{i+1}} = y_{1:t_{i+1}}\right) \sim \mathrm{N}(a_{t_{i+1}}, C_{t_{i+1}})$ where $a_{t_{i+1}}$ and $C_{t_{i+1}}$ are given by (4.12) and (4.13);

Output: $\pi(y_{1:t_{i+1}}|\theta)$, $a_{t_{i+1}}$ and $C_{t_{i+1}}$.

tails). Then, the latent process x can be constructed deterministically from n and the initial values x_0 using Equation (3.1).

4.3.2 Pseudo-marginal methods via Poisson leap

We will now consider the task of performing Bayesian inference under the Poisson leap. Unfortunately, when the Poisson leap is used as an inferential model, the observed data likelihood term in Equation (4.6) does not have an analytically tractable form. Progress can be made by considering pseudo-marginal methods (see Sections 2.6 and 2.7), where we instead make use of an unbiased estimate of $\pi(y|\theta)$ to target the joint density $\pi(\theta, u)$, which can be marginalised to obtain samples from the target.

To this end, consider the intractable observed data likelihood $\pi(y|\theta)$ which can be factorised as

$$\pi(y|\theta) = \pi(y_{t_1}|\theta) \prod_{i=2}^{L} \pi(y_{t_i}|y_{1:t_{i-1}}, \theta) \tag{4.14}$$

where $y_{1:t_{i-1}} = (y_{t_1}, \ldots, y_{t_{i-1}})'$. The terms in Equation (4.14) can be recursively estimated using a particle filter, such as the bootstrap particle filter of Section 2.8.3, in such a way that realisations of a non-negative unbiased estimator of the full likelihood are obtained. We denote this estimator by

$$\hat{\pi}_U(y|\theta) = \hat{\pi}_{U_1}(y_{t_1}|\theta) \prod_{i=2}^{L} \hat{\pi}_{U_{t_i}}(y_{t_i}|y_{1:t_{i-1}}, \theta)$$

where the flattened vector $U = (U_1', \ldots, U_{t_L}')' \sim g(u)$ denotes all random variables used in the construction of the estimator. Hence, unbiasedness here means that $E_{U\sim g}\{\hat{\pi}_U(y|\theta)\} = \pi(y|\theta)$. Algorithm 14 gives step t_{i+1} of the particle filter and can be executed for $t_0, \ldots, t_L$ upon initialising with particles $\{x_0^{(k)}, k = 1, \ldots, N\}$. Note

that, if the initial state is assumed fixed and known, each $x_0^{(k)}$ will be the same for $k = 1, \ldots, N$, and therefore the initial state x_0 should only be entered once. Upon iterating Algorithm 14 over t, the product (over observation times) of the average unnormalised weight gives an unbiased estimator of $\pi(y|\theta)$ (Del Moral, 2004) and is key to the construction of a pseudo-marginal scheme that we now describe.

Recall from Section 2.6 that pseudo-marginal Metropolis-Hastings methods are a class of Metropolis-Hastings (MH) scheme that target the joint density

$$\pi(u, \theta) \propto \pi(\theta) g(u) \hat{\pi}_u(y|\theta)$$

for which it is easily checked that marginalising over U gives the marginal parameter posterior $\pi(\theta|y)$. Hence, an MH scheme with proposal density $q(\theta^*|\theta)g(u^*)$ and

Algorithm 14 Step t_{i+1} of the Particle Filter

Input: Parameter vector θ, next observation $y_{t_{i+1}}$, N particles $\{n_{t_i}^{(k)}, x_0^{(k)}, k = 1, \ldots, N\}$.

1. Forward propagation. For $k = 1, \ldots, N$, run the Poisson leap algorithm (see Algorithm 9) over $(t_i, t_{i+1}]$ to obtain $\Delta n_{t_{i+1}}^{(k)}$;
2. Compute the weights. For $k = 1, \ldots, N$:

$$w_{t_{i+1}}^{(k)} = \pi\left(y_{t_{i+1}} | \Delta n_{t_{i+1}}^{(k)}, \phi\right), \qquad \tilde{w}_{t_{i+1}}^{(k)} = \frac{\tilde{w}_{t_{i+1}}^{(k)}}{\sum_{j=1}^{N} \tilde{w}_{t_{i+1}}^{(j)}};$$

3. Resample N particles with replacement using the weights $w_{t_{i+1}}^{(k)}$, $k = 1, \ldots, N$ as probabilities.

Output: N particles $\{n_{t_{i+1}}^{(k)}, k = 1, \ldots, N\}$ to be used in step t_{i+1}, an estimate for the current marginal likelihood term $\hat{\pi}_{u_{t_{i+1}}}(y_{t_{i+1}}|y_{1:t_i}, \theta) = \frac{1}{N}\sum_{k=1}^{N} w_{t_{i+1}}^{(k)}$.

acceptance probability

$$\alpha\left(\{\theta^*, u^*\}|\{\theta, u\}\right) = \min\left\{1, \frac{\pi(\theta^*)\hat{\pi}_{u^*}(y|\theta^*)}{\pi(\theta)\hat{\pi}_u(y|\theta)} \times \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right\}$$

targets the joint density $\pi(u, \theta)$ for which retaining draws of θ gives (dependent) samples from the marginal parameter posterior.

As noted in Section 2.7, the efficiency of the PMMH scheme can be improved by proposing the auxiliary variable from a g-reversible kernel $K(u^*|u)$ that induces positive correlation between u and u^*, and in turn, $\hat{\pi}_u(y|\theta)$ and $\hat{\pi}_{u^*}(y|\theta^*)$, so that the variance of the acceptance probability is reduced. Suppose that $g(u) = \mathrm{N}\left(u;\, 0\,,\, I_{\dim(u)}\right)$ and note that where necessary, the inverse CDF method can be used to transform the auxiliary variable to be Gaussian where a uniform draw is required. A practical choice of $K(u^*|u)$ is the g-reversible Crank-Nicolson kernel

$$K(u^*|u) = \mathrm{N}\left(u^*;\, \zeta u\,,\, \left(1 - \zeta^2\right) I_{\dim(u)}\right)$$

for which the tuning parameter ζ controls correlation between u and u^*. The resulting correlated PMMH scheme (CPMMH, Dahlin et al., 2015; Deligiannidis et al., 2018) is a MH scheme targeting $\pi(u, \theta)$ with proposal density $q(\theta^*|\theta)K(u^*|u)$ and acceptance probability as above. CPMMH can result in significant gains in computational efficiency over PMMH (see e.g. Golightly et al., 2019, in the context of stochastic kinetic models), provided that the positive correlation between u and u^* induces positive correlation between successive likelihood estimates. To alleviate the issue of resampling in the particle filter potentially eroding this correlation, we follow Choppala et al. (2016) by sorting particles (according to Euclidean distance from the particle with the smallest first component) before propagation. The general procedure at an arbitrary time t is as follows. Suppose at time t, the set of particles

is given by $\{x_t^i\}_{i=1}^N$, where each x_t^i is the n-dimensional vector $x_t^i = (x_{1,t}^i, \ldots, x_{n,t}^i)'$. First, find the particle with the smallest first component, i.e. find the particle x_t^j for which $x_{1,t}^j = \min_i \left(\{x_{1,t}^i\}_{i=1}^N \right)$, and put the particle x_t^j first in the new ordered set. Then, for each remaining particle, calculate the Euclidean distance between each x_t^i and the previously found x_t^j. Sorting these remaining particles by their corresponding Euclidean distances (smallest to largest), gives the order of particles in the new sorted set.

Finally, we note that when interest lies in the posterior for the latent incidence process, samples can be obtained via modification of the (C)PMMH scheme (Andrieu et al., 2010) by drawing a particle path at each algorithm iteration. Note that this requires storing the ancestral lineages of the particles in each run of the particle filter.

Tuning pseudo-marginal Metropolis-Hastings

As discussed in Section 2.6, the estimator's variance can be seen as a tuning parameter for the scheme. Golightly and Wilkinson (2015) note that, when considering particle marginal schemes, the chain mixing is dependent on N, the number of particles used in the particle filter when estimating the observed data likelihood. Furthermore, by increasing the number of particles, we can reduce the estimator's variance.

Thus, one option is to make N arbitrarily large, which would ensure that the variance is reduced (compared to using just one particle) at the cost of increasing the computational power required to generate the estimate. Clearly, if N is chosen to be too large, then the computational cost of generating the estimate will become infeasible for relatively little improvement in mixing, and so it is not recommended

to take this approach. Instead, Sherlock et al. (2015) suggests an optimal value of N leads to the log-estimate variance being around 2. Therefore, after obtaining (approximate) posterior mean values for the parameters via a short pilot run of the scheme, N can be found through trial and error by fixing the parameter values at the posterior means and running the scheme, storing the log-estimate for each iteration.

Furthermore, once an optimal value of N is found, additional tuning of the scheme can be made to optimise the acceptance rate of proposed values. Recall from Section 2.5.2 that, for random walk Metropolis schemes, the innovation variance V can be seen as a tuning parameter; large values lead to a chain which makes infrequent but significant jumps around the parameter space whereas small values result in a chain which makes small but frequent moves around the parameter space, leading to poor posterior coverage. It is recommended (see Sherlock et al., 2015) to take

$$V = \frac{2.56^2}{d}\text{Var}(\theta),$$

as a rule of thumb. Optimal acceptance rates can be found in a table in Schmon et al. (2021), to which we refer the reader for more information. As discussed in Section 2.5.2, we typically substitute the unknown posterior variance $\text{Var}(\theta)$ with an estimate $\widehat{\text{Var}}(\theta)$ obtained from a short pilot run of the scheme.

4.4 Discussion

This chapter considered the inference task for epidemic models: given partial and noisy observations, learn about the parameters and dynamic processes in the Bayesian paradigm. From the perspective of a batch analysis, this requires sampling the joint posterior over all unknown quantities, a problem made difficult by the intractability

of the observed data likelihood.

If the observation model can be approximated as Gaussian, with a mean that is linear in the latent state, then an LNA model of the latent transmission process allows direct approximation of the observed data likelihood.

Additional inferential accuracy is possible at increased computational cost via pseudo-marginal methods, which exactly target the parameter posterior for a given inferential model. These methods are compared and contrasted regarding overall efficiency in Chapter 6.

Chapter 5

Bayesian inference for SKMs via sequential methods

In this chapter, we consider the problem of performing full sequential Bayesian inference for the parameters (and unobserved dynamic processes) governing the dSEIR model based on incidence observations that may be incomplete and subject to measurement error. We briefly recall the observation model before considering the inference task. We perform sequential Bayesian inference via a particle filter, a key ingredient of which is a novel construct that allows approximate draws of the conditioned dSEIR process between observation instants.

5.1 Observation model

Recall from Section 4.1 that the data $y = (y_{t_1}, \ldots, y_{t_L})'$ are partial observations on the cumulative incidence process $\Delta N_{t_i} = N_{t_i} - N_{t_{i-1}}$ at regular times t_i, for $i = 1, ..., L$. Further, recall that these observations are assumed noisy, with noise

typically characterised by one of the following three observation models

$$Y_{t_i}|\Delta N_{t_i} \sim \mathrm{N}\left(P'\Delta N_{t_i}, \sigma^2\right), \tag{5.1}$$

$$Y_{t_i}|\Delta N_{t_i} \sim \mathrm{Bin}\left(P'\Delta N_{t_i}, \rho\right), \tag{5.2}$$

$$Y_{t_i}|\Delta N_{t_i} \sim \mathrm{NegBin}\left(\mu_i = \rho P'\Delta N_{t_i}, \sigma_i^2 = \mu_i + \mu_i^2/\nu\right) \tag{5.3}$$

for $i = 1, \ldots, L$. We denote by ϕ the parameters governing the observation model and explicitly link the data to the latent incidence process via the mass function $\pi(y_{t_i}|\Delta n_{t_i}, \phi)$.

5.2 Inference task

As in Section 4.2, we assume that interest lies in the vector of all static parameters $\psi = (\theta', \lambda', \phi')'$, the latent incidence process $\{N_t, t_0 \leq t \leq t_L\}$, and additionally the contact rate process $\{\tilde{\beta}_t, t_0 \leq t \leq t_L\}$ and the initial state vector x_{t_0}.

Recall that we opt to partition each inter-observation time interval as

$$t_{i-1} = \tau_{i,0} < \tau_{i,1} < \cdots < \tau_{i,m} = t_i$$

with $\tau_{i,j} - \tau_{i,j-1} = \Delta\tau = \Delta t/m$, so that the reaction hazards can be assumed constant over each sub-interval $(\tau_{i,j-1}, \tau_{i,j}]$.

We let $\tilde{\beta} = (\tilde{\beta}_{\tau_{1,0}}, \tilde{\beta}_{\tau_{1,1}}, \ldots, \tilde{\beta}_{\tau_{L,m}})'$ and $\Delta n = (\Delta n_{\tau_{1,1}}, \Delta n_{\tau_{1,2}}, \ldots, \Delta n_{\tau_{L,m}})'$ denote the latent contact rate process and collection of incidences over sub-intervals $(\tau_{i,j-1}, \tau_{i,j}]$, for $i = 1, \ldots, L$ and $j = 1, \ldots, m$, again noting that $\Delta n_{t_i} = \sum_{j=1}^{m} \Delta n_{\tau_{i,j}}$ gives the cumulative incidence over $(t_{i-1}, t_i]$. Now, upon ascribing a prior density $\pi(\psi)$ to ψ,

the inference tasks proceeds via the joint posterior

$$\pi(\psi, \tilde{\beta}, \Delta n|y) \propto \pi(\psi)\pi(\tilde{\beta}|\lambda)\pi(\Delta n|\tilde{\beta}, \theta)\pi(y|\Delta n, \phi). \tag{5.4}$$

Here, the joint density of the latent contact rate process is

$$\begin{aligned}\pi(\tilde{\beta}|\lambda) &= \pi(\tilde{\beta}_{\tau_{1,0}}) \prod_{i=1}^{L}\prod_{j=1}^{m} \pi(\tilde{\beta}_{\tau_{i,j}}|\tilde{\beta}_{\tau_{i,j-1}}, \lambda) \\ &= \pi(\tilde{\beta}_{\tau_{1,0}}) \prod_{i=1}^{L}\prod_{j=1}^{m} \mathrm{N}(\tilde{\beta}_{\tau_{i,j}}; a(\tilde{\beta}_{\tau_{i,j-1}}, \lambda)\Delta\tau\ ,\ b^2(\tilde{\beta}_{\tau_{i,j-1}}, \lambda)\Delta\tau),\end{aligned}$$

where $\mathrm{N}(\cdot; m, v^2)$ denotes the density of a normal random variable with mean m and variance v^2. The joint probability of the incidence process is

$$\pi(\Delta n|\tilde{\beta}, \theta) = \prod_{i=1}^{L}\prod_{j=1}^{m}\prod_{k=1}^{3} \mathrm{Po}(\Delta n_{k,\tau_{i,j}}; h_k(x_{\tau_{i,j-1}}, \tilde{\beta}_{\tau_{i,j-1}})\Delta\tau)$$

where $\mathrm{Po}(\cdot; h)$ denotes the probability mass function of a Poisson random variable with mean h. Finally,

$$\pi(y|\Delta n, \phi) = \prod_{i=1}^{L} \pi(y_{t_i}|\Delta n_{t_i}, \phi)$$

with $\pi(y_{t_i}|\Delta n_{t_i}, \phi)$ as the probability mass function arising from either (5.1), (5.2) or (5.3).

Since the joint posterior in (5.4) will be intractable, we resort to Monte Carlo methods for generating samples of the parameters and latent dynamic processes. In particular, we wish to perform inference sequentially and develop a particle filter approach in the next section.

5.3 Particle filter approach

It will be helpful here to introduce the shorthand notation $y_{[1,i]} = (y_{t_1}, \ldots, y_{t_i})'$ to denote the observations up to (and including) time t_i. Similarly,

$$\tilde{\beta}_{[0,i]} = (\tilde{\beta}_{\tau_{1,0}}, \tilde{\beta}_{\tau_{1,1}}, \ldots, \tilde{\beta}_{\tau_{i,m}})'$$

and

$$\Delta n_{(0,i]} = (\Delta n_{\tau_{1,1}}, \Delta n_{\tau_{1,2}}, \ldots, \Delta n_{\tau_{i,m}})'$$

denote respectively, the latent contact rate process and collection of incidence increments over the corresponding time horizon.

Now, by applying Bayes theorem sequentially, we obtain the posterior distribution at time t_{i+1} as

$$\begin{aligned}\pi(\psi, \tilde{\beta}_{[0,i+1]}, \Delta n_{(0,i+1]} | y_{[1,i+1]}) \propto \pi(\psi, \tilde{\beta}_{[0,i]}, \Delta n_{(0,i]} | y_{[1,i]}) \pi(\tilde{\beta}_{(i,i+1]} | \tilde{\beta}_{t_i}, \lambda) \\ \times \pi(\Delta n_{(i,i+1]} | \tilde{\beta}_{[i,i+1)}, \theta) \pi(y_{t_{i+1}} | \Delta n_{t_{i+1}}, \phi) \quad (5.5)\end{aligned}$$

and note that, $\tilde{\beta}_{(i,i+1]} = (\tilde{\beta}_{\tau_{i+1,1}}, \ldots, \tilde{\beta}_{\tau_{i+1,m}})'$ whereas $\tilde{\beta}_{[i,i+1)} = (\tilde{\beta}_{\tau_{i+1,0}}, \ldots, \tilde{\beta}_{\tau_{i+1,m-1}})'$. Now, given an equally weighted sample of 'particles' $\left(\psi^{(1:N)}, \tilde{\beta}^{(1:N)}_{[0,i]}, \Delta n^{(1:N)}_{(0,i]}\right)$ from $\pi(\psi, \tilde{\beta}_{[0,i]}, \Delta n_{(0,i]} | y_{[1,i]})$, the form of (5.5) immediately suggests an importance resampling step that extends the dynamic process particles via $\tilde{\beta}^{(k)}_{(i,i+1]} \sim \pi(\cdot | \tilde{\beta}^{(k)}_{t_i}, \lambda^{(k)})$, $\Delta n^{(k)}_{(i,i+1]} \sim \pi(\cdot | \tilde{\beta}^{(k)}_{[i,i+1)}, \theta^{(k)})$, weights the resulting particles $\left(\psi^{(k)}, \tilde{\beta}^{(k)}_{[0,i+1]}, \Delta n^{(k)}_{(0,i+1]}\right)$ by $w^{(k)}_{i+1} \propto \pi(y_{t_{i+1}} | \Delta n^{(k)}_{t_{i+1}}, \phi^{(k)})$, before sampling with replacement among the particles (using the weights as probabilities). Applying this sequence of steps recursively in time, and outputting the samples at each observation time (summarising the sequence of filtering distributions $\pi(\psi, \tilde{\beta}_{t_i}, \Delta n_{t_i} | y_{[1,i]})$, $i = 1, \ldots, L$), gives the

bootstrap particle filter (see e.g. Gordon et al., 1993). Two problems are apparent here. Firstly, repeated resampling of static parameter particles will lead to sample impoverishment (with marginal parameter posteriors collapsing to point masses). Secondly, drawing the incidence process 'blindly' (that is, via forward simulation) is likely to lead to many particle trajectories with negligible weight.

To alleviate the problem of sample impoverishment, we follow the approach of Storvik (2002) (and see also Fearnhead (2002)) by noting that the conditional posterior of a subset of parameter components is tractable for a particular choice of prior. To this end, consider the factorisation

$$\begin{aligned}\pi(\psi, \tilde{\beta}_{[0,i]}, \Delta n_{(0,i]}|y_{[1,i]}) &= \pi(\theta, \lambda|\phi, \tilde{\beta}_{[0,i]}, \Delta n_{(0,i]}, y_{[1,i]})\pi(\phi, \tilde{\beta}_{[0,i]}, \Delta n_{(0,i]}|y_{[1,i]}) \\ &= \pi(\theta, \lambda|\Delta n_{(0,i]}, \tilde{\beta}_{[0,i]})\pi(\phi, \tilde{\beta}_{[0,i]}, \Delta n_{(0,i]}|y_{[1,i]}) \\ &= \pi(\theta|\Delta n_{(0,i]})\pi(\lambda|\tilde{\beta}_{[0,i]})\pi(\phi, \tilde{\beta}_{[0,i]}, \Delta n_{(0,i]}|y_{[1,i]})\end{aligned}$$

where the last two lines follow from the conditional independencies present in the model. Assuming an independent prior specification for the components of $\theta = (\kappa, \gamma)'$, with $\kappa \sim \text{Gamma}(a_\kappa, b_\kappa)$ and $\gamma \sim \text{Gamma}(a_\gamma, b_\gamma)$ gives

$$\kappa|\Delta n_{(0,i]} \sim \text{Gamma}\left(a_\kappa + \sum_{l=1}^{i}\sum_{j=1}^{m} \Delta n_{2,\tau_{l,j}}, b_\kappa + \sum_{l=1}^{i}\sum_{j=1}^{m} g_2(x_{\tau_{l,j-1}})\Delta\tau\right), \tag{5.6}$$

$$\gamma|\Delta n_{(0,i]} \sim \text{Gamma}\left(a_\gamma + \sum_{l=1}^{i}\sum_{j=1}^{m} \Delta n_{3,\tau_{l,j}}, b_\gamma + \sum_{l=1}^{i}\sum_{j=1}^{m} g_3(x_{\tau_{l,j-1}})\Delta\tau\right), \tag{5.7}$$

where $g_2(x) = h_2(x)/\kappa$ and $g_3(x) = h_3(x)/\gamma$, that is, the combinatorial factors in the hazard functions governing the infection and removal events. Similarly, if the discretised SDE (3.19) is adopted for the contact rate process $\tilde{\beta}_t$, then we may

ascribe the prior $\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$ leading to the conditional posterior

$$\lambda|\tilde{\beta}_{[0,i]} \sim \text{Gamma}\left(a_\lambda + \frac{im}{2}, b_\lambda + \frac{1}{2}\sum_{l=1}^{i}\sum_{j=1}^{m}(\tilde{\beta}_{\tau_{l,j-1}} - \tilde{\beta}_{\tau_{l,j}})^2/\Delta\tau\right). \tag{5.8}$$

Hence, the conditional posterior $\pi(\theta, \lambda|\Delta n_{(0,i]}, \tilde{\beta}_{[0,i]})$ can be summarised by a vector of sufficient statistics $T_i := T_i(\Delta n_{(0,i]}, \tilde{\beta}_{[0,i]})$ given by the three shape and three rate hyperparameters in (5.6)–(5.8). This vector can then be updated recursively within the particle filter upon initialising with $T_0 = (a_\kappa, a_\gamma, a_\lambda, b_\kappa, b_\gamma, b_\lambda)$. In turn, particles $\theta^{(k)}$ and $\lambda^{(k)}$ can be updated by drawing from (5.6)–(5.8) conditional on $T_i^{(k)} := T_i(\Delta n_{(0,i]}^{(k)}, \tilde{\beta}_{[0,i]}^{(k)})$.

Our treatment of ϕ depends on the observation model used. In the case of the Binomial model in (5.2), with $\phi = \rho$, the conditional posterior $\pi(\phi|\Delta n_{(0,i]}, y_{[1,i]})$ is tractable upon ascribing a $\text{Beta}(a_\phi, b_\phi)$ prior to ϕ. We obtain

$$\phi|\Delta n_{(0,i]}, y_{[1,i]} \sim \text{Beta}\left(a_\phi + \sum_{l=1}^{i} y_{t_l}, b_\phi + P'\sum_{l=1}^{i}\Delta N_{t_l} - \sum_{l=1}^{i} y_{t_l}\right).$$

Unfortunately, in the case of the Negative Binomial model (5.3) with $\phi = (\rho, \nu)'$, $\pi(\phi|\Delta n_{(0,i]}, y_{[1,i]})$ remains intractable. In this case, we treat ϕ as a dynamic parameter and induce artificial evolution of ϕ by following the approach of Liu and West (2001). That is, if $\phi_i^{(k)}$ denotes the kth particle at time t_i, we propagate forward to time t_{i+1} via

$$\phi_{i+1}^{(k)} \sim N\left(m_i^{(k)}, s^2 V_i\right)$$

where $m_i^{(k)} = a\phi_i^{(k)} + (1-a)\bar{\phi}_i$, $V_i = \sum_{k=1}^{N}(\phi_i^{(k)} - \bar{\phi}_i)(\phi_i^{(k)} - \bar{\phi}_i)'/N$ and $\bar{\phi}_i = \sum_{k=1}^{N}\phi_i^{(k)}/N$. The practitioner can choose the hyperparameters a and s to control shrinkage and over-dispersion. See Section 2.8.4 for more details.

We propagate the dynamic $\tilde{\beta}$ process blindly by drawing $\tilde{\beta}^{(k)}_{(i,i+1]} \sim \pi(\cdot|\tilde{\beta}^{(k)}_{t_i}, \lambda^{(k)})$. To mitigate against jump process trajectories being inconsistent with the next observation (the second problem described above), we propagate $\Delta n^{(k)}_{(i,i+1]}$ conditional on $y_{t_{i+1}}$, by drawing

$$\Delta n^{(k)}_{(i,i+1]} \sim q(\cdot|\tilde{\beta}^{(k)}_{[i,i+1)}, \phi^{(k)}_{i+1}, \theta^{(k)}, y_{t_{i+1}})$$

for some suitable construct $q(\cdot|\tilde{\beta}^{(k)}_{[i,i+1)}, \phi^{(k)}_{i+1}, \theta^{(k)}, y_{t_{i+1}})$, which we discuss in Section 5.4. The complete particle filtering scheme (as appropriate for observation model (5.3)) is summarised in Algorithm 15.

Algorithm 15 Particle Filter

Input: data $y = (y_{t_1}, \ldots, y_{t_L})'$, number of particles N, initial draws $\psi^{(1:N)} \sim \pi(\psi)$, $x^{(1:N)}_{t_0} \sim \pi(x_{t_0})$, $\tilde{\beta}^{(1:N)}_{t_0} \sim \pi(\tilde{\beta}_{t_0})$ and sufficient statistic $T_0^{1:N}$.
For $i = 0, \ldots, L-1$ and $k = 1, \ldots, N$:

1. **Propagate** dynamic processes:
   a. draw $\phi^{(k)}_{i+1} \sim N(m^{(k)}_i, s^2 V_i)$;
   b. draw $\tilde{\beta}^{(k)}_{(i,i+1]} \sim \pi(\cdot|\tilde{\beta}^{(k)}_{t_i}, \lambda^{(k)})$;
   c. draw $\Delta n^{(k)}_{(i,i+1]} \sim q(\cdot|\tilde{\beta}^{(k)}_{[i,i+1)}, \phi^{(k)}_{i+1}, \theta^{(k)}, y_{t_{i+1}})$;
2. **Resample** with replacement among $\left(\psi^{(1:N)}, \tilde{\beta}^{(1:N)}_{t_{i+1}}, \Delta n^{(1:N)}_{t_{i+1}}\right)$ using the weights

$$w^{(k)}_{i+1} \propto \frac{\pi(\Delta n^{(k)}_{(i,i+1]}|\tilde{\beta}^{(k)}_{[i,i+1)}, \theta^{(k)})\pi(y_{t_{i+1}}|\Delta n^{(k)}_{t_{i+1}}, \phi^{(k)})}{q(\Delta n^{(k)}_{(i,i+1]}|\tilde{\beta}^{(k)}_{[i,i+1)}, \phi^{(k)}_{i+1}, \theta^{(k)}, y_{t_{i+1}})}$$

   as probabilities;
3. **Update** sufficient statistic $T^{(k)}_{i+1} := T_{i+1}(T^{(k)}_i, \Delta n^{(k)}_{(i,i+1]}, \tilde{\beta}^{(k)}_{(i,i+1]})$;
4. **Sample** $\theta^{(k)} \sim \pi(\theta|T^{(k)}_{i+1})$ and $\lambda^{(k)} \sim \pi(\lambda|T^{(k)}_{i+1})$ using (5.6)–(5.8);

Output: particle representation $\left(\psi^{(1:N)}_i, \tilde{\beta}^{(1:N)}_{t_i}, \Delta n^{(1:N)}_{t_i}\right)$ of the filtering distributions $\pi(\psi, \tilde{\beta}_{t_i}, \Delta n_{t_i}|y_{[1,i]})$, $i = 1, \ldots, L$.

5.4 A novel bridge construct

As previously discussed, we wish to propagate particle trajectories conditional on the next observation. To this end, we derive an approximate instantaneous rate or hazard function conditioned on the next observation. The derivation involves the construction of a Gaussian approximation to the joint distribution of the incidence over a time window whose right end-point is the next observation time and the next observation itself. The conditioned hazard is then taken to be proportional to the expectation of the incidence given the observation. The effect of this is to essentially steer proposed values towards the observed values, reducing the variance of the importance weight. Since we are conditioning on the observed values, we also obtain proposed trajectories that are consistent with the data.

Suppose we receive an observation y_{t_i} and have simulated as far as $\tau_{i,j} \in (t_{i-1}, t_i]$ so that $x_{\tau_{i,j}}$ and $\Delta n_{(t_{i-1},\tau_{i,j}]}$ are fixed and known. By analogy with the unconditioned hazard function, we denote the conditioned hazard function by $h^*\left(x_{\tau_{i,j}}|y_{t_i}\right)$, with dependence on the parameters θ and incidence process suppressed for notational simplicity.

Assuming a Normal approximation to the Poisson distribution, the number of reaction events in the interval $(\tau_{i,j}, t_i]$, denoted by $\Delta N_{(\tau_{i,j},t_i]}$, is

$$\Delta N_{(\tau_{i,j},t_i]} \overset{\text{approx.}}{\sim} \mathrm{N}\left\{h(x_{\tau_{i,j}})(t_i - \tau_{i,j}), H(x_{\tau_{i,j}})(t_i - \tau_{i,j})\right\} \tag{5.9}$$

where $H(x_{\tau_{i,j}}) = \mathrm{diag}\left\{h(x_{\tau_{i,j}})\right\}$. Moreover, if we take a Normal approximation to the observation model, we obtain

$$Y_{t_i}|\Delta N_{(\tau_{i,j},t_i]} \overset{\text{approx.}}{\sim} \mathrm{N}\left\{\mu(\Delta n_{(t_{i-1},\tau_{i,j}]}), \sigma^2(\Delta n_{(t_{i-1},\tau_{i,j}]})\right\}$$

where $\mu(\Delta n_{(t_{i-1},\tau_{i,j}]}) = \rho P'(\Delta n_{(t_{i-1},\tau_{i,j}]} + \Delta N_{(\tau_{i,j},t_i]})$ for both Binomial and Negative Binomial observation models, with dependence on $\Delta N_{(\tau_{i,j},t_i]}$ suppressed for notational ease. The variance term is model dependent, taking the form

$$\sigma^2(\Delta n_{(t_{i-1},\tau_{i,j}]}) = \rho(1-\rho)P'(\Delta n_{(t_{i-1},\tau_{i,j}]} + \Delta \hat{N}_{(\tau_{i,j},t_i]})$$

in the case of a Binomial observation model and

$$\sigma^2(\Delta n_{(t_{i-1},\tau_{i,j}]}) = \hat{\mu}(\Delta n_{(t_{i-1},\tau_{i,j}]}) + \hat{\mu}(\Delta n_{(t_{i-1},\tau_{i,j}]})^2/\nu$$

in the case of a Negative Binomial observation model. Note that $\hat{\mu}$ has $\Delta N_{(\tau_{i,j},t_i]}$ replaced by an estimate $\Delta \hat{N}_{(\tau_{i,j},t_i]}$, leading to a linear Gaussian structure, which is essential for the tractability of the conditioned hazard function. We take $\Delta \hat{N}_{(\tau_{i,j},t_i]}$ to be the mean of the approximating Normal distribution in (5.9).

Making use of the linear Gaussian structure, we now form the approximate joint distribution of $\Delta N_{(\tau_{i,j},t_i]}$ and Y_{t_i}. This is

$$\begin{pmatrix} \Delta N_{(\tau_{i,j},t_i]} \\ Y_{t_i} \end{pmatrix} \sim N\left\{ \begin{pmatrix} h(x_{\tau_{i,j}})(t_i - \tau_{i,j}) \\ \hat{\mu}(\Delta n_{(t_{i-1},\tau_{i,j}}) \end{pmatrix}, \right.$$
$$\left. \begin{pmatrix} H(x_{\tau_{i,j}})(t_i - \tau_{i,j}) & \rho H(x_{\tau_{i,j}})P(t_i - \tau_{i,j}) \\ \rho P' H(x_{\tau_{i,j}})(t_i - \tau_{i,j}) & \rho^2 P' H(x_{\tau_{i,j}})P(t_i - \tau_{i,j}) + \sigma^2(\Delta n_{(t_{i-1},\tau_{i,j}]}) \end{pmatrix} \right\}.$$

Now, we condition on $Y_{t_i} = y_{t_i}$, take the expectation of the resulting distribution and divide by $(t_i - \tau_{i,j})$ to obtain the conditioned hazard function as $h^*(x_{\tau_{i,j}}, \theta | y_{t_i})$

$$h^*(x_{\tau_{i,j}}|y_{t_i}) = h(x_{\tau_{i,j}}) + \rho P' H(x_{\tau_{i,j}})(\rho^2 P' H(x_{\tau_{i,j}})P(t_i - \tau_{i,j}) + \sigma^2)^{-1}$$
$$\times (y_{t_i} - \hat{\mu}(\Delta n_{(t_{i-1},\tau_{i,j}]})).$$

It is then straightforward to generate end-point conditioned trajectories, for example, over an interval $(t_{i-1}, t_i]$ by executing Algorithm 9, with h replaced by h^* and time step $\Delta\tau$. Moreover, the corresponding likelihood $q(\cdot|\tilde{\beta}_{[i-1,i)}, \phi_i, \theta, y_{t_i})$ is simply a product of Poisson probability mass functions, each with rate $h^*(x_{\tau_{i,j}}|y_{t_i})\Delta\tau$, for $j = 0, \ldots, m-1$.

5.5 Simulated data example

To demonstrate the effectiveness of this bridge construct, we will consider its implementation over a single observation interval using synthetic data from the SIR model (see Section 3.2).

The observation, denoted by y for this example, is a noisy observation of the cumulative incidence at time $t = 1$, where we assume that observations for this toy epidemic would be made on unit intervals. That is, we simulate a single realisation from the true, unobserved, cumulative incidence process $\{N_t, t \geq 0\}$ using the Poisson leap and then subject this to error according to a Binomial observation regime with reporting rate ρ chosen arbitrarily to be 0.9. The initial state x_0 and parameter vector θ are chosen to be consistent with those from Section 3.3.1. Specifically, $x_0 = (762, 5)'$ and $\theta = (\exp(-6), 0.5)'$.

Presented in Figure 5.1 are summaries (mean and 95% credible interval) from 10^4 realisations of the cumulative incidence of infections over the interval $(0, 1]$. The top left panel shows summaries from the vanilla Poisson leap, produced using Algorithm 9, the top right panel shows summaries from the Poisson leap with the bridge, and the bottom panel shows summaries from the conditioned process.

By comparing the proposed trajectories to the conditioned process, we can immediately see that the bridge provides us with trajectories consistent with the observation and, hence, are much more desirable. Also visualised is the overwhelming difference in concentration of endpoints about the observation. Without implementing the bridge, we allow the trajectories to fan out across the interval, leading to many trajectories ending a significant distance away from the true value of the process. When the bridge is used, we can see that all trajectories are pushed towards the observation, giving a much more concentrated distribution of endpoints about the true value of the process. In practice, any trajectory simulated under the bridge construct is much more likely to yield a significant importance weight than one simulated via the vanilla Poisson leap.

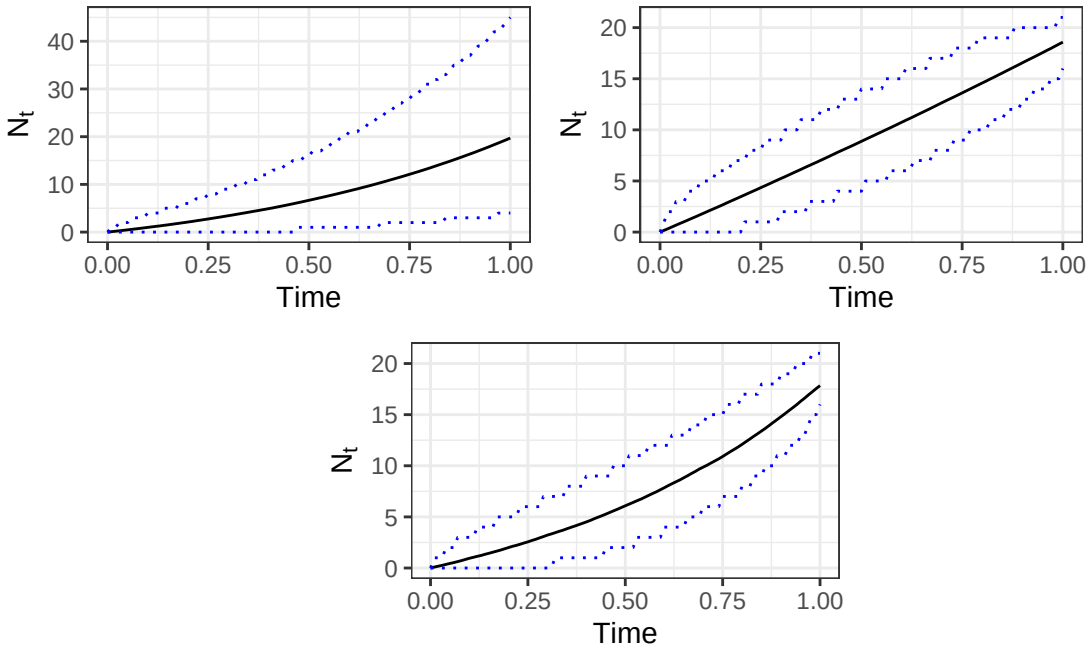

Figure 5.1: Mean (black) and 95% credible interval (blue) from 10^4 simulations of $N_{1,t}$ using the Poisson leap with the unconditioned hazard function (top left) and conditioned hazard function (top right). The bottom panel shows the same summaries from $1e4$ realisations of the conditioned process. All plots produced using $\Delta t = 0.01$, $\theta = (\exp(-6), 0.5)'$ and $x_0 = (762, 5)'$.

5.6 Discussion

This chapter considered a sequential approach to continued parameter and state inference. The Poisson leap model permits a tractable conditioned parameter posterior for the static rate constants and, in some situations, for the parameters governing the dynamic infection rate and observation models. This permits implementation of the approach in Storvik (2002), which we further modified by propagating particles conditioned on the following observation via a novel bridge construct. We underline that the Storvik approach can alleviate, but not wholly overcome, particle degeneracy due to the eventual collapse of the discrete representation of the sufficient statistic distribution (see e.g. Chopin et al., 2010, for a discussion of particle learning, and our discussion in Chapter 2). This is also evidenced by our simulation experiments (see Chapter 6) where many millions of parameter particles are required to give reasonable effective sample sizes. We also note that the bridge construct effectively gives a linear push towards the observation, which can be unsatisfactory when the dynamics between observations are nonlinear. Nevertheless, for the real data applications in Chapter 6, the bridge construct offers an effective method for propagating particles.

Chapter 6

Applications

In this chapter, we consider five applications of the methodology presented in Chapters 4 and 5. We begin by considering three different applications of the particle filter (Algorithm 15) using both synthetic and real data to demonstrate the effectiveness of the scheme. The first application uses synthetic data generated using the SIR model, allowing us to compare the accuracy of the resulting marginal parameter posteriors with a pseudo-marginal Metropolis-Hastings scheme. The other two applications are real data examples, taken from Fintzi et al. (2022) and Spannaus et al. (2022), respectively. In all cases, we set the time-step of the dS(E)IR model to be $\Delta\tau = 0.1$, which gave a reasonable balance between accuracy and computational cost. All algorithms were coded in R and run on a PC with a 2.6 GHz clock speed across twenty cores. Computer code to reproduce these experiments can be downloaded from https://github.com/Sam-Whitaker/Sequential_Bayes_Epi. In the final two applications, we make use of the methodology described in Chapter 4. Firstly, using synthetic data generated from the SIR model, we compare the performance of the two marginalisation techniques; these are the correlated pseudo-marginal

Metropolis-Hastings (CPMMH) and forward filtering Metropolis-Hastings (henceforth FFMH) based inference schemes from Sections 4.3.2 and 4.3.1 respectively. We compare the accuracy of posterior output from these schemes with inferences obtained by assuming the most natural Markov Jump process as the inferential model. We fit this model using the pseudo-marginal Metropolis-Hastings (PMMH) scheme described in Golightly and Wilkinson (2011). Additionally, we include inferences based on the deterministic ODE model of latent incidence (fit via MH). In the second application, we use FFMH to fit two models with two different choices of observation model, to a real data set consisting of pest removals from trees in a London park. The first model is the SIR model, as described in Section 3.2, and the second is an extension to the SIR model, namely the SIRS model, which allows for members of the removed species to become susceptible again through the addition of a suitable pseudo-reaction. The schemes used for these two applications all use random walk proposals with Gaussian innovations for the log-transformed parameters. For CPMMH, we fixed $\zeta = 0.99$, which we found to give a good balance between mixing over the auxiliary variable and parameter chains. We chose the number of particles N by following the practical advice of Deligiannidis et al. (2018). That is, we choose N so that the variance of $\log \hat{\pi}_{u^*}(y|\theta^*) - \log \hat{\pi}_u(y|\theta) \approx 1$. For CPMMH and FFMH, we took the random walk innovation variance to be $\widehat{\text{Var}}(\log \theta|y)$ estimated from a pilot run, and subsequently scaled to meet a desired empirical acceptance rate (see e.g. Schmon et al. (2021) for (C)PMMH and Schmon and Gagnon (2022) for Metropolis-Hastings). CPMMH was run for $50,000$ iterations and the remaining schemes were run for $10,000$ iterations, which we found gave reasonable mixing efficiency as measured by effective sample size (see e.g. Plummer et al., 2006).

6.1 Simulation study I

We consider the dSIR model and synthetic data consisting of the (noisy) number of new infections in $(t-1, t]$ for $t = 1, \ldots, 10$. The data were generated by applying Algorithm 9 with a time step of 0.001 over the time interval $[0, 10]$ and summing the number of infections over equally spaced sub-intervals of unit length. To emulate the influenza data set described in BMJ News and Notes (1978), we took the initial state to be $(i_0, s_0)' = (762, 5)'$, a removal rate of $\gamma = 0.5$ and assumed a time-varying infection rate whose logarithm is of the form

$$\mathrm{d}\tilde{\beta}_t = \lambda^{-1/2}\mathrm{d}W_t, \quad \tilde{\beta}_0 = -6 \quad \text{and} \quad \lambda = 100$$

so that the infection rate itself is scaled Brownian motion. Finally, we corrupted the data via a Binomial observation model with parameter $\rho = 0.9$. The data are shown in Figure 6.1.

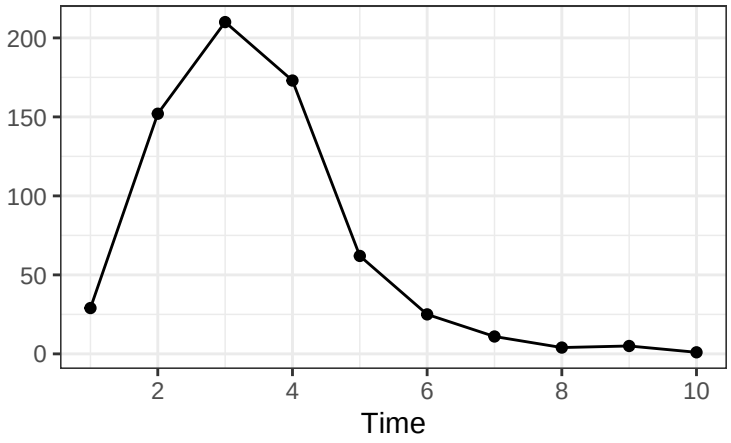

Figure 6.1: Synthetic data application. Number of new infections in $(t-1, t]$ for $t = 1, \ldots, 10$.

We adopted a weakly informative prior specification by taking $\log(\beta_0) \sim \mathrm{N}(-6.5, 0.5^2)$, $\gamma \sim \mathrm{Gamma}(11, 20)$, $\lambda \sim \mathrm{Gamma}(15, 0.14)$ and $\rho \sim \mathrm{Beta}(90, 15)$, and note that our

results are relatively insensitive to modest changes in the prior specification. The particle filter of Chapter 5 was run with $N = 10^k$ particles, for $= 1, 2, \ldots, 6$. The wall clock CPU time versus N is shown in Figure 6.2 for serial and parallel implementations; a single run with $N = 10^6$ takes approximately 16 minutes. The benefit of parallelising the particle filter is clear, with an order of magnitude speed-up (over a serial implementation) achieved for $N \geq 10^4$ particles.

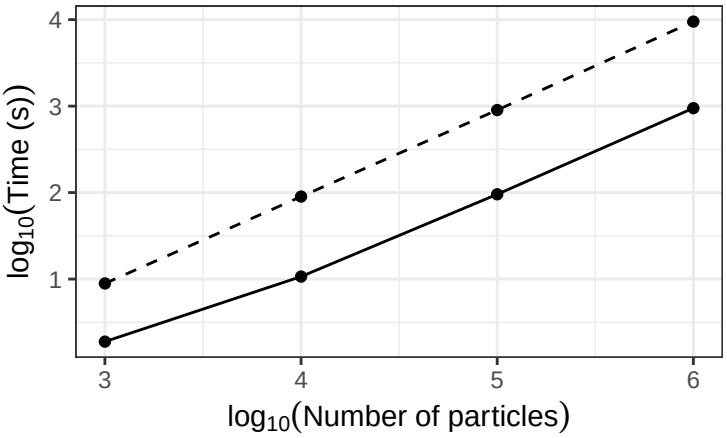

Figure 6.2: Synthetic data application. Wall clock CPU times (in seconds) for runs of the scheme in serial (dashed) and parallel (solid) versus the number of particles.

To benchmark accuracy for different values of N, we generated samples from the marginal parameter posteriors via a long run (10^6 iterations) of a pseudo-marginal Metropolis-Hastings scheme (PMMH, e.g. Andrieu et al., 2010) targeting the posterior under the dSIR model. For each component of the parameter vector $\psi = (\gamma, \lambda, \rho)'$ we computed "gold standard" estimates of the marginal posterior expectations $\mathrm{E}(\psi_i|y)$ and standard deviations $\mathrm{SD}(\psi_i|y)$, denoted by e_i and s_i, respectively. Replicate runs of the particle filter are then used to obtain corresponding estimates $\hat{e}_i^{(N,j)}$ and $\hat{s}_i^{(N,j)}$ where N denotes the number of particles used and $j = 1, \ldots, R$ indexes the replicate run. Bias and root mean squared error (RMSE) of the particle

filter's estimator of e_i is then calculated as

$$\text{Bias}(\hat{e}_i^{(N)}) = \frac{1}{R}\sum_{j=1}^{R}(\hat{e}_i^{(N,j)} - e_i) \quad \text{and} \quad \text{RMSE}(\hat{e}_i^{(N)}) = \sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{e}_i^{(N,j)} - e_i)^2}$$

with similar expressions obtained for the estimator of s_i. Bias and RMSE for the estimator of the marginal posterior expectation and standard deviation of γ (removal rate) and ρ (reporting rate) can be found in Figure 6.3. Unsurprisingly, Bias

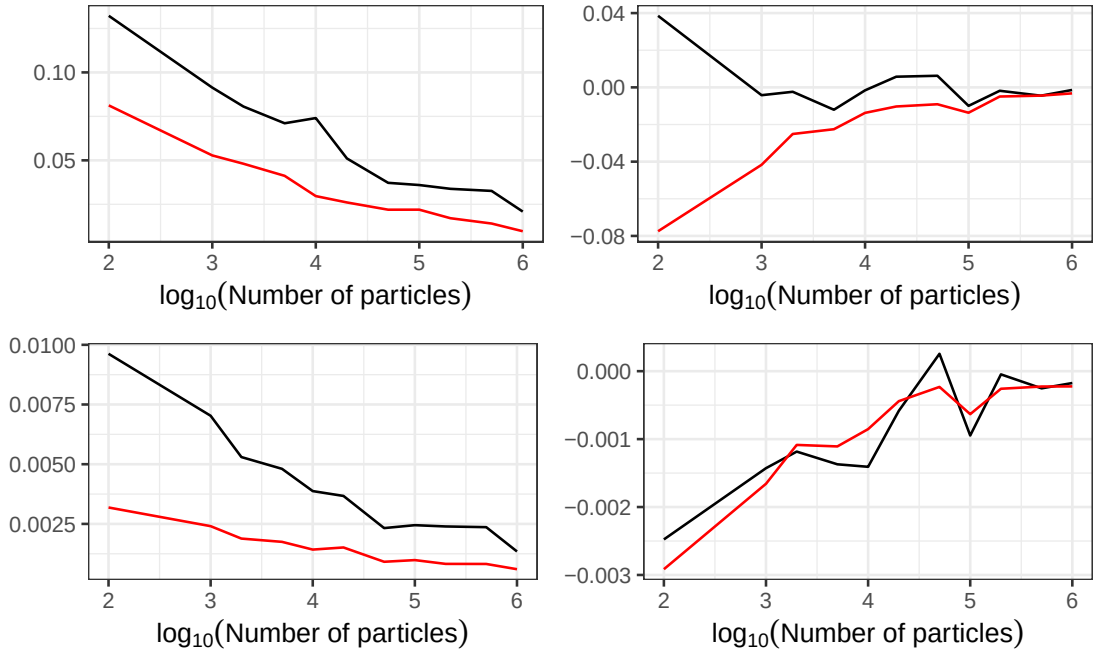

Figure 6.3: Synthetic data application. RMSE (left) and bias (right) versus $\log_{10}(N)$ for the particle filter's estimator of $e_1 = \text{E}(\gamma|y)$ (black, top row) and $s_1 = \text{SD}(\gamma|y)$ (red, top row) and $e_3 = \text{E}(\rho|y)$ (black, bottom row) and $s_1 = \text{SD}(\rho|y)$ (red, bottom row).

and RMSE reduce as the number of particles increases. There is relatively little improvement in accuracy beyond $N = 50K$, suggesting that for this scenario, this value of N gives a reasonable balance between accuracy and efficiency. This remark is further supported by Figure 6.4, which shows the particle filter sample output (after assimilating all observations) for $N = 50K$, with the gold standard PMMH

marginal posterior densities overlaid.

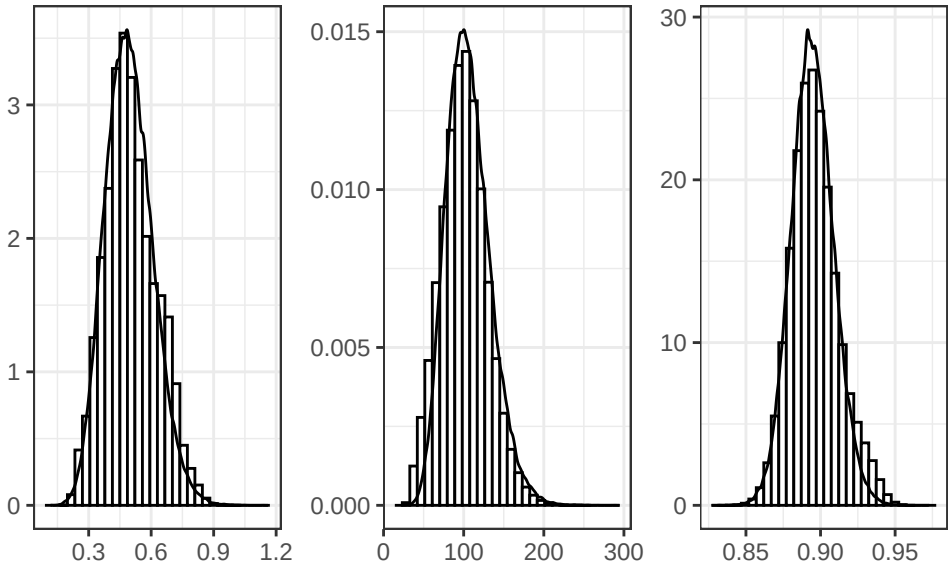

Figure 6.4: Synthetic data application. Posterior output from the particle filter (histograms) and ground truth posterior densities (kernel density estimates overlaid) were obtained via PMMH (10^6 iterations). Panels left to right are γ, λ and ρ respectively.

Summaries (mean and 95% credible interval) of the marginal filtering distributions $\pi(\psi_i|y_{[1,t]})$ and $\pi(\beta_t|y_[1,t])$ are shown in Figure 6.5. Samples from the filtering distributions are consistent with the ground truth parameter values that produced the data. Posterior uncertainty generally reduces as more data points are assimilated. However, there is little difference between the prior distribution of the precision parameter λ governing the time-varying infection rate and the filtering distribution of λ at each observation time-point; we anticipate that this parameter will be particularly sensitive to the choice of prior.

The particle filter gives samples of cumulative incidence over $(t_{i-1}, t_i]$, $\Delta n_{t_i}^{(1:N)}$, from the (marginal) filtering distributions $\pi(\Delta n_{t_i}|y_{[1,i]})$, $i = 1, \ldots, 10$. Given corresponding samples of $x_{t_{i-1}}^{(1:N)}$, the prevalence at time t_i can be computed for each sample k

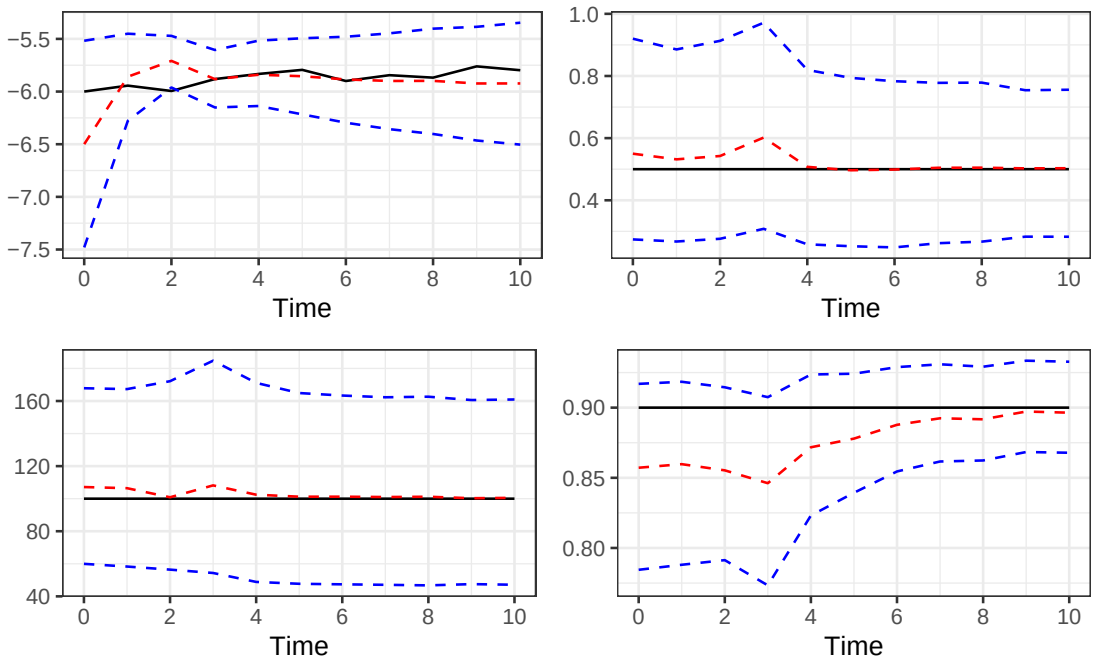

Figure 6.5: Synthetic data application. Filtering mean (red) and 95% interval (blue) of the log latent infection rate (top left), removal rate (top right), infection precision (bottom left) and reporting rate (bottom right). The ground truth is indicated (black).

via

$$x_{t_i}^{(k)} = x_{t_{i-1}}^{(k)} + \sum_j A'_j \Delta n_{j,t_i}^{(k)}.$$

Hence, samples from the filtering distributions for prevalence, $\pi(x_{t_i}|y_{[1,i]})$, are easily generated as part of the particle filter. Summaries of these filtering distributions are shown in Figure 6.6, from which we see that the filter output is consistent with the ground truth.

6.2 Ebola in West Africa

We consider the first of our three real data examples by applying the proposed dSEIR model and inference scheme to a subset of the data found in Fintzi et al.

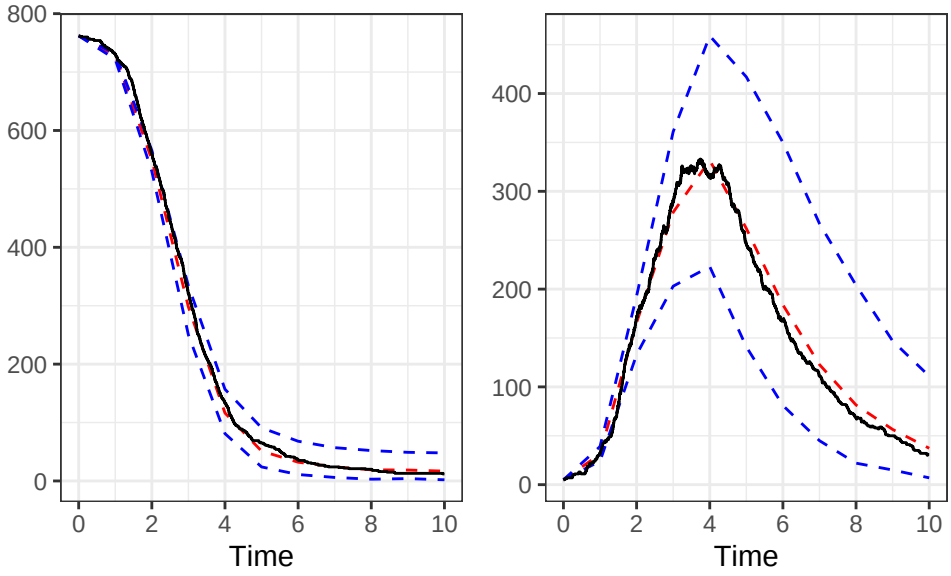

Figure 6.6: Synthetic data application. Filtering mean (red) and 95% credible interval (blue) for the susceptible (left) and infective (right) species. The ground truth is shown in black.

(2022). In particular, we only consider data from Sierra Leone. The data are 53 weekly observations of the incidence of Ebola from May 2014 to May 2015; these are shown in Figure 6.7.

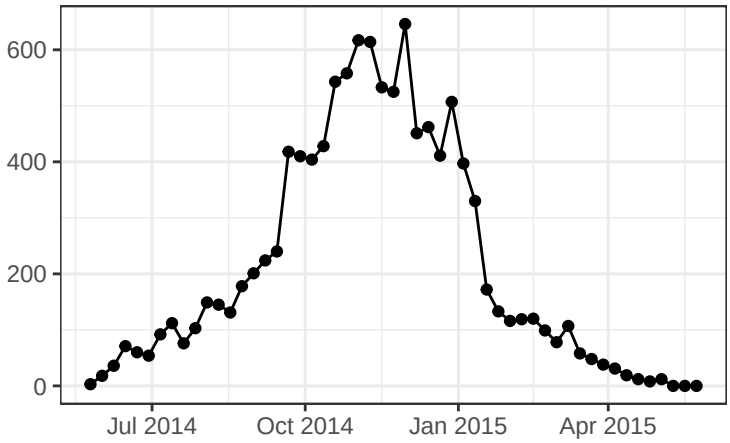

Figure 6.7: Ebola application. Weekly incidence data from the Ebola outbreak in Sierra Leone.

Following Fintzi et al. (2022), we specify a Negative Binomial observation model

to link the true unobserved incidence process to observations; see Equation (4.3). The authors include a constant immigration of cases, occurring with rate α, via a modification of the hazard of the first reaction, $h_1(x_t)$. As we only consider a single country, we ignore this proposed immigration, assuming that the import and export rates will be approximately equal. Instead, we fit the dSEIR model with a constant contact rate as described in Chapter 3.

Our choice of prior is consistent with the specification used in Fintzi et al. (2022); we take $\beta \sim \text{Gamma}(2, 50000)$, $\kappa \sim \text{Gamma}(5, 4.6)$, $\gamma \sim \text{Gamma}(10, 10)$, $\text{logit}(\rho) \sim \text{N}(0.85, 0.75^2)$ and $\nu \sim \text{Gamma}(5, 0.2)$. We assume the initial state is fixed and known to be $x_0 = (44326, 15, 10)'$. We found that taking the number of particles to be at least $N = 4 \times 10^6$ is necessary to avoid degeneracy in the parameter samples governing the observation model (for which we resort to the jittering approach described in Section 5.3). A single run of Algorithm 15 (parallelised over 20 cores) with this choice of N took approximately 321 minutes (or around 5.3 hours).

As in Section 6.1, we provide summaries (mean and 95% credible interval) of the marginal filtering distributions of each parameter in Figure 6.8. We see that the uncertainty reduces from prior to posterior for most parameters, with the exception of the reporting rate ρ, which remains almost unchanged. This reduction in uncertainty is most prominent for both the contact rate, β and the over-dispersion parameter ν.

Given in Figure 6.9 are summaries of the filtering distributions for the prevalence. We see that the credible region for the susceptible species narrows towards the middle of the data (around Nov-Dec 2014) before widening towards the end again. Conversely, the uncertainty increases towards the middle of the data in both the exposed and infectious species. The peaks in both the exposed and infectious species

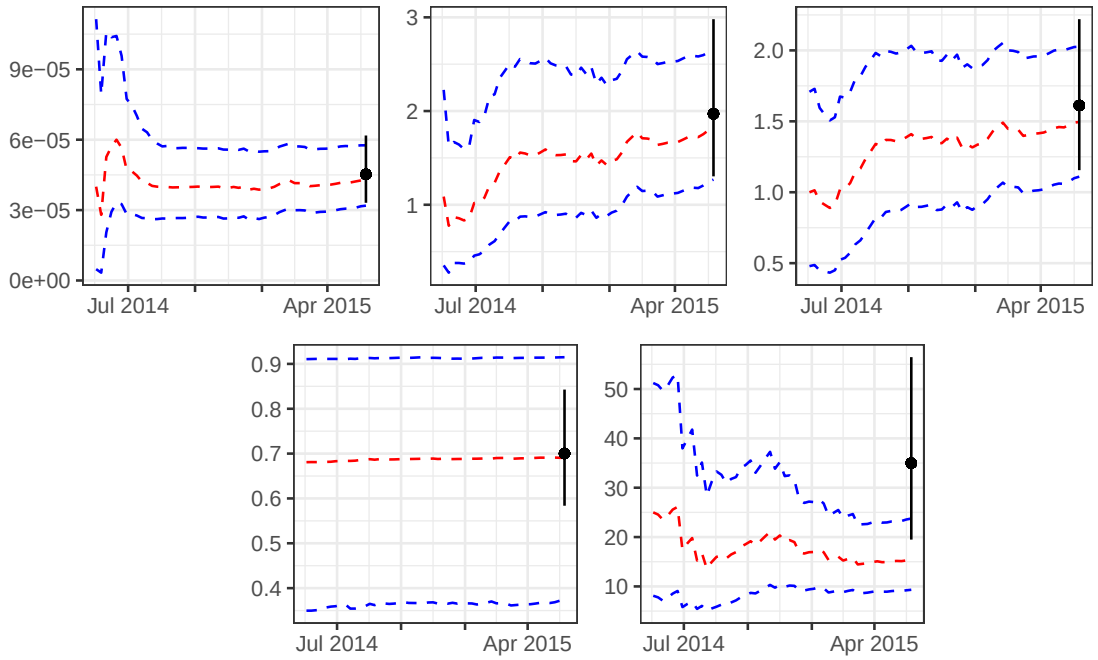

Figure 6.8: Ebola application. Filtering mean (red) and 95% credible interval (blue) of the contact rate (top left), infection rate (top centre), removal rate (top right), reporting rate (bottom left) and overdispersion parameter (bottom right) for the dSEIR model. Overlaid are filtering means (black dots) and 95% credible intervals (black lines) at the final time, using the linear noise approximation (LNA) as the transmission model.

correspond to peaks in the data, which is to be expected; as the population of the exposed/infectious species increases, so will the weekly incidence.

Finally, we produce five one-step-ahead forecasts of the weekly incidence for the last five non-zero observations in the data. To produce a forecast for the observed weekly incidence at an arbitrary time T, we run Algorithm 15 up to time $T-1$, and then, using the particle representation of $\pi(\phi, \tilde{\beta}_{[0,T-1]}, \Delta n_{(0,T-1]}|y_{[1,T-1]})$ from the output of Algorithm 15, we generate N realisations of the cumulative incidence process over $(T-1, T]$. The final cumulative incidences are subjected to noise as per the observation model, and the resulting samples are then summarised using a box and whisker plot. We repeat this process for each of the five forecast intervals of interest. The resulting box and whisker plots are provided in Figure 6.10. The

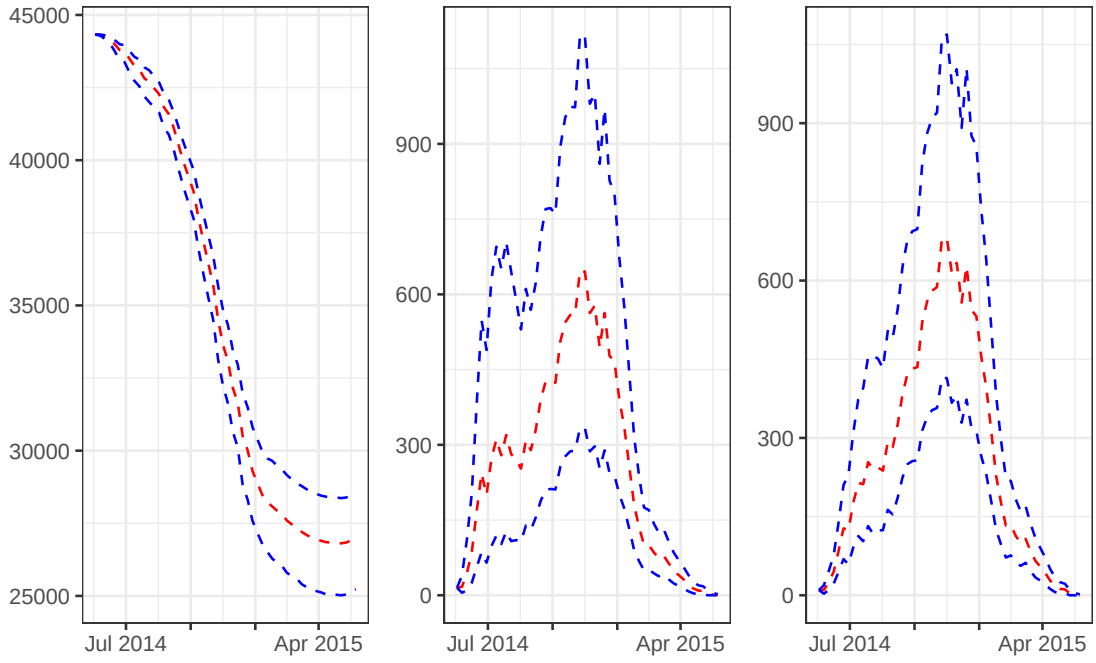

Figure 6.9: Ebola application. Filtering mean (red) and 95% credible interval (blue) for the susceptible, exposed and infective species under the dSEIR model.

observed values, given as dashed lines, lie close to the median value for each of the five forecasts and fall between the lower and upper quartiles, suggesting that the dSEIR model (in conjunction with the proposed inference scheme) can accurately forecast Ebola incidence.

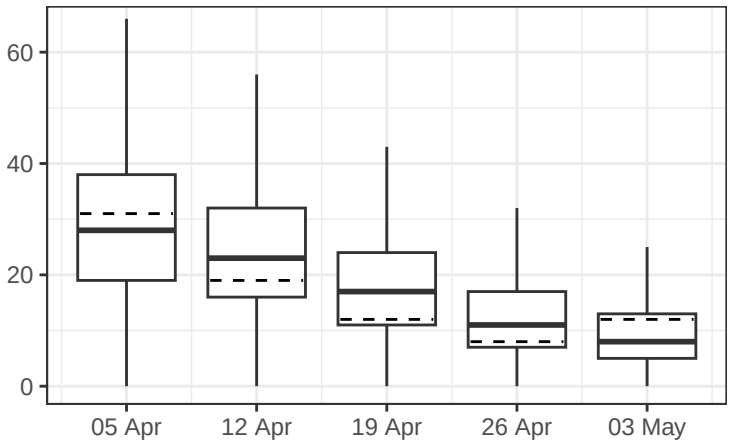

Figure 6.10: Ebola application. Five one step ahead dSEIR forecasts of the final five (non-zero) observed incidences overlaid with the true observed values (dashed).

Comparison with a linear noise approximation (LNA)

Here, we compare forecast accuracy of the dSEIR model to the approach taken by Fintzi et al. (2022), who describe SEIR dynamics using a linear noise approximation (see Chapter 3). That is, the LNA is used as the transmission model and the previously described Negative Binomial model is used as the observation model. In order to generate samples from the marginal parameter posterior, an MCMC scheme is used. In particular, and for reasons of computational efficiency, we implement the marginal Metropolis-Hastings scheme of Golightly et al. (2023); see Chapter 4 for brief details regarding the LNA and the MCMC scheme used in this setting.

The LNA requires a solution of a coupled ordinary differential equation (ODE) system. We used a simple Euler solver with a time step of 0.01, which gave a reasonable computational efficiency and accuracy balance. The prior specification is as previously described. Using a correlated random walk proposal, we ran the MCMC scheme for 100k iterations and tuned using the posterior variance from a shorter pilot run of 10k iterations. The resulting computational cost is approximately equal (regarding CPU wall clock time) to the particle filter run with $N = 4 \times 10^6$.

Posterior means and 95% credible intervals for the model parameters under the LNA are compared to the same posterior summaries obtained from the particle filter applied to the dSEIR model in Figure 6.8. It is evident that posterior summaries under the LNA are mostly consistent with those obtained under the dSEIR model, except for the overdispersion parameter, ν, whose posterior mass is concentrated around larger values than that from the particle filter. There is a clear reduction in reporting rate posterior variance under the LNA compared to dSEIR, although the mean from both approaches appears consistent.

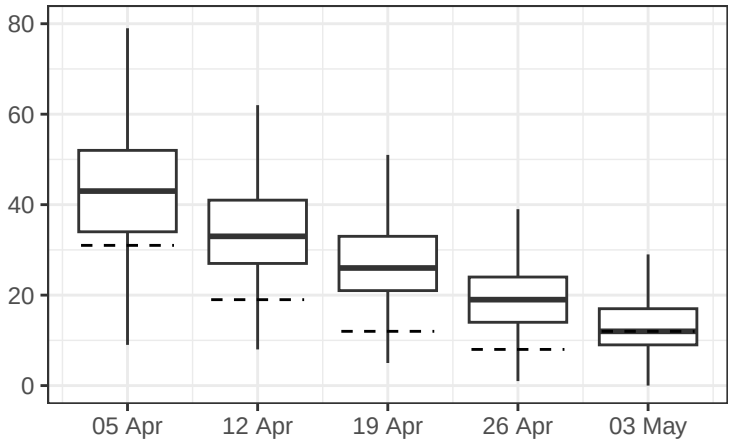

Figure 6.11: Ebola application. Five one step ahead LNA forecasts of the final five (non-zero) observed incidences overlaid with the true observed values (dashed).

As a final comparison, we now generate five one-step-ahead forecasts of the last five non-zero observations under the LNA. To generate a forecast at time T, we ran the LNA MCMC scheme using all the data up to and including time $T-1$, storing each accepted parameter vector along with a corresponding draw from $N_{T-1}|Y_{1:T-1}$. We then propagate a thinned sample forward over the interval $(T-1, T)$ to obtain estimates of the cumulative incidence at time T, which we subjected to noise according to the observation model. Box and whisker plots in Figure 6.11 summarise the forecasts. A comparison of Figures 6.10 and 6.11 suggests that the LNA tends to overestimate the observation, with four out of five median forecasts lying below the lower quartile.

6.3 Covid-19 in New York

We consider a data set taken from Spannaus et al. (2022) for the second real data application. The data consists of twenty-five observations of the weekly incidence of COVID-19 in New York, starting in March 2020 and ending in mid-August 2020.

The data are presented in Figure 6.12.

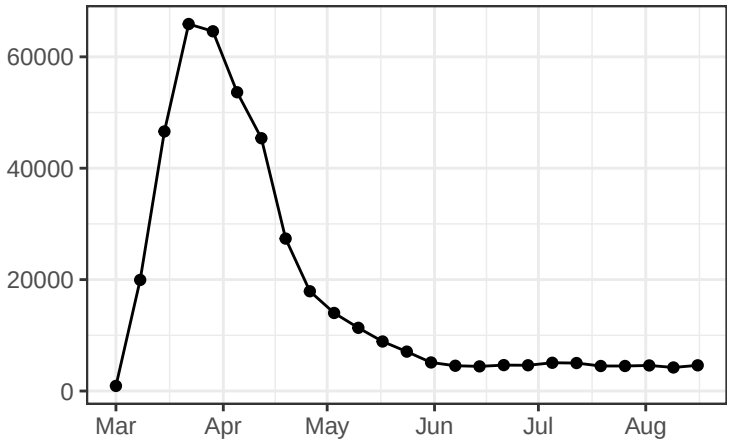

Figure 6.12: COVID-19 application. Weekly incidence in New York.

We adopt an approach consistent with the authors by modelling the data using an SIR model with dynamic contact and reporting rates, and a Negative Binomial observation model. Specifically, we allow the logit of the reporting rate to vary through time according to a scaled Brownian motion process whose precision is given by λ_ρ. To avoid ambiguity, we denote the scaled Brownian motion process's precision governing the contact rate log by λ_β. Hence, the dSIR transmission model is

$$\begin{cases} X_{t+\Delta t} = x_t + A'\Delta N_t, \\ \log \beta_{t+\Delta t} = \log \beta_t + \lambda_\beta^{-1/2} \Delta W_{1,t}, \\ \text{logit}\rho_{t+\Delta t} = \text{logit}\rho_t + \lambda_\rho^{-1/2} \Delta W_{2,t}. \end{cases} \tag{6.1}$$

where $W_{1,t}$ and $W_{2,t}$ are uncorrelated Brownian motion processes. The observation model is

$$Y_{t_i} | \Delta N_{t_i} \sim \text{NegBin}(\mu_i = \rho_{t_i} P' \Delta N_{t_i}, \sigma_i^2 = \mu_i + \mu_i^2/\nu), \quad i = 1, \ldots, L. \tag{6.2}$$

We follow Spannaus et al. (2022) by adopting a prior specification as follows. We

let $\log(\beta_0) = 0$, $\gamma \sim \text{Gamma}(11.088, 2.192)$, $\lambda_\beta \sim \text{Gamma}(4, 1)$, $\lambda_\rho \sim \text{Gamma}(4, 1)$, $\rho_0 \sim \text{Beta}(3, 2)$ and $1/\sqrt{\nu} \sim \text{U}(0, 0.5)$ *a priori*. We found that taking the number of particles to be at least $N = 4 \times 10^6$ avoids sample impoverishment. A single run of Algorithm 15 (parallelised over 20 cores) with this choice of N took approximately 49 minutes.

Summaries of the filtering distributions of the parameters and prevalence process are given in Figures 6.13 and 6.14. We can see that the uncertainty about the parameters reduces from prior to posterior in all but one of the cases, indicating that the analysis is informative.

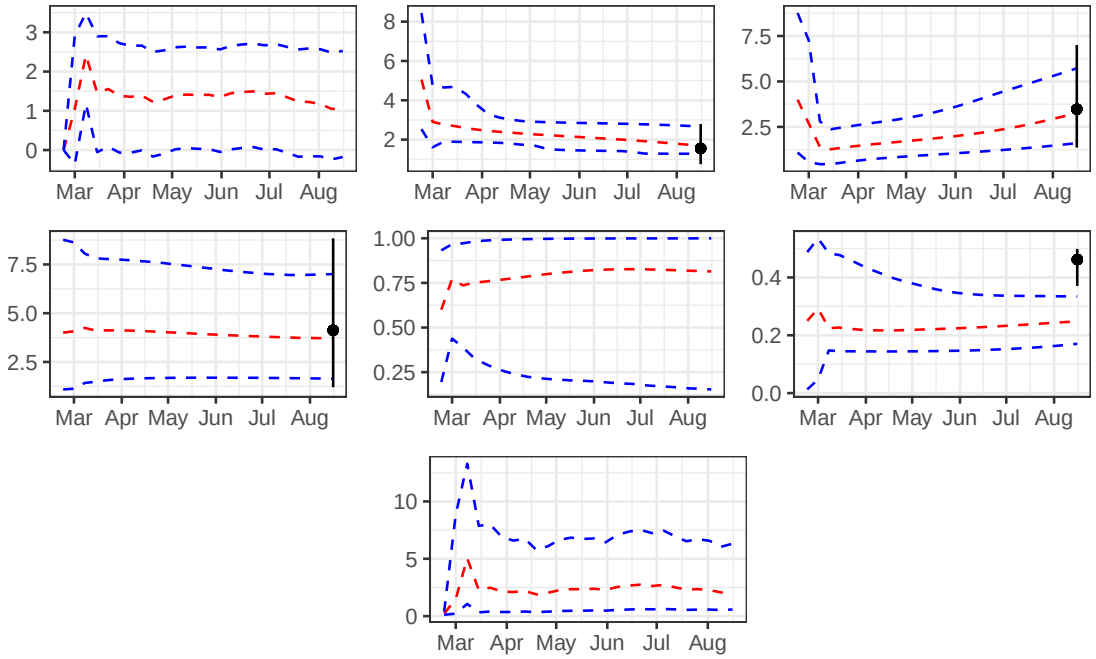

Figure 6.13: COVID-19 application. Mean (red) and 95% interval (blue) from the filtering distributions of $\log(N\beta_t)$ (top left), γ (top centre), λ_β (top right), λ_ρ (centre left), ρ_t (centre), $1/\sqrt{\nu}$ (centre right) and the basic reproductive number R_0 (bottom) under the dSIR model. In black are the mean (dots) and 95% credible intervals (lines) for each static parameter under the transmission model of Spannaus et al. (2022).

Finally, we assess the forecast accuracy of the dSIR model by providing five one-step-ahead forecasts for the final five non-zero observations, using the Monte Carlo

method described in Section 6.2. It is evident that all forecast medians are consistent with the observations, suggesting good forecast accuracy. The forecast for the final observation is comparable to that given in Spannaus et al. (2022) (see Figure 5 therein). Further comparisons with the transmission model used in the latter are considered in the next section.

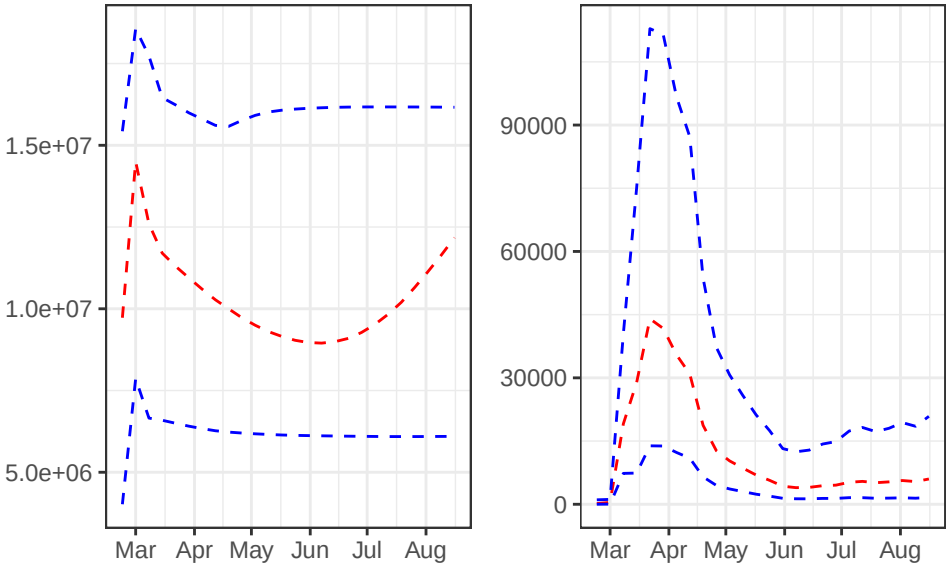

Figure 6.14: COVID-19 application. Mean (red) and 95% interval (blue) for the filtering distributions of the Susceptible (left) and Infective (right) species.

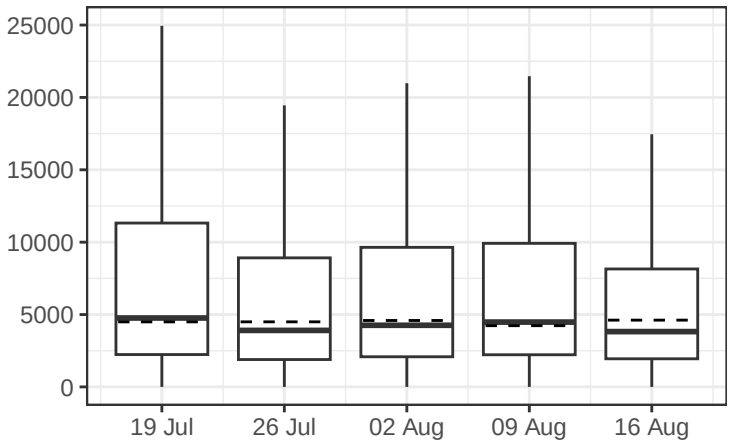

Figure 6.15: COVID-19 application. Five one-step-ahead dSIR forecasts of the final five observed incidences overlaid with the true observed values (dashed).

Comparison with Spannaus et al.

The SIR transmission model in Spannaus et al. (2022) takes the form of the well-studied Kermack-McKendrick ordinary differential equation (ODE) system, albeit with a time-varying infection rate. That is

$$\begin{cases} \mathrm{d}S_t = -\beta_t S_t I_t \mathrm{d}t, \\ \mathrm{d}I_t = (\beta_t S_t I_t - \gamma I_t)\mathrm{d}t, \\ \mathrm{d}R_t = \gamma I_t \mathrm{d}t, \end{cases} \tag{6.3}$$

where $\log \beta_t$ follows an SDE with solution as in Equation (6.1). To facilitate comparison with the dSIR model, we assume that the logit reporting rate $\mathrm{logit}\rho_t$ follows an SDE with solution as in Equation (6.1) and a Negative Binomial observation model as in Equation (6.2).

To generate samples from the marginal parameter posterior under the assumption of Equation (6.3), we ran the pseudo-marginal Metropolis-Hastings (PMMH) scheme described in Spannaus et al. (2022) for 10^6 iterations. We report the posterior mean and 95% credible interval for each of the static parameters in Figure 6.13. Parameter samples are consistent with those obtained under the dSIR model (via the particle filter) for γ, λ_β and λ_ρ. However, for the over-dispersion parameter ν, there is some disagreement between the two marginal posterior distributions, with samples of $1/\sqrt{\nu}$ typically smaller under dSIR than when using the transmission model in Equation (6.3).

We also provide five one-step-ahead forecasts under the assumption of the transmission model in Equation (6.3) for the final five non-zero observations in the data set in Figure 6.16. Comparing Figures 6.15 and 6.16, we see agreement between forecast

medians and the observations (for both schemes), although use of Equation (6.3) as the inferential and forecasting model appears to result in increased forecast uncertainty (relative to dSIR). This is likely due to the increased values of the inverse over-dispersion samples, resulting in a larger observation variance.

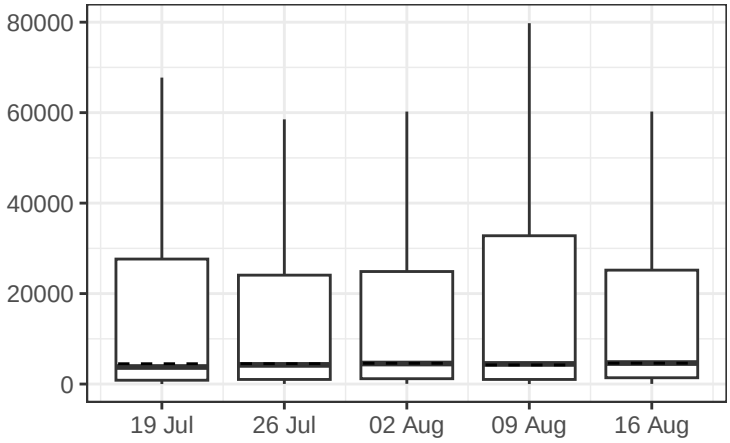

Figure 6.16: COVID-19 application. Five one-step-ahead forecasts of the final five observed incidences assuming the model of Spannaus et al. (2022), overlaid with the observed value (dashed).

6.4 Simulation study II

We now consider the first application of the methodology presented in 4, in the form of a second simulation study. To proceed, we generate three synthetic data sets from the SIR model, replicating epidemic outbreaks of increasing size, and for each we compare the performance of the competing marginalisation techniques, namely the correlated pseudo-marginal Metropolis-Hastings (CPMMH) and the forward filtering Metropolis-Hastings (FFMH) based inference schemes.

We generate the three synthetic data sets (denoted $\mathcal{D}_i$, $i = 1, 2, 3$) from the SIR model, each consisting of the number of new infections in time intervals $(t, t + 10]$ for $t = 10, 20, \ldots, 70$. For $\mathcal{D}_1$, we used $x_0 = (119, 1)'$ and $(\beta, \gamma)' = (0.00091, 0.082)'$;

these choices are consistent with inferences from the well-studied Abakaliki smallpox data (see e.g. Bailey, 1975). For $\mathcal{D}_2$, we constructed a larger outbreak by scaling the total population size N_{pop} and removal rate by a factor of 3, resulting in $x_0 = (359, 1)'$ and $(\beta, \gamma)' = (0.00091, 0.246)'$. For $\mathcal{D}_3$, we scaled N_{pop} by a factor of 10 (compared to $\mathcal{D}_1$) and set $x_0 = (1180, 20)'$. We scaled both the infection and removal rates to give $(\beta, \gamma)' = (0.00018, 0.164)'$. Note that all data sets have the same basic reproduction number, $R_0 = N_{\text{pop}}\beta/\gamma = 1.33$. We corrupted the resulting incidences via the Binomial observation model given in Equation (4.2) with $\rho = 0.8$ in each case. The data sets are shown in Figure 6.17 alongside the underlying traces of S_t and I_t (assumed unobserved).

When analysing $\mathcal{D}_1$, we adopted an independent prior specification with Gamma and Uniform components by taking $\beta \sim \text{Gamma}(10, 10^4)$, $\gamma \sim \text{Gamma}(10, 10^2)$ and $\rho \sim \text{Unif}(0, 1)$. The first two choices have been used by Fearnhead and Meligkotsidou (2004) and many others when analysing the Abakaliki smallpox data. For $\mathcal{D}_2$ and $\mathcal{D}_3$ we adopted a more diffuse prior for the removal rate to better reflect the increase in the ground truth value; specifically, $\gamma \sim \text{Gamma}(10, 30)$, with the prior specification for the remaining components as for $\mathcal{D}_1$. We assume that the initial state x_0 is fixed and known but note that inference for x_0 is possible by augmenting θ to include the components of x_0.

Table 6.1 and Figure 6.18 summarise the posterior output from each scheme (MJP based PMMH, LNA based CPMMH and LNA based FFMH). For data set $\mathcal{D}_1$ (population size $N_{\text{pop}} = 120$), although all inferential models give posterior output that is consistent with the ground truth parameter values, there are noticeable inconsistencies between LNA- and MJP-based inferences. Using the LNA to model the latent incidence process but with the correct observation model (LNA / (C)PMMH) results

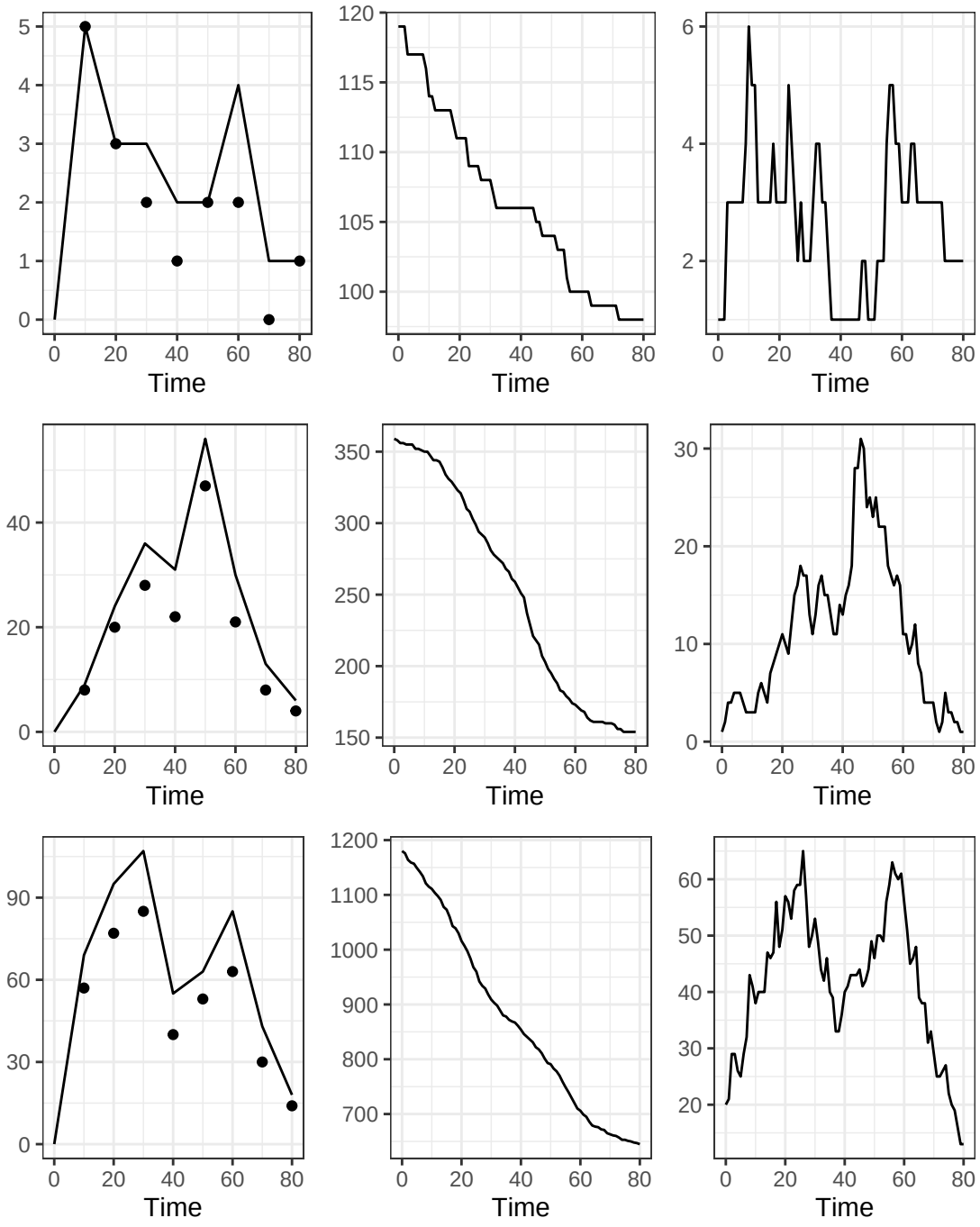

Figure 6.17: Synthetic data sets $\mathcal{D}_1$ (top panel), $\mathcal{D}_2$ (middle panel) and $\mathcal{D}_3$ (bottom panel). Left: noisy numbers of new infecteds in a 10-day interval (circles) and latent values (line). Middle and right: corresponding susceptible and infected states.

in an overestimation of the infection and removal rates, although there is little difference between this approach and the MJP when considering the basic reproduction number R_0. However, differences are more pronounced when further approximating the observation model as Gaussian (LNA / FFMH, see Section 4.3.1), which results in underestimation of R_0. Nevertheless, these differences are relatively small, and the advantages (in terms of overall efficiency) of analytically integrating out the latent process (as per LNA / FFMH) are clear.

We measure overall efficiency using minimum (over each parameter chain) effective sample size (ESS) per second (mESS/s). Given the small population size for $\mathcal{D}_1$, using the MJP inside a PMMH scheme is computationally more efficient than using the LNA (which, as implemented, requires numerical integration of 5 coupled ODEs per particle per iteration). Correlating successive likelihood estimates (LNA / CPMMH vs PMMH) increases overall efficiency by a factor of two. However, the largest gains in overall efficiency are obtained by LNA / FFMH, which improves on MJP / PMMH by a factor of 4 and on LNA / CPMMH by a factor of almost 50.

For data set $\mathcal{D}_2$ (population size $N_{\text{pop}} = 360$), the pseudo-marginal schemes require more particles due to the intrinsic stochasticity of realisations of the latent process generated inside the particle filter, which is large compared to observation noise. The increased population size (and corresponding parameter values that generated the data) leads to many more reaction occurrences between observation instants (compared to $\mathcal{D}_1$), reducing the relative efficiency of MJP / PMMH versus LNA / (C)PMMH and LNA / FFMH. We note that for this data set, using the LNA to model the latent process and additionally taking a linear Gaussian approximation of the observation model leads to an inference scheme that is both efficient (with an mESS/s 40 times larger than that of the next best performing scheme) and accurate

(see Figure 6.18, bottom panel).

The magnitude of typical observations in the data set $\mathcal{D}_3$ (population size $N_{\text{pop}} = 1200$) is broadly consistent with that of $\mathcal{D}_2$, and we find that the pseudo-marginal schemes require similar particle numbers. Using the LNA with the correct observation model gives parameter inferences that are consistent with the MJP. Using LNA / FFMH appears to result in underestimates of the infection and removal rates, although the basic reproduction number appears to be accurately estimated. In terms of overall efficiency, the advantage of LNA / FFMH over competing schemes is clear, with an mESS/s that is approximately 80 times larger than MJP / PMMH. There is relatively little difference between the performance of LNA / CPMMH and MJP / PMMH.

Table 6.1 also includes summarised posterior output when using a deterministic ODE model of latent incidence combined with a Gaussian approximation to the Binomial observation model (ODE / MH). This approach requires only the solution of the ODE system in Equation (3.15), as opposed to Equations (3.15), (3.16) and (3.17) when using the LNA. Consequently, an approximate 4-fold increase in overall efficiency is achieved for ODE / MH compared to LNA / FFMH. However, ignoring intrinsic stochasticity leads to a clear loss of inferential accuracy. In particular, the reporting rate is underestimated (which is unsurprising, as this leads to a larger observation variance, which can somewhat offset the inability of the latent ODE model to capture intrinsic stochasticity) and the basic reproduction number is typically overestimated. As N_{pop} increases, posterior uncertainty for all static parameters is underestimated (relative to the gold standard MJP approach) irrespective of each data set.

Model / Scheme	ζ	N	mESS/s	Mean (Std. Dev.)			
				β	γ	ρ	R_0
				Data set $\mathcal{D}_1$			
				0.00091	0.082	0.8	1.33
MJP / PMMH	0.00	30	0.682	0.00091 (0.00024)	0.088 (0.022)	0.62 (0.22)	1.29 (0.35)
LNA / PMMH	0.00	25	0.039	0.00107 (0.00027)	0.094 (0.025)	0.61 (0.24)	1.37 (0.41)
LNA / CPMMH	0.99	15	0.064	0.00111 (0.00024)	0.101 (0.027)	0.64 (0.22)	1.36 (0.40)
LNA / FFMH	–	–	3.075	0.00102 (0.00021)	0.117 (0.029)	0.82 (0.15)	1.09 (0.26)
ODE / MH	–	–	12.091	0.00203 (0.00024)	0.192 (0.042)	0.48 (0.16)	1.35 (0.42)
				Data set $\mathcal{D}_2$			
				0.00091	0.246	0.8	1.33
MJP / PMMH	0.00	125	0.091	0.00087 (0.00016)	0.225 (0.041)	0.75 (0.12)	1.35 (0.19)
LNA / PMMH	0.00	120	0.018	0.00092 (0.00017)	0.231 (0.050)	0.77 (0.12)	1.51 (0.23)
LNA / CPMMH	0.99	60	0.030	0.00089 (0.00016)	0.228 (0.056)	0.77 (0.12)	1.44 (0.25)
LNA / FFMH	–	–	3.680	0.00094 (0.00021)	0.234 (0.060)	0.78 (0.12)	1.49 (0.22)
ODE / MH	–	–	13.440	0.00087 (0.00005)	0.188 (0.021)	0.65 (0.04)	1.68 (0.10)
				Data set $\mathcal{D}_3$			
				0.00018	0.164	0.8	1.33
MJP / PMMH	0.00	120	0.045	0.00032 (0.00006)	0.290 (0.059)	0.82 (0.09)	1.33 (0.10)
LNA / PMMH	0.00	120	0.012	0.00034 (0.00007)	0.318 (0.070)	0.83 (0.10)	1.30 (0.11)
LNA / CPMMH	0.99	60	0.019	0.00034 (0.00006)	0.308 (0.062)	0.81 (0.12)	1.32 (0.12)
LNA / FFMH	–	–	3.533	0.00023 (0.00005)	0.217 (0.049)	0.85 (0.10)	1.29 (0.11)
ODE / MH	–	–	12.984	0.00019 (0.00001)	0.159 (0.008)	0.63 (0.01)	1.45 (0.04)

Table 6.1: Synthetic data application. Inferential model/scheme, correlation parameter, number of particles, minimum effective sample size per second and marginal parameter posterior summaries. The ground truth parameter values are indicated for each data set.

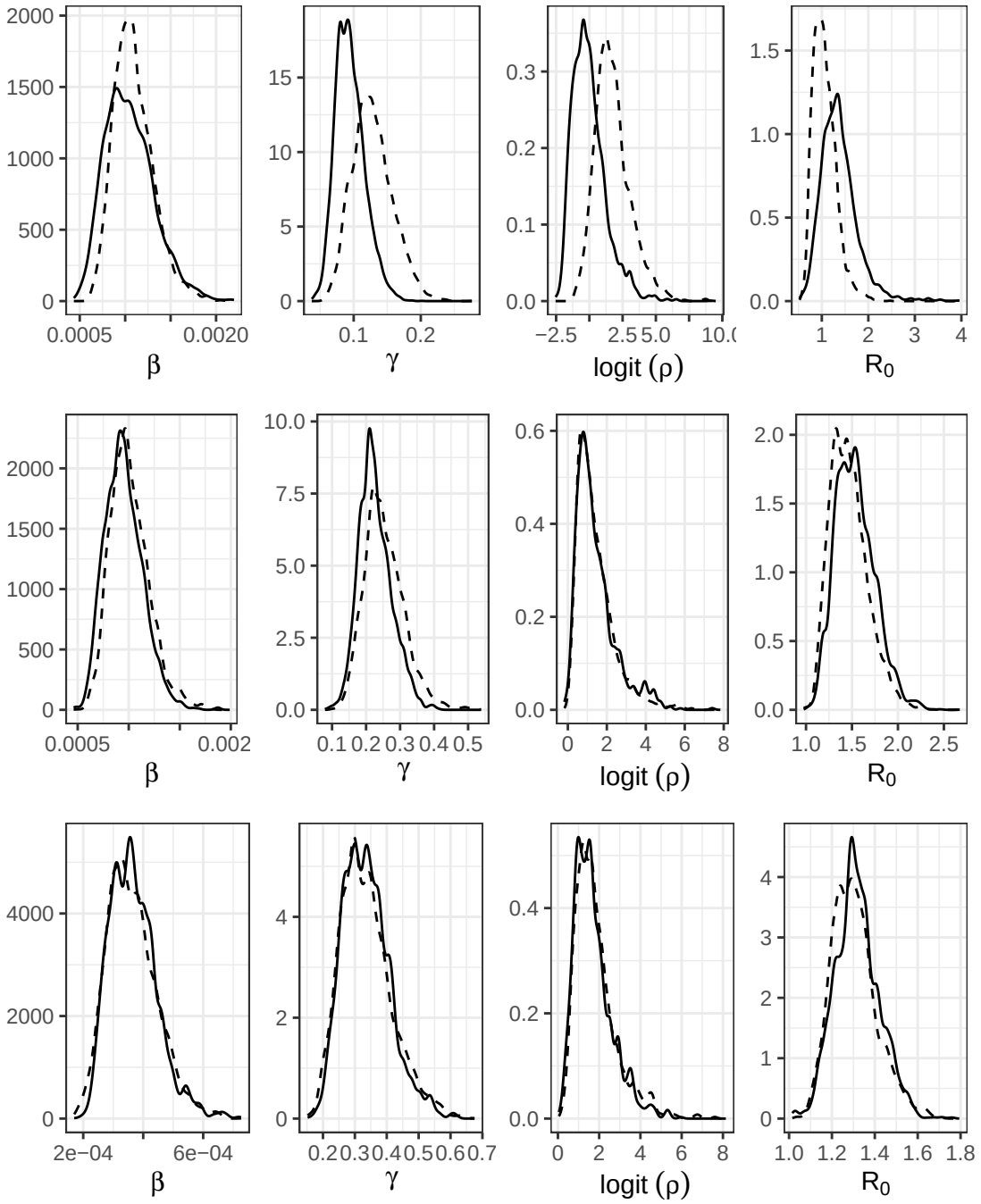

Figure 6.18: Synthetic data application. Marginal posterior densities based on $\mathcal{D}_1$ (top panel), $\mathcal{D}_2$ (middle panel) and $\mathcal{D}_3$ (bottom panel), and using the output of MJP / PMMH (solid line), LNA / CPMMH (dashed line).

6.5 Oak processionary moth in Richmond Park, London

In this section, we consider the application of the methodology presented in Chapter 4 to the infestation of the oak processionary moth (OPM), *Thaumetopoea processionea*, in Richmond Park, London. OPM is an invasive pest, destructive to oak trees and toxic to humans and animals (Maier et al., 2003, 2004; Gottschling and Meyer, 2006; Rahlenbeck and Utikal, 2015). The moth was first established in the UK in 2006 and despite efforts to initially eradicate, and then contain the infestation, OPM has continued to spread (Suprunenko et al., 2021; Wadkin et al., 2022).

The Royal Parks charity carries out surveys and control strategies for Richmond Park, and this data is then shared with the governmental Oak Processionary Moth Control Programme (Mainprize and Straw, 2021). The data records the numbers of OPM nests removed from trees (with recorded locations) between 2013 and 2020, allowing the formation of a time series for the yearly removal incidence of infested trees; see Table 6.2. The removal prevalence of the same set of trees (constructed under the assumption of known initial conditions) was considered in an SIR model in Wadkin et al. (2022). However, upon the manual removal of the OPM nests, the trees can become susceptible to re-infestation, and thus, we additionally consider the SIRS model below.

Year	2013	2014	2015	2016	2017	2018	2019	2020
No. removals	1024	1414	958	540	594	557	587	1029

Table 6.2: OPM data. Number of "removed trees" in a given year, Richmond park, London, 2013–2020.

6.5.1 Model and prior distribution

To allow for removed trees re-entering the susceptible class, we consider the SIRS compartment model shown graphically in Figure 6.19. Transitions between com-

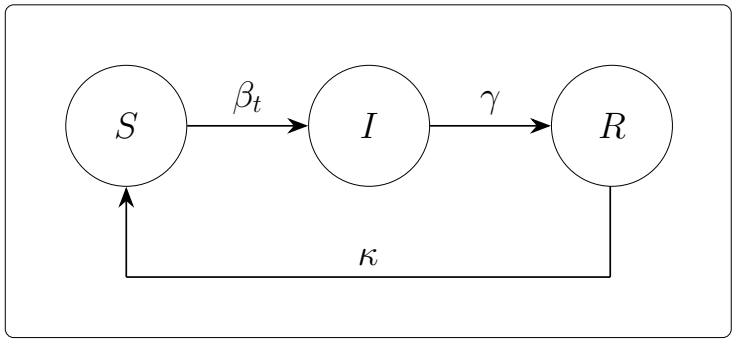

Figure 6.19: SIRS compartment model.

partments can be described by the set of pseudo-reactions given by

$$S + I \xrightarrow{\beta_t} 2I, \quad I \xrightarrow{\gamma} R, \quad R \xrightarrow{\kappa} S$$

where β_t is a time-varying infection rate whose natural logarithm is described by the SDE

$$d \log \beta_t = \sigma_\beta \, dW_t.$$

That is, the log infection rate is a scaled Brownian motion process. We note that setting $\kappa = 0$ gives the SIR model and in what follows we fit both SIR and SIRS models under the assumption the latent incidence process are well-described by the linear noise approximation. Details of the LNA for the SIRS compartment model is as follows.

Let $X_t = (S_t, I_t)'$ denote the numbers of susceptibles and infectives at time t. Similarly, let $n_t = (n_{1,t}, n_{2,t}, n_{3,t})'$ denote the cumulative number of infection, removal and loss of immunity (that is, removal to susceptible) events at time t. Let β_t, γ

and κ denote the corresponding event rates. The cumulative incidence $\{N_t, t \geq 0\}$ is an MJP governed by the transition probabilities

$$\mathbb{P}(N_{t+\Delta t} = (n_{1,t}+1, n_{2,t}, n_{3,t})' | n_t, x_t, \theta) = \beta_t s_t i_t \, \Delta t + o(\Delta t),$$
$$\mathbb{P}(N_{t+\Delta t} = (n_{1,t}, n_{2,t}+1, n_{3,t})' | n_t, x_t, \theta) = \gamma i_t \, \Delta t + o(\Delta t),$$
$$\mathbb{P}(N_{t+\Delta t} = (n_{1,t}, n_{2,t}, n_{3,t}+1)' | n_t, x_t, \theta) = \kappa(N_{pop} - s_t - i_t) \, \Delta t + o(\Delta t),$$
$$\mathbb{P}(N_{t+\Delta t} = (n_{1,t}, n_{2,t}, n_{3,t})' | n_t, x_t, \theta) = 1 - (\beta_t s_t i_t + \gamma i_t + \kappa[N_{pop} - s_t - i_t]) \, \Delta t + o(\Delta t)$$

and recall that N_{pop} is the total population size (assumed fixed and known). The stoichiometry matrix is given by

$$S = \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

and the hazard function is

$$h(x_t) = (\beta_t s_t i_t, \gamma i_t, \kappa[N_{pop} - s_t - i_t])'.$$

Using Equation (3.1) we may write the hazard function in terms of the incidence process as

$$\tilde{h}(n_t) = (\beta[s_0 - n_{t,1} + n_{t,3}][i_0 + n_{t,1} - n_{t,2}], \gamma[i_0 + n_{t,1} - n_{t,2}], \kappa[N_{pop} - s_0 - i_0 - n_{t,3} + n_{t,2}])'.$$

Now define $N_{4,t} = \log \beta_t$ as a Brownian motion process scaled by σ_β. This leads to the CLE for the SIRS model (with time-varying infection rate) as

$$dN_t = \left\{ \tilde{h}_1(n_t), \tilde{h}_2(n_t), \tilde{h}_3(n_t), 0 \right\} dt + \text{diag} \left\{ \sqrt{\tilde{h}_1(n_t)}, \sqrt{\tilde{h}_2(n_t)}, \sqrt{\tilde{h}_3(n_t)}, \sigma_\beta \right\} dW_t$$

where W_t is a length-4 vector of uncorrelated Brownian motion processes. The LNA then follows from Equations (3.15), (3.16) and (3.17), with the (transpose of the) Jacobian matrix F_t given by

$$F_t' = \begin{pmatrix} \exp(\eta_{4,t})(s_0 - i_0 - 2\eta_{t,1} + \eta_{t,2} + \eta_{3,t}) & \gamma & 0 & 0 \\ \exp(\eta_{4,t})(\eta_{t,1} - \eta_{t,3} - s_0) & -\gamma & \kappa & 0 \\ \exp(\eta_{4,t})(i_0 + n_{t,1} - n_{t,2}) & 0 & -\kappa & 0 \\ \exp(\eta_{4,t})(s_0 - \eta_{t,1} + \eta_{t,3})(i_0 + \eta_{t,1} - \eta_{t,2}) & 0 & 0 & 0 \end{pmatrix}.$$

We additionally consider two observation models: these are the Binomial and Negative Binomial models given by Equations (4.2) and (4.3). This leads to 4 competing models, which we compare using the deviance information criterion (DIC, see e.g. Gibson et al., 2018, for a discussion of DIC in the epidemic context) given by

$$\text{DIC} = -2\text{E}_\theta\{\log \pi(y|\theta)|y\} + p_D$$

where $p_D = -2\text{E}_\theta\{\log \pi(y|\theta)|y\} + 2\log \pi(y|\bar{\theta})$ measures the effective number of parameters in the model. Note that the observed data likelihood $\pi(y|\theta)$ is tractable under the Gaussian approximation approach to inference described in Section 4.3.1, which we employ here, and that $\text{E}_\theta\{\log \pi(y|\theta)|y\}$ is the posterior mean of the log-likelihood. Hence, DIC is easily calculated and the model with the smallest DIC is preferred.

We follow Wadkin et al. (2022) by fixing $N_{\text{pop}} = 40,000$, $x_0 = (38600, 1400)'$ and $\log \beta_0 = -10$. We adopt an independent prior specification by taking $\log \gamma \sim \text{N}(0, 0.5^2)$, $\log \kappa \sim \text{N}(0, 1)$, $\log \sigma_\beta \sim \text{N}(1, 1)$, $\lambda \sim \text{Unif}(0, 1)$ and, when using a Negative Binomial observation model, $\log \phi \sim \text{N}(0, 1)$. Note that the choice of β_0 and prior for the removal rate γ induces a prior on the basic reproduction number

$R_0 = N_{\text{pop}}\beta_0/\gamma$ at time 0 as lognormal $\text{logN}(0.6, 0.5^2)$. This gives a 95% equitailed credible interval of $(0.7, 4.8)$ for R_0, which reflects our belief that OPM spread is likely to persist without precluding $R_0 < 1$. We note also that the prior for κ gives a 95% credible interval of $(0.14, 7.10)$ years, reflecting vague prior beliefs on the time taken for a removed tree to re-enter the susceptible class.

6.5.2 Results

We ran the marginal Metropolis-Hastings scheme (as described in Section 4.3.1) for $50,000$ iterations, which we found took approximately 20 minutes, with the resulting parameter chains suggesting adequate mixing. Tables 6.4–6.5 and Figures 6.20-6.22 summarise the posterior output under each competing model. We also perform a short simulation study to ensure that the DIC, as defined above, is an appropriate model selection criterion in this setting. To this end, we simulate ten synthetic data points consisting of the (noisy) number of removals from the SIRS model via the Poisson leap, using the parameter posterior means from Table 6.2, and subject each to Binomial noise. Each of the four model combinations are used within the Metropolis-Hastings scheme, and the DIC values are computed.

Table 6.3 shows estimated DIC for the SIR and SIRS models, assuming either a Binomial or Negative Binomial observation model for the synthetic data described above. As previously discussed, the model with the smallest DIC is the preferred model. Thus, according to the DIC, the preferred model is the SIRS with Binomial noise regime, which is precisely the setup used to generate the data and hence we can conclude that the DIC is appropriate.

Table 6.4 shows estimated DIC for the SIR and SIRS models, assuming either a

Binomial or Negative Binomial observation model. A Binomial observation model is preferred irrespective of the assumed underlying compartment model. This is consistent with the inferred values of the (inverse) dispersion parameter ϕ, which are typically small; see Table 6.5. Hence, it appears that the true but unobserved removal incidence is much larger than the observed incidence, which is unsurprising given surveyed areas of Richmond Park in each year, which typically constitute a small fraction of the total area. Although our findings support the hypothesis that trees can become susceptible to re-infestation over the time scales of the data set considered, we note that the analysis has not been particularly informative for the parameter κ, governing the rate of R to S transitions; see Figure 6.20 showing marginal posterior densities for parameters in the SIRS model and the prior specification. Since removed trees are treated with insecticide, this parameter will likely interest practitioners. Nevertheless, improved data collection protocols and a longer study period may provide a partial record of the number of R to I transitions, which would greatly improve inferences on κ.

Figure 6.21 summarises the within-sample predictive distributions for the susceptible and infective prevalence processes (which are easily reconstructed from the predicted incidences, not shown) and the log infection process, $\log \beta_t$. These results are broadly consistent with those of Wadkin et al. (2022), which suggest a plausibly constant infection rate and an uptick in infected trees from 2018. Figure 6.22 summarises the marginal posterior distribution of the basic reproduction number R_0 against year. Sampled values of R_0 appear largely consistent across years. However, the marginal posterior distributions in 2018, 2019 and 2020 have the greatest support for $R_0 > 1$, suggesting that OPM will continue propagating in Richmond Park.

Model	SIR (Bin)	SIR (Neg Bin)	SIRS (Bin)	SIRS (Neg Bin)
DIC	172.4	180.2	161.2	165.5

Table 6.3: OPM data application (synthetic data). Estimated DIC for the SIR and SIRS models, assuming either a Binomial (Bin) or Negative Binomial (Neg Bin) observation model.

Model	SIR (Bin)	SIR (Neg Bin)	SIRS (Bin)	SIRS (Neg Bin)
DIC	115.4	120.4	113.4	119.8

Table 6.4: OPM data application. Estimated DIC for the SIR and SIRS models, assuming either a Binomial (Bin) or Negative Binomial (Neg Bin) observation model.

	Mean (Standard Deviation)			
Parameter	SIR (Bin)	SIR (Neg Bin)	SIRS (Bin)	SIRS (Neg Bin)
γ	0.90 (0.30)	1.44 (0.44)	0.96 (0.31)	1.45 (0.47)
κ	–	–	1.57 (2.02)	1.26 (1.72)
σ_β	0.64 (0.24)	0.58 (0.58)	0.56 (0.22)	0.51 (0.39)
λ	0.64 (0.16)	0.57 (0.20)	0.61 (0.16)	0.62 (0.20)
ϕ	–	0.30 (0.38)	–	0.25 (0.32)

Table 6.5: OPM data application. Marginal parameter posterior summaries.

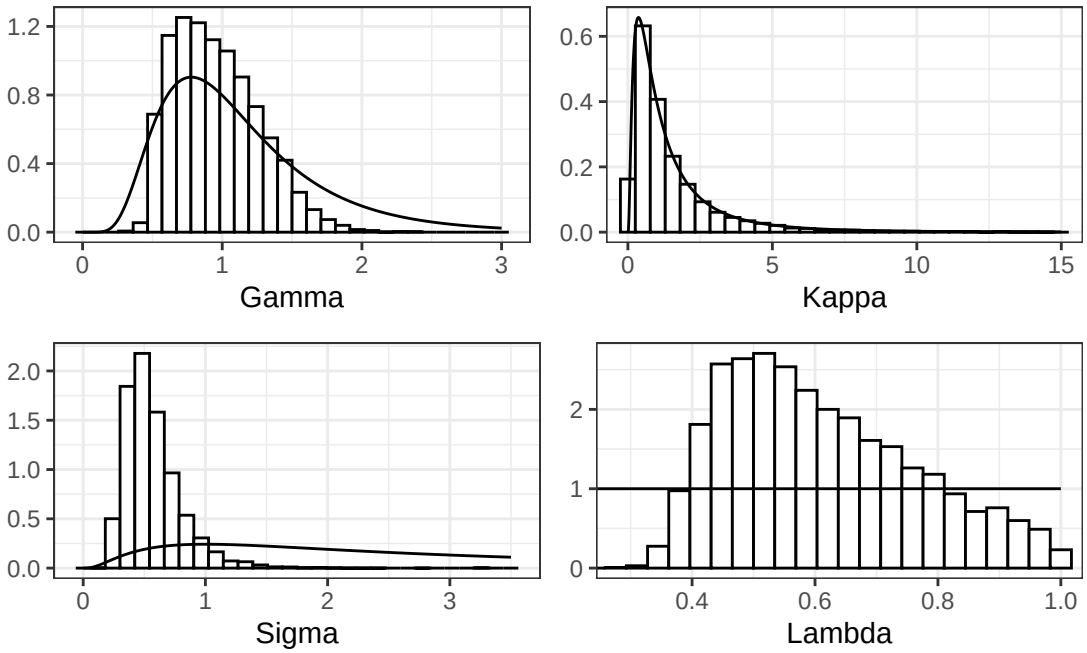

Figure 6.20: OPM data application. Marginal posterior densities (histograms) and prior (solid line), of the parameters in the SIRS model assuming binomial observations.

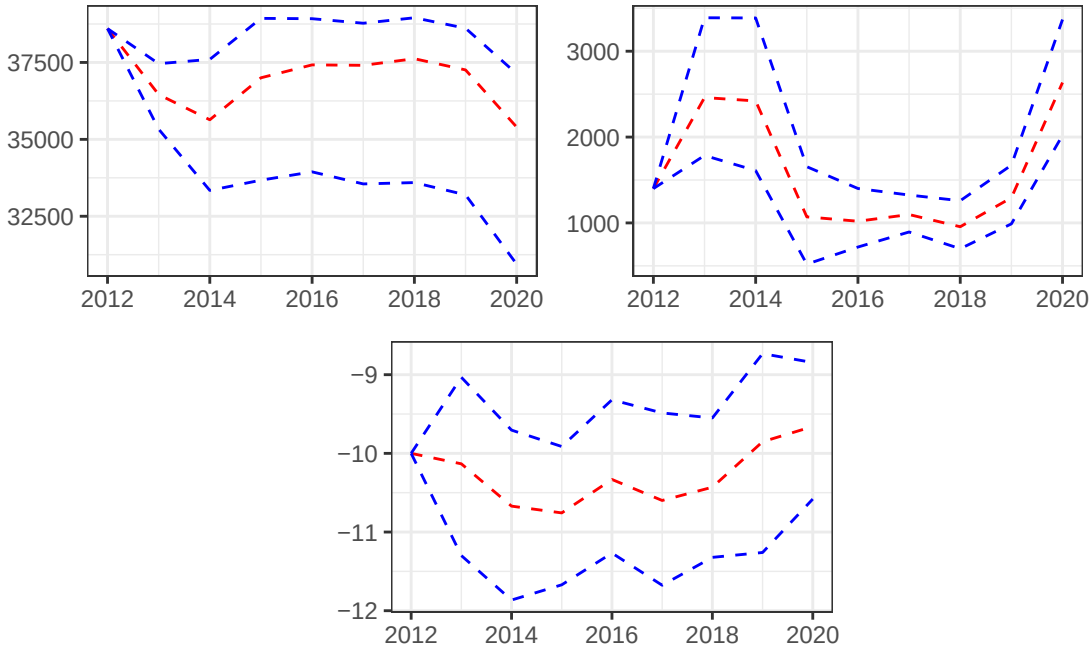

Figure 6.21: OPM data application. Within-sample predictive distributions (mean and 95% credible intervals) for S_t (top left) and I_t (top right) and $\log \beta_t$ (bottom).

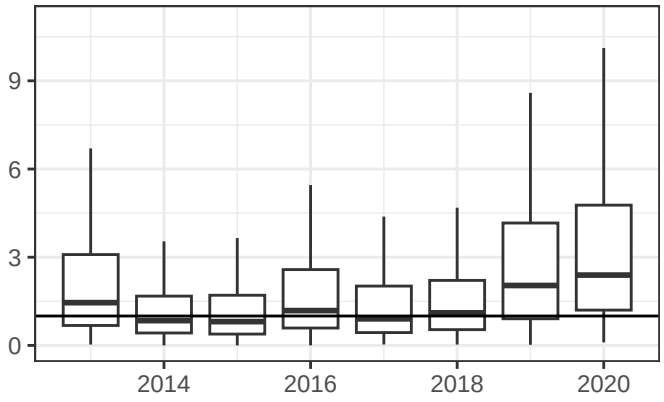

Figure 6.22: OPM data application. Boxplots summarising the marginal posterior distribution of the basic reproduction number R_0 against year.

Chapter 7

Conclusions

In this thesis, we have considered the challenging problem of performing Bayesian inference for the static parameters and latent, dynamic components of a stochastic representation of an epidemic compartment model. The inference task was developed assuming that only imperfect, discrete-time incidence values are available.

In the first contribution, the most natural continuous-time Markov jump process (MJP) representation of cumulative incidence was replaced by a discrete-time approximation in which the numbers of each transition event (e.g. exposure, infection or removal in an SEIR model) are assumed to be Poisson distributed over a time interval whose length is specified by the practitioner. The model allows for the straightforward incorporation of time-varying parameters via time-discretised stochastic differential equations (SDEs). The resulting dSEIR model admits several inferential advantages over other (stochastic) approximations of the MJP. Conditional on complete information regarding the cumulative incidence process, the static parameters governing the latent transmission model, and, in some cases, the observation model, follow a tractable posterior distribution. We leveraged this tractability

via a sequential Monte Carlo scheme (particle filter) that rejuvenates parameter samples by drawing from their conditional parameter posterior after propagating and resampling particle trajectories of the latent incidence process. This approach essentially follows the framework of Storvik (2002) (see also Carvalho et al., 2010, for a related particle learning approach). We further modified the particle filter by propagating particle trajectories conditional on the next observation by adapting the conditioned hazard of Golightly and Wilkinson (2015) to the cumulative incidence setting.

We applied the resulting methodology to synthetic and real data examples, benchmarking against competing models and inference schemes. The proposed particle filter benefits from a parallel implementation, as demonstrated using synthetic data generated from a stochastic SIR model with a time-varying infection rate and Binomial observation model. We observed an order of magnitude speed-up (over a serial implementation) for $N \geq 10^4$ particles. We also achieve satisfactory inferential accuracy, as measured by bias and RMSE for key posterior quantities, relative to corresponding values obtained from long runs of a pseudo-marginal Metropolis-Hastings scheme. We compared the forecast accuracy of the dS(E)IR model to two competing approaches via real data applications involving the spread of Ebola in West Africa (Fintzi et al., 2022) and COVID-19 in New York (Spannaus et al., 2022). In the former, a linear noise approximation was used as a tractable approximation to the MJP representation of cumulative incidence. For a fixed computational budget, the dSEIR model better reflects the end of the epidemic, likely due to relatively low incidence numbers, for which a continuous-valued approximation is likely to give a relatively poor approximation. For the third application involving COVID-19, we compared against a transmission model in which the infection rate of a deterministic ordinary differential equation (ODE) system is replaced with the solution of an SDE.

We found that this approach leads to accurate point-summary forecasts but overestimates uncertainty (relative to the dSIR model). Our empirical findings suggest that using the particle filter allows for assimilating a new observation in around six minutes for the dSEIR model and two minutes for the simpler dSIR model.

In the second contribution, we have proposed a fast and efficient method for inferring the parameters governing a stochastic epidemic model's linear noise approximation (LNA), using incidence data consisting of the cumulative number of new infections (or removals) in fixed-length windows. This setting is considered in Fintzi et al. (2022) who combines the LNA of the incidence process with a Negative Binomial observation model and develops an efficient MCMC scheme targeting the joint density of the parameters and latent incidence process. Our contribution, on the other hand, is a framework for marginalising out the latent incidence process, either by exactly targeting the marginal parameter posterior via a (correlated) pseudo-marginal method or analytically through a Gaussian approximation of the observation process. We additionally allow for a flexible, time-varying stochastic infection rate, naturally handled within the LNA framework. Our experiments demonstrated that use of the LNA and a further Gaussian approximation of an observation model can be both accurate and efficient. Using parameter values inspired by the Abakaliki smallpox outbreak, we investigated the accuracy and efficiency of the analytically marginalised LNA as the population size increased (and with the parameters scaled appropriately). In the 'large epidemic' setting ($N_{\text{pop}} = 1200$), the analytic marginalisation scheme outperforms the next best performing scheme by about a factor of 80. In this scenario, using the most natural Markov jump process representation of the epidemic is computationally prohibitive.

We further illustrated our approach via an application with real data consisting of

numbers of trees infested with oak processionary moth (OPM) nests in Richmond Park, London. Typical observations consist of around 500–1500 removals in a given year, with a total population size of around 40,000 trees, thus necessitating the efficient inference methods developed here. As well as inferring key quantities of interest, such as the basic reproduction number and latent susceptible and infective trajectories, our approach allows for easy computation of the observed data likelihood, which can be used, for example, to compute a deviance information criterion (DIC). We used DIC to compare two different compartment models (SIR versus SIRS) and two observation models (Binomial versus Negative Binomial). Our analysis suggests the SIRS model as the best-fitting compartment model, suggesting trees can re-enter the susceptible class following removal (via treatment). Although improved data collection protocols, which include observation of the number of removed trees which subsequently become infected, will greatly improve predictive power, our approach demonstrates that meaningful conclusions on the spread of OPM can be drawn despite a data-poor scenario.

7.1 Limitations and extensions

Here, we consider several limitations of the proposed approach and make some suggestions for future work.

The dS(E)IR model requires the specification of a step size, which balances accuracy (relative to the most natural MJP representation of cumulative incidence) and computational cost. We have assumed a fixed step size but note that an optimal choice is likely to depend on the phase of the epidemic. Conditions under which the Poisson approximation is reasonable have been discussed by Gillespie (2001) among

others; for methods that allow a step size to be chosen adaptively, see e.g. Cao et al. (2006); Sandmann (2009).

The particle filter used in this thesis propagates particles conditional on the next observation by replacing the Poisson rate with a hazard function derived via a Gaussian approximation of the cumulative incidence between the current time and next observation, conditional on the observation itself. A further assumption in deriving the resulting conditioned hazard is linearity of the latent process over a time interval of length at most given by the observation interval. In scenarios in which the epidemic is sparsely observed in time, to the extent that cumulative incidence does not evolve in a linear fashion, the conditioned hazard is likely to result in trajectories that are inconsistent with the true conditioned process. Recent work on sampling conditioned jump processes (in a chemical kinetics setting) is relevant here (see e.g. Golightly and Sherlock, 2019; Corstanje et al., 2023) and could be adapted to the dS(E)IR model.

A key feature of the proposed inference scheme is leveraging a tractable conditional parameter posterior, which depends (given complete information on the latent process up to time t_i) on a low-dimensional sufficient statistic T_i. This approach also forms the basis of particle learning (Carvalho et al., 2010). Particle degeneracy in the latter suite of algorithms has been well documented (see e.g. Section 4 of Chopin et al., 2010). That is, the number of different values of the statistic at a fixed time point prior to t_i that contribute to T_i is decreasing in i at an exponential rate. The applications presented here involve relatively short time-series coupled with a parallel implementation that allows sufficiently many particles to avoid degeneracy. Nevertheless, application to data sets involving longer time horizons remains of interest, as does a comparison with SMC schemes that use resample-move steps, such

as SMC^2.

Furthermore, within the stochastic kinetic models context, the LNA can be derived directly from the most natural Markov jump process (MJP) representation (Kurtz, 1970, 1972) but is perhaps most intuitively viewed as a tractable Gaussian process approximation of the Itô Stochastic differential equation (SDE) that best matches the MJP representation (Ferm et al., 2008; Fearnhead et al., 2014). As advocated by Fuchs (2013) among others, judging the validity of these continuous-valued approximations should involve comparison with the MJP (e.g. via simulations) for the specific system considered. Nevertheless, we expect, in general, that the best matching SDE and LNA approaches are likely to approximate the MJP particularly poorly when species numbers are comparatively small (e.g. in the few tens). In such situations, we envisage that our approach will likely be of most practical benefit in providing initial values and tuning choices for simulation-based inference schemes that target the posterior under the MJP. For inherently multi-scale epidemics, it may be possible to leverage hybrid simulation techniques (see e.g. Sherlock et al., 2015) whereby the LNA is used to model species which frequently change state, coupled with a discrete stochastic updating procedure for species which change state less often.

This work can be further extended in several ways. Of particular interest to us is the use of the proposed approach within a spatio-temporal setting and with application to OPM spread, for example, by allowing the importation of pests from nearby locations. Extension of the methodology to allow incorporation of multiple data streams (see e.g. Corbella et al., 2022) also merits further attention.

Bibliography

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

Andersson, H. and Britton, T. (2012). *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697 – 725.

Bailey, N. (1975). The mathematical theory of infectious diseases and its applications. *The mathematical theory of infectious diseases and its applications. 2nd edition.*, (2nd edition).

Begon, M., Bennett, M., Bowers, R., French, N., Hazel, S., and Turner, J. (2002). A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiology and Infection*, 129(1):147–153.

BMJ News and Notes (1978). Influenza in a boarding school. *British Medical Journal*, pages 1–587.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC press.

Cao, Y., Gillespie, D. T., and Petzold, L. R. (2006). Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(4).

Carvalho, C. M., Johannes, M. S., Lopes, H. F., and Polson, N. G. (2010). Particle Learning and Smoothing. *Statistical Science*, 25(1):88 – 106.

Cauchemez, S. and Ferguson, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of the Royal Society Interface*, 5(25):885–897.

Chang, C.-H. (2008). Skew-normal approximation to the negative binomial distribution.

Chopin, N., Iacobucci, A., Marin, J.-M., Mengersen, K., Robert, C. P., Ryder, R., and Schäfer, C. (2010). On particle learning. *arXiv preprint arXiv:1006.0554*.

Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo*, volume 4. Springer.

Choppala, P., Gunawan, D., Chen, J., Tran, M.-N., and Kohn, R. (2016). Bayesian inference for state space models using block and correlated pseudo marginal methods. *arXiv preprint arXiv:1612.07072*.

Corbella, A., Presanis, A. M., Birrell, P. J., and De Angelis, D. (2022). Inferring epidemics from multiple dependent data via pseudo-marginal methods. *arXiv preprint arXiv:2204.08901*.

Corstanje, M., van der Meulen, F., and Schauer, M. (2023). Conditioning continuous-time Markov processes by guiding. *Stochastics*, 95(6):963–996.

Dahlin, J., Lindsten, F., Kronander, J., and Schön, T. B. (2015). Accelerating pseudo-marginal Metropolis-Hastings by correlating auxiliary variables. *arXiv preprint arXiv:1511.05483*.

Del Moral, P. (2004). *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). The correlated pseudo-marginal method. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(5):839–870.

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking epidemics with Google flu trends data and a state-space SEIR model. *Journal of the American Statistical Association*, 107(500):1410–1426.

Dureau, J., Kalogeropoulos, K., and Baguelin, M. (2013). Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, 14(3):541–555.

Elvira, V., Martino, L., and Robert, C. P. (2022). Rethinking the Effective Sample Size. *International Statistical Review*, 90(3):525–550.

Fearnhead, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics*, 11(4):848–862.

Fearnhead, P., Giagos, V., and Sherlock, C. (2014). Inference for reaction networks using the linear noise approximation. *Biometrics*, 70(2):457–466.

Fearnhead, P. and Meligkotsidou, L. (2004). Exact filtering for partially-observed continuous-time models. *Journal of the Royal Statistical Society, Series B*, 66:771–789.

Ferm, L., Lötstedt, P., and Hellander, A. (2008). A hierarchy of approximations of the master equation scaled by a size parameter. *Journal of Scientific Computing*, 34(2):127–151.

Finkenstädt, B., Woodcock, D. J., Komorowski, M., Harper, C. V., Davis, J. R. E., White, M. R. H., and Rand, D. A. (2013). Quantifying intrinsic and extrinsic noise in gene transcription using the Linear Noise Approximation: An application to single cell data. *The Annals of Applied Statistics*, 7(4):1960–1982.

Fintzi, J., Wakefield, J., and Minin, V. N. (2022). A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. *Biometrics*, 78(4):1530–1541.

Fuchs, C. (2013). *Inference for diffusion processes: with applications in life sciences*. Springer Science & Business Media.

Gibson, G. J., Streftaris, G., and Thong, D. (2018). Comparison and assessment of epidemic models. *Statistical Science*, 33(1):19–33.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Gillespie, D. T. (2000). The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306.

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733.

Golightly, A., Bradley, E., Lowe, T., and Gillespie, C. S. (2019). Correlated pseudo-marginal schemes for time-discretised stochastic kinetic models. *Computational Statistics & Data Analysis*, 136:92–107.

Golightly, A. and Gillespie, C. S. (2013). Simulation of stochastic kinetic models. *In Silico Systems Biology*, pages 169–187.

Golightly, A. and Kypraios, T. (2018). Efficient SMC^2 schemes for stochastic kinetic models. *Statistics and Computing*, 28(6):1215–1230.

Golightly, A. and Sherlock, C. (2019). Efficient sampling of conditioned Markov jump processes. *Statistics and Computing*, 29:1149–1163.

Golightly, A., Wadkin, L. E., Whitaker, S. A., Baggaley, A. W., Parker, N. G., and Kypraios, T. (2023). Accelerating Bayesian inference for stochastic epidemic models using incidence data. *Statistics and Computing*, 33(6):134.

Golightly, A. and Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface focus*, 1(6):807–820.

Golightly, A. and Wilkinson, D. J. (2015). Bayesian inference for Markov jump processes with informative observations. *Statistical applications in genetics and molecular biology*, 14(2):169–188.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113.

Gottschling, S. and Meyer, S. (2006). An epidemic airborne disease caused by the oak processionary caterpillar. *Pediatric dermatology*, 23(1):64–66.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, 42(4):599–653.

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4(3):465 – 496.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45.

Kantas, N., Doucet, A., Singh, S. S., and Maciejowski, J. M. (2009). An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. *IFAC Proceedings Volumes*, 42(10):774–785.

Karppinen, S., Singh, S. S., and and, M. V. (2024). Conditional Particle Filters with Bridge Backward Sampling. *Journal of Computational and Graphical Statistics*, 33(2):364–378.

Kermack, W. O. and McKendrick, A. G. (1927). 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.

Kloeden, P. E. and Platen, E. (1992). *Introduction to Stochastic Time Discrete Approximation*, pages 305–337. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kloek, T. and van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: An application of integration by monte carlo. *Econometrica*, 46(1):1–19.

Komorowski, M., Finkenstädt, B., Harper, C. V., and Rand, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC bioinformatics*, 10:1–10.

Kurtz, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of applied Probability*, 7(1):49–58.

Kurtz, T. G. (1972). The relationship between stochastic and deterministic models for chemical reactions. *The Journal of Chemical Physics*, 57(7):2976–2978.

Kypraios, T., Neal, P., and Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. *Mathematical Biosciences*, 287:42–53.

Lewis, P. W. and Shedler, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413.

Lin, J. and Ludkovski, M. (2014). Sequential Bayesian inference in hidden Markov stochastic kinetic models with application to detection and response to seasonal epidemics. *Statistics and Computing*, 24:1047–1062.

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo methods in practice*, pages 197–223. Springer.

Lloyd-Smith, J. O. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PloS one*, 2(2):e180.

Maier, H., Spiegel, W., Kinaciyan, T., and Hönigsmann, H. (2004). Caterpillar dermatitis in two siblings due to the larvae of Thaumetopoea processionea L., the oak processionary caterpillar. *Dermatology*, 208(1):70–73.

Maier, H., Spiegel, W., Kinaciyan, T., Krehan, H., Cabaj, A., Schopf, A., and Hönigsmann, H. (2003). The oak processionary caterpillar as the cause of an

epidemic airborne disease: survey and analysis. *British Journal of Dermatology*, 149(5):990–997.

Mainprize, N. and Straw, N. (2021). Forestry Commision, Oak Processionary Moth (Thaumetopoea processionea) Contingency Plan. [Online; accessed 28-October-2021].

McKinley, T. J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2018). Approximate Bayesian Computation and Simulation-Based Inference for Complex Stochastic Epidemic Models. *Statistical Science*, 33(1):4 – 18.

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.

Minas, G. and Rand, D. A. (2017). Long-time analytic approximation of large stochastic oscillators: Simulation, analysis and inference. *PLoS Computational Biology*, 13(7):e1005676.

Minter, A. and Retkute, R. (2019). Approximate Bayesian Computation for infectious disease modelling. *Epidemics*, 29:100368.

Murray, L. M., Lee, A., and Jacob, P. E. (2016). Parallel resampling in the particle filter. *Journal of Computational and Graphical Statistics*, 25(3):789–805.

Oksendal, B. (2013). *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media.

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.

Plummer, M., Best, N., Cowles, K., Vines, K., et al. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11.

Pooley, C. M., Bishop, S. C., and Marion, G. (2015). Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes. *Journal of The Royal Society Interface*, 12(107):20150225.

Rahlenbeck, S. and Utikal, J. (2015). The oak processionary moth: a new health hazard? *British Journal of General Practice*, 65(637):435–436.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351 – 367.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20 – 71.

Ross, J., Pagendam, D., and Pollett, P. (2009). On parameter estimation in population models II: multi-dimensional processes and transient dynamics. *Theoretical Population Biology*, 75(2-3):123–132.

Sandmann, W. (2009). Streamlined formulation of adaptive explicit-implicit tau-leaping with automatic tau selection. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 1104–1112. IEEE.

Särkkä, S. and Solin, A. (2019). *Applied stochastic differential equations*, volume 10. Cambridge University Press.

Schmon, S. M., Deligiannidis, G., Doucet, A., and Pitt, M. K. (2021). Large-sample asymptotics of the pseudo-marginal method. *Biometrika*, 108(1):37–51.

Schmon, S. M. and Gagnon, P. (2022). Optimal scaling of random walk Metropolis

algorithms using Bayesian large-sample asymptotics. *Statistics and Computing*, 32(2):28.

Sherlock, C., Thiery, A. H., Roberts, G. O., and Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics*, 43(1):238 – 275.

Spannaus, A., Papamarkou, T., Erwin, S., and Christian, J. B. (2022). Inferring the spread of COVID-19: the role of time-varying reporting rate in epidemiological modelling. *Scientific Reports*, 12(1):10761.

Stathopoulos, V. and Girolami, M. A. (2013). Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110541.

Stockdale, J. E., Kypraios, T., and O'Neill, P. D. (2021). Pair-based likelihood approximations for stochastic epidemic models. *Biostatistics*, 22(3):575–597.

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on signal Processing*, 50(2):281–289.

Suprunenko, Y. F., Castle, M. D., Webb, C. R., Branson, J., Hoppit, A., and Gilligan, C. A. (2021). Estimating expansion of the range of oak processionary moth (Thaumetopoea processionea) in the UK from 2006 to 2019.

Swallow, B., Birrell, P., Blake, J., Burgman, M., Challenor, P., Coffeng, L. E., Dawid, P., De Angelis, D., Goldstein, M., Hemming, V., et al. (2022). Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling. *Epidemics*, 38:100547.

Wadkin, L. E., Branson, J., Hoppit, A., Parker, N. G., Golightly, A., and Baggaley, A. W. (2022). Inference for epidemic models with time-varying infection rates: Tracking the dynamics of oak processionary moth in the UK. *Ecology and Evolution*, 12(5):e8871.

West, M. and Harrison, J. (2006). *Bayesian Forecasting and Dynamic Models*. Springer Science & Business Media.

Whiteley, N. and Rimella, L. (2021). Inference in stochastic epidemic models via multinomial approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1305. PMLR.

Wilkinson, D. J. (2018). *Stochastic modelling for systems biology*. CRC press.