Bayesian Inference and Model Selection for Partially-Observed, Continuous-Time, Stochastic Epidemic Models

James Nicholas Walker

Thesis submitted for the degree of

Doctorate of Philosophy

in

Applied Mathematics and Statistics

at

The University of Adelaide

(Faculty of Engineering, Computer and Mathematical Sciences)

School of Mathematical Sciences

December, 2019

Contents

List of Figures

Abstract

Emerging infectious diseases are an ongoing threat to the health of populations around the world. In response, countries such as the USA, UK and Australia, have outlined data collection protocols to surveil these novel diseases. One of the aims of these data collection protocols is to characterise the disease in terms of transmissibility and clinical severity in order to inform an appropriate public health response. This kind of data collection protocol is yet to be enacted in Australia, but such a protocol is likely to be tested during a seasonal influenza (flu) outbreak in the next few years. However, it is important that methods for characterising these diseases are ready and well understood for when an epidemic disease emerges. The epidemic may only be characterised well if its dynamics are well described (by a model) and are accurately quantified (by precisely inferred model parameters).

This thesis models epidemics and the data collection process as partially-observed continuous-time Markov chains and aims to choose between models and infer parameters using early outbreak data. It develops Bayesian methods to infer epidemic parameters from data on multiple small outbreaks, and outbreaks in a population of households. An exploratory analysis is conducted to assess the accuracy and precision of parameter estimates under different epidemic surveillance schemes, different models and different kinds of model misspecification. It describes a novel Bayesian model selection method and employs it to infer two important characteristics for understanding emerging epidemics: the shape of the infectious period distribution; and, the time of infectiousness relative to symptom onset. Lastly, this thesis outlines a method for jointly inferring model parameters and selecting between epidemic models. This new method is compared with an existing method on two epidemic models and is applied to a difficult model selection problem.

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I also give permission for the digital version of my thesis to be made available on the web, via the Universitys digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

SIGNED: .... DATE: ....

Acknowledgements

I would like to thank my supervisors, Joshua Ross and Andrew Black, for the time, effort and advice they lent to help this thesis come to fruition. I sincerely thank Dr. Michael and Heather Dunne for the generous funding provided through the Dr Michael and Heather Dunne Applied Mathematics Supplementary Scholarship. I also acknowledge the support I have received through the Australian Government Research Training Program Scholarship. This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide and the help provided by the Phoenix Support Team (particularly Fabien Voisin). I would also like to thank Philip O'Neil, Theodore Kypraios and Chris Drovandi for listening to my ideas, giving advice and for their stimulating research conversations. I thank my friends and family for their support throughout my PhD candidature. Finally, I thank Peter Hochs for providing me with the sole quote of this thesis: "*A quote makes a thesis appear sophisticated*".

Publications

The work described in Chapter 3 (Section 3.1) and Chapter 5 of this thesis was done in collaboration with Andrew Black and Joshua Ross. This has been published as:

J. N. Walker, J. V. Ross, and A. J. Black, Inferenceof epidemiological parameters from household stratified data, PLOS ONE, vol. 12, pp. 121, 10 2017.

J. N. Walker, A. J. Black, and J. V. Ross, Bayesian model discrimination for partially-observed epidemic models, Mathematical Biosciences, vol. 317, p. 108266, 2019.

Chapter 1

Introduction

Since the 2009 swine flu' outbreak, countries around the world, such as the USA, UK and Australia, have outlined data collection protocols, sometimes referred to as First Few Hundred studies (FF100), for emerging infectious diseases [1–6]. When a novel infectious disease emerges these intensive data collection protocols may be enacted; in Australia, FF100 study protocol involves surveilling contacts of symptomatic individuals to obtain the time of symptom onset of the first few hundred cases at a daily resolution. These studies are not resourced to surveil every contact of every symptomatic individual, so only contact from the same household, workplace or school may be surveilled. Contacts within these subpopulations are easily surveilled, and, are most likely to become infected. Although an FF100 study is yet to take place in Australia, it is pertinent that methods are ready for the analysis of these data in case of a pandemic outbreak. Further, tests of FF100 protocol are likely to take place during a seasonal flu' outbreak in the next few years, so this thesis provides methods that can be readily applied during these studies. It develops and analyses methods for choosing between epidemiological models and inferring their parameters using simulated FF100 type data, so that these methods may be well understood and applied during the emergence of an infectious disease.

Using FF100 study data to characterise the spread of novel diseases both retrospectively and during an outbreak allow authorities to understand, control and respond in an informed way. One aim of FF100 studies is to characterise the disease in terms of clinical severity and transmissibility [7–9]. Clinical severity can be measured in a variety of ways, possibly involving specific symptoms, though it has been suggested that it may be appropriate to consider visibility as a measure of severity [3, 8, 9]. Transmissibility can be considered in terms of the average number of exposures caused by an infectious individual (the reproduction number) or the exponential rate at which infections occur in the population (exponential growth rate) [10–13]. Once a model for the epidemic is chosen, these measures of transmissibility can be used to estimate the final size and peak time of an outbreak.

Biological and epidemiological processes can have highly complex behaviour and in most cases we do not observe all of their underlying dynamics. For example, FF100 data only contains daily cases of symptom onset, so we do not observe the exact times where individuals contract the disease, become infectious, or, recover. As our observations may depend on dynamics we do not see, it can be difficult to infer properties of this partially-observed process. To add complexity to the problem, there may be several theories about the behaviour of the epidemic process and it may not be obvious as to which of these models is most appropriate given some set of observations. For example in epidemiology, it is widely recognised that the variance of the infectious period is critical to understanding historical disease incidence, and also for accurate evaluation of control measures for public health use [14, 15]. Similarly, it has been identified that the relative timing of symptoms and infectiousness is a key determinant of ability to control an outbreak [16]. Investigating these features is most easily done by encoding the structure of the model in a certain way rather than parametrising it, hence to discriminate different possibilities we need to select between competing models in an informed way.

For emerging diseases there is a reasonable chance that an outbreak will fade-out immediately, or, that there is a delay before the epidemic goes into a phase of exponential growth, so modelling these stochastic features is important. Stochastic modelling is particularly well suited to outbreaks in small populations (such as households), and is therefore appropriate for modelling epidemics in subpopulations which give rise to FF100 study data. Continuous-time Markov chains (CTMCs) are a variety of stochastic models that have seen wide use in mathematical epidemiology [9,12,17,18], partially due to their convenient mathematical properties (discussed in Chapter 2), and, partially due to their ability to model random, discrete events in continuous-time. Hence, this thesis uses CTMCs to model epidemic outbreaks. These CTMCs are defined by a set of possible states and transition rates between the states, which are a function of the state of the process and some model parameters.

Given an epidemic model and data, a disease can be characterised by inferring the models parameters. In the early stages of an epidemic there may be little data available, but there may be prior knowledge of dynamics of the infectious disease; for example, for influenza it may be reasonable to assume the average infectious period is less than a week. Bayesian statistics is a framework in which parameters have a distribution which changes as data are obtained. Bayesian inference incorporates prior knowledge of a process, via a *prior distribution*, and data, via a *likelihood function*, to obtain a *posterior distribution*. The prior distribution is chosen based on assumptions about parameters, whereas the likelihood function is defined as the probability density of observing data given a parameter set. Given limited data and reasonable prior knowledge in the emerging stages of a disease, a Bayesian framework is well suited. Hence, this thesis considers inference of parameters of epidemic models in a Bayesian framework.

As many events in the underlying epidemics are unobserved the epidemic models and observed data are considered as partially-observed CTMCs, in which the underlying CTMC process is related to data via an observation process. Unfortunately, while expressions can be written down for the likelihood function of partially-observed CTMCs, these are infeasible to compute in all but the simplest cases; they require evaluation of a matrix exponential [19,20] that in some cases may even be too large to fit in computer memory. As such, we build on and utilise Bayesian inference methods that circumvent pointwise calculation of the likelihood. These include: data-augmented Markov chain Monte Carlo (DA-MCMC) [17,21], which augments observed data with the transitions which are unobserved and imputes these; particle filters [22,23], which are methods based around simulating the underlying epidemic; and, sequential Monte Carlo squared (SMC-squared) [24–26], which is a method that simulates the underlying epidemic in important regions of the parameter space. Although these methods do not rely on direct calculation of the likelihood function, they target quantities of interest without introducing any error. These models and methods for inferring parameters of epidemic models are discussed in Chapter 2.

Inference methods need to be adapted to suit FF100 study data and in particular, as data are stratified by subpopulations, they need to incorporate models that consider the epidemic in a population of subpopulations (household epidemic models) or as multiple independent outbreaks in subpopulations. Inference for household epidemic models has been considered in the literature previously. For example, [27] considers outbreaks of measles in households, however, this assumed that recovery times were observed (which is typically not the case in FF100 studies), and only considered households in which only a single secondary infection is possible. Inference for models with two levels of mixing, such as household models, with data on final epidemic size has also been considered [28, 29]. The final epidemic size data are typically unavailable in the earliest stages of emerging diseases, further, FF100 study data contains temporal information which is potentially useful for inference. Bayesian inference for emerging infectious diseases in a population of farms was considered in [30]; however this relied on deterministic modelling of outbreaks within farms, which is a sensible modelling choice when there are many susceptible animals in close contact, but is less valid for modelling outbreaks in households containing few individuals. Parameters of a model of financial crises analogous to a household epidemic model were inferred via DA-MCMC in [31]; however, this assumed that the recovery rate was known, which is typically not the case for emerging infectious diseases. In Chapter 3 we present computationally-efficient implementations of DA-MCMC for epidemics in a population with household structure and apply them to simulated FF100 type data, some of which we have published [6]. As a FF100 study has not yet been enacted in Australia, we conduct simulation studies, which allows us to generate data sets for which we know the parameter values and validate results by checking whether inferred parameters correspond to the input values. One limitation of DA-MCMC is that incorporating new data requires inference to begin anew, whereas sequential methods, such as particle filters allow current estimates to be updated sequentially as new data are obtained. We describe a sequential version of DA-MCMC which can allow for more efficient mixing than standard DA-MCMC. This gives solutions to inference problems where mixing using DA-MCMC alone is prohibitively slow, for example, inferring the shape of an infectious period distribution.

FF100 studies are resourced to collect data on only the first few hundred symptomatic individuals and their contacts [3]. Given limited resources to surveil a population in the emerging stages of a disease, a decision needs to be made as to how the population should be surveilled to infer measures of transmissibility most accurately. Chapter 4 conducts an exploratory analysis to compare the accuracy and precision of inferred parameters under various surveillance schemes. That is, we infer epidemiological parameters assuming that infections are observed in different kinds of subpopulations: many small households, few schools or workplaces, or, a mix. This kind of analysis has not been considered before in the context of FF100 studies and is an important step in being prepared for an emerging infectious disease. We test the robustness of estimates from each of these surveillance schemes under model misspecification by performing inference based on a model that did not generate the data. Through a simulation study, we assess how different kinds of model misspecification on outbreaks in subpopulations bias estimates. In particular we consider misspecification of the exposed and infectious period distributions. Model misspecification has previously been considered in the context of incorrectly assuming data came from an SIR model, where it is actually simulated from an SI model [32], whereas here we consider SEIR and SIR models with different infectious and exposed period distributions. Biases with respect to the basic reproduction number under assumptions of misspecified mixing structure has been previously investigated [33]. In comparison, Chapter 4 investigates bias with misspecification of the exposed and infectious period distributions on both the reproduction number and the exponential growth rate. We find that under the true model, surveilling many small households is the optimal strategy, however surveilling a mix of small and large subpopulations was the most robust under the kinds of model misspecification considered in this thesis. However, model misspecification did lead to bias, and competing sources of bias makes it difficult to determine the best strategy in general. For example, the reproduction number is negatively biased for surveillance schemes where early fade out in subpopulations is likely, but positive bias can be introduced if the infectious period distribution is assumed to be exponential when it is actually Erlang. Further, the number of samples from final size distributions and the length of temporal data is likely to effect the bias in estimates. The presence of bias under misspecification, and, the presence of confounding sources of bias makes it difficult to choose one optimal scenario in general. Choosing a surveillance scenario that protects against bias highlights the need to effectively choose between competing epidemic models. The ability to both choose an appropriate model and infer model parameters will allow for an optimal surveillance scheme to be chosen effectively and robustly.

Selecting a model is typically done by comparing measures based upon the maximum value of the likelihood function, such as Akaike information criterion (AIC) [34] and Bayesian information criterion (BIC) [35]; once a model is chosen, parameters are fit to only that model. This is at odds with the Bayesian paradigm, which dictates that the space of models should have an associated posterior distribution. Bayesian model selection is a Bayesian method for choosing between models in a way that incorporates prior knowledge of the epidemic process; this is naturally the best approach if data are not highly informative but we have prior knowledge of some model parameters. This approach compares models in terms of their *posterior model probability*, that is, the probability that model was the true model given the data. The outputs from Bayesian model selection are useful for quantifying uncertainty in a way that incorporates uncertainty in the model choice, for example, posterior distribution of the reproduction number can be calculated under each model and these can be combined by weighting these distributions by their posterior model probabilities and adding them (this is known as Bayesian model averaging) [36].

Methods for performing Bayesian model selection include importance sampling approaches [23, 37], Reversible Jump MCMC (RJ-MCMC) [38] and SMC-squared [24]. Important sampling approaches typically involve sampling from the parameter space and evaluating the likelihood function to estimate the probability that the data was generated under each model (the model evidence). The model evidences can then be weighted and normalised to obtain posterior model probabilities. This approach can be effective for selecting between competing models but requires a sensible way of evaluating or estimating the likelihood function. RJ-MCMC involves proposing moves within a parameter space and between parameter spaces of different models; each of these proposals requires an evaluation of the likelihood function. The proposals between parameter spaces can be difficult to implement effectively if the parameter spaces are dissimilar, for example, the dimensions of the parameter spaces may be completely different. The method becomes even less effective if the likelihood function is computationally intractable and hence is ill-suited for our needs. The SMC-squared approach is attractive as it jointly estimates the posterior distribution and gives estimates of model evidence, however, the stochasticity in model evidence estimates can make effectively choosing between models impossible based upon a single SMC-squared run [25]. Chapter 5 describes an importance sampling based method for choosing between partially-observed epidemic models via an efficient particle filter and importance sampling scheme [39]. The novel feature of this method is that it uses efficient simulations that are constrained to always match observed data to estimate the likelihood function. Further, the method makes use of a stopping criterion to ensure accuracy of model selection. We apply this approach to two important problems for understanding emerging diseases: inferring the time of symptom onset relative to the time of infectiousness; and, inferring the shape of the infectious period distribution. We find that our approach is effective for these model selection problems and that FF100 type data are informative enough to effectively distinguish between these models.

Inferring both parameters and choosing an appropriate model is important for effectively characterising an infectious disease; SMC-squared simultaneously does both [24–26]. Further, it is a sequential method, so rather than beginning inference anew when new data are available, the algorithm can update the current parameter and model evidence estimates. A main drawback of using SMC-squared for model selection is that there are no error bounds on model evidence and choosing sensible inputs can be difficult. Estimating the error in model evidence estimates may require multiple runs of SMC-squared, which can be computationally taxing. Once the error is estimated, it may be too large to effectively select between models, so the process will need to begin again with different input values. Chapter 6 describes a novel kind of SMC-squared which allows uncertainty in the model evidence to be quantified, and, allows error in model evidence to be reduced. These two features allow for effective model selection to occur after a single run of SMC-squared. We provide a comparison with a standard SMC-squared algorithm on two standard epidemic models and show that the new method outperforms in terms of runtime, the accuracy of parameter estimates, and, the accuracy of model evidence estimates. The method is then applied to a difficult inference and model selection problem: inferring the time of infection relative to symptom onset from multiple outbreak data in large subpopulations.

Chapter 2

Background Material

2.1 Markov Chains

Markov chains are a stochastic process that have a property known as the *Markov property*, in essence this means that the future of the process depends only on the current state of the process, not states from further in the past. This property allows probability expressions to be simplified by ignoring dependencies across multiple time points. A common alternative to Markov chain models are differential equation models. These models are widely used because they are efficient to solve, however, models on discrete state spaces are more realistic to the true process and allow data to be related to the model in a more intuitive way. Further, the stochastic effects of infectious processes are important in the emerging stages of an epidemic; stochasticity allows delays before the epidemic reaches an exponential growth phase and allows the epidemic to fade out. Once the disease is established these stochastic effects are less important. As this thesis is concerned with inference for emerging epidemics it considers only stochastic models.

Section 2.1.1 discusses some important definitions and properties of Markov chains, Section 2.1.2 defines partially-observed Markov chains and outlines some theoretical considerations for inference and Section 2.1.3 outlines how to simulate from Markov chains via the Doob-Gillespie algorithm.

2.1.1 Continuous-time Markov chain theory

This section defines continuous-time Markov chains (CTMCs), introduces the kinds of Markov chains that will be considered in this thesis and discusses some of their key properties. A CTMC is a stochastic process on the positive real numbers, $\mathbb{R}^+$, that satisfies the Markov property. That is, a stochastic process X_t defined on some state space $\mathcal{S}$, for $t \in \mathbb{R}^+$ is a CTMC if it satisfies

$$\Pr(X_t \in A | X_s, X_u) = \Pr(X_t \in A | X_s)$$

for all $s, u, t \in \mathbb{R}^+$ with $u \leq s \leq t$ and $A \subseteq \mathcal{S}$. Throughout this thesis, inference is based on stochastic epidemic models which are CTMCs defined on finite state spaces. A CTMC, X_t, on a finite state space, $\mathcal{S}$, can be described in terms of an infinitesimal transition rate matrix, also referred to as a "Q-matrix" given by

$$[Q(t)]_{i,j} = \begin{cases} \lim_{h \to 0^+} \dfrac{\Pr(X_{t+h} = j | X_t = i)}{h} & \text{for } i \neq j \\ -\sum_{k \neq i} [Q(t)]_{i,k} & \text{otherwise.} \end{cases} \tag{2.1}$$

The entries of $Q(t)$ are the "instantaneous transition rates" from one state to another, that is $[Q(t)]_{i,j}$ is the rate at which the CTMC jumps to state j given that it is in state i at time t. A Markov chain is called *time homogeneous* if it satisfies

$$\Pr(X_t = j | X_s = i) = \Pr(X_{t-s} = j | X_0 = i),$$

for all $s \leq t$. A fundamental property of time-homogeneous CTMCs is that the times between transitions are exponentially distributed, so "transition rate" actually refers to the exponential rate at which events occur and $Q(t)$ is constant with respect to time, that is, $Q(t) = Q$ for all t [40].

The transition function, $P(t)$, of a CTMC is defined as

$$[P(t)]_{i,j} = \Pr(X_t = j | X_0 = i).$$

The Kolmogorov forward equation relates the Q-matrix to the transition function of a CTMC. It satisfies

$$\frac{dP(t)}{dt} = P(t)Q(t). \tag{2.2}$$

For a time homogeneous CTMC, defined on a finite state space, with infinitesimal transition rate matrix Q, Equation (2.2) has solution

$$P(t) = e^{Qt};$$

this is a matrix exponential operation. The matrix exponential function on some finite dimensional matrix A denoted e^A is given by

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!},$$

where A^n is a matrix product. Let p_t be the probability mass function of the state of the CTMC, $\{X_t\}_{t\in\mathbb{R}^+}$, at time t, such that p_t is a vector with entries given by

$$[p_t]_i = \Pr(X_t = i).$$

Given an initial distribution of the CTMC, p_0, one can express p_t as

$$p_t = p_0 e^{Qt}. \tag{2.3}$$

Matrix exponential operations may be computed using methods such as the software package Expokit for MATLAB [41]. Expokit makes these calculations efficient by calculating the vector $p_0 e^{Qt}$ by the use of Krylov subspace projection methods, rather then calculating the matrix e^{Qt} and premultiplying the answer by p_0. However, this can still be computationally infeasible to evaluate for large Q-matrices [19]. Hence the distribution p_t can be computationally infeasible to evaluate for CTMCs with large state spaces, such as models of epidemics in large populations.

If the times at which transitions occur and the corresponding transition types, or events, are observed we can calculate the probability of the trajectory of X_t as a product of event probabilities and inter-arrival time probabilities,

$$\begin{aligned}\Pr\left(X_{t_1} = j, X_{(t_0,t_1)} = i | X_{t_0} = i\right) &= \frac{[Q]_{i,j}}{[Q]_{i,i}}[Q]_{i,i}e^{[Q]_{i,i}(t_1-t_0)},\\ &= [Q]_{i,j}e^{[Q]_{i,i}(t_1-t_0)}. \end{aligned} \tag{2.4}$$

These kinds of probabilities can allow inference to be performed without the need for matrix exponential calculations; this is explained in greater detail in Section 2.3.3.

2.1.2 Partially-observed Markov chains

Partially-observed Markov chains are models constructed from an underlying Markov chain, X_t, and an observation process, Y_t, where observations are conditionally-independent given the underlying Markov chain; such a process is illustrated in Figure 2.1. For example, if an epidemic were to spread according to a continuous-time Markov chain, we may observe events related to symptom onset at a daily resolution, but we may not observe recovery events; as we do not have perfect observation of the Markov chain this is considered a partially-observed Markov chain. In this thesis we consider processes where the observation depends on the hidden process through how the underlying state changed since the last observation, that is, $\Pr(Y_t|X_{[0,t]}, Y_{1:t-1}) = \Pr(Y_t|X_{t-1}, X_t)$. Given a set of observations $y = (y_0, \ldots, y_T)$ from a partially-observed Markov chain, the distribution of an observation given states of the underlying process, $\Pr(y_t|X_t, X_{t-1})$, the distribution of the initial state $\Pr(X_0)$ and a Q-matrix which depends on model parameters, Q_θ, we may calculate the likelihood function of a partially-observed Markov chain as

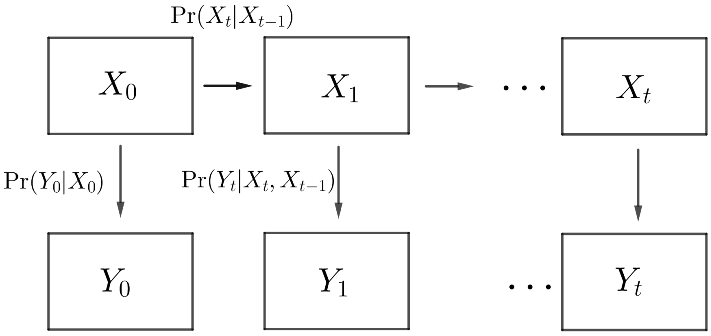

Figure 2.1: Illustration of a partially-observed Markov chain with observation process Y_t and hidden process X_t.

$$\begin{aligned}\Pr(y|\theta) &= \sum_{(i_0,\ldots,i_T)\in\mathcal{S}^{n+1}} \Pr(y_0|X_0 = i_0)P(X_0 = i_0)\prod_{j=1}^{T}\Pr(y_j|X_j = i_j, X_{j-1} = i_{j-1}) \\ &\qquad \times \Pr(X_j = i_j, |X_{j-1} = i_{j-1}) \\ &= \sum_{(i_0,\ldots,i_T)\in\mathcal{S}^{n+1}} \Pr(y_0|X_0 = i_0)\Pr(X_0 = i_0)\prod_{j=1}^{T}\Pr(y_j|X_j = i_j, X_{j-1} = i_{j-1})) \\ &\qquad \times \left[e^{Q_\theta}\right]_{i_{j-1},i_j},\end{aligned}$$

where $\mathcal{S}$ is the state space of X_t. Note that this formula has been described only for observations that are made at times $1, 2, \ldots, T$, and, throughout this thesis we only consider processes where observations are made at constant time increments. However, more generally, if observations are made at irregular times, where the observation times are independent of X_t, the only difference in the likelihood is that the matrix exponential is taken to the power of the time between observations. That is, $\left[e^{Q_\theta}\right]_{i_{j-1},i_j}$ is replaced by $\left[e^{Q_\theta(t_j - t_{j-1})}\right]_{i_{j-1},i_j}$, where t_j is the time of the jth transition. Although one can write an analytical expression for the likelihood function, for all but small state space models it is computationally intractable to calculate directly. This is because large state spaces lead to computationally intractable matrix exponential calculations, as well as the need for a sum over $|\mathcal{S}|^{T+1}$ terms. This motivates the need for methods that estimate the likelihood or avoid calculation of the full likelihood, such as particle-marginal MCMC and data-augmented MCMC, respectively.

2.1.3 Doob-Gillespie algorithm

The Doob-Gillespie algorithm is an algorithm for simulating CTMCs [42]. The algorithm uses the fact that inter-arrival times of events are exponentially distributed according to the sum of transition rates, and each transition occurs with a probability that is proportionate to the rate associated with that transition. An efficient implementation of the Doob-Gillespie algorithm is given in Algorithm 1. The Doob-Gillespie algorithm requires some kind of stopping condition; in this thesis stopping conditions include: the process hits an absorbing state, the process hits some time threshold, or the process hits a given set of states.

```
Initialization:
Generate an initial state, $x$, according to $p(x_0)$;
Set time $t = 0$;
Iterations:
while some condition holds do
    Calculate a vector of rates, $r_x$, such that entries correspond to non-zero
      transition rates from state $x$;
    Calculate the cumulative sum vector of $r_x$, $R_x$, let $n$ denote the length of $R_x$;
    Time change:
    Sample an Exponential($[R_x]_n$) distributed time increment, $\tau$, via inverse
    transform sampling. That is, take $u_1 \sim$ Uniform$(0, 1)$ and set
    $\tau = -\frac{1}{[R_x]_n}\log(1 - u_1)$;
    Set new time $t = t + \tau$;
    State Change:
    Sample a random variable to decide the new state, $u_2 \sim$ Uniform$(0, 1)$;
    Find the index, $i$, of the first entry in $R_x$ such that $R_x > u_2[R_x]_n$;
    Set new state $x = f(x, i)$, where $f(x, i)$ is a function that maps index, $i$, from
      $r_x$ to the state space ;
end
```

Algorithm 1: Doob-Gillespie algorithm

2.2 Stochastic Epidemic Models

Throughout this thesis we consider SIR-type and SEIR-type models, described in Sections 2.2.1 and 2.2.2 respectively. These models are compartmental models that consider every individual in a population to be either susceptible (S), exposed (E), infectious (I) or recovered (R). Susceptible individuals may become exposed to the disease if they make "effective contact" (contact of the kind that allows transmission to occur) with an infectious individual. For an individual in a SIR-type model an exposed individual instantaneously becomes infectious, while for a SEIR-type model the exposed individual becomes infectious after some random period of time. Infectious individuals recover after a random period of time and remain immune to the disease for the rest of the epidemic. Throughout this thesis we assume that the exposed and infectious period is Erlang distributed. The assumptions about how "effective contact" occur are important for providing realistic transmission dynamics; in this thesis we concentrate on homogeneous mixing of individuals and household mixing processes, the latter of which is described in Section 2.2.3.

For any model, the initialisation is important for modelling and inference. In this thesis models are either initialised by seeding an infectious individual in the population at a Uniform$(0, 1)$ distributed time, seeding an infectious individual at time 0 or generating an initial state at time 0 via simulation methods.

2.2.1 Homogeneous SI(n)R model

The SIR model is one of the simplest and most used epidemic models. It models infectious diseases where individuals can spread the disease soon after being exposed and are immune to the disease once no longer infectious. More generally we describe a SI(n)R model here, in which the infectious period is Erlang-n distributed; where the standard SIR model is equivalent to the SI(1)R model. We describe two kinds of SI(n)R model: the standard SI(n)R model (sometimes referred to as the lumped SI(n)R model); and, the labelled SI(n)R model. The standard model groups individuals into compartments whereas the labelled model distinguishes between individuals, which allows for a more natural description of transitions in terms of infectious periods, rather than recovery times. However, this labelled representation generally increases the size of the state space. The utility of the labelled model is discussed and shown in Section 2.3.3.

The standard SI(n)R model in a population of N individuals tracks the number of individuals who are classified as either susceptible (S), infectious phase k for $k = 1, \ldots, n$ (I^k) or recovered (R). Note that the infectious phases are simply used to create a CTMC with an Erlang-n infectious period, the phase need not have a physical interpretation. The state of the population at time t, $X_t = (S_t, I_t^1, \ldots, I_t^n, R_t)$, is a vector giving the number of individuals in each of the classes at time t. As $S_t + \sum_{k=1}^n I_t^k + R_t = N$, the state can be simplified to $X_t = (S_t, I_t^1, \ldots, I_t^n)$. The state space of the standard SI(n)R model is given by

$$\mathcal{S} = \left\{ X_t \in \{0, \ldots, N\}^{n+1} : S_t + \sum_{k=1}^{n} I_t^k \leq N \right\}.$$

The SI(n)R model has three kinds of transitions, being transmission, phase change and recovery which are governed by two parameters, the effective contact rate, β, and the recovery rate, γ. Infectious individuals each make effective contact with the other $N-1$ individuals in the population at rate β, however infections only occur if the infectious individuals make effective contact with a susceptible individual. Hence the instantaneous infection rate in the population at time t is $\frac{\beta S_t \sum_{k=1}^n I_t^k}{N-1}$. We define γ such that $1/\gamma$ is the mean infectious period, hence the rate at which phase changes occur for infectious individuals is γn. Therefore the instantaneous phase change rate in the population at time t is $\gamma n I_t^k$, for $k = 1, \ldots, n-1$, and the instantaneous recovery rate at time t is $\gamma n I_t^n$.

The labelled SI(n)R model in a population of N individuals tracks the state of each individual as susceptible (s), infectious (i), or recovered (r). The state of the population at time t, $X_t := \left(X_t^1, \ldots, X_t^N\right)$, is a vector giving the state of each individual at time t, that is, $X_t^k \in \{s, i, r\}$ for $k = 1, \ldots, N$. Let the total number of susceptible and infectious individuals in the population at time t be given by $S_t = \sum_{k=1}^N 1_{\left\{X_t^k = s\right\}}$ and $I_t = \sum_{k=1}^N 1_{\left\{X_t^k = i\right\}}$, respectively. The state space of the labelled SI(n)R model is

$$\mathcal{S} = \{s, i, r\}^N.$$

The labelled SI(n)R model has two kinds of transitions, being transmission and recovery which are governed by the effective contact rate, β, and the recovery rate, γ. The infection rate in the population is as before, however, as individuals are labelled at time t there are S_t individuals each of which become infected at rate $\frac{\beta I_t}{N-1}$. Each infectious individual then recovers with a Gamma$(n, n\gamma)$ infectious period. Note that the labelled representation of the SI(n)R model is a semi-Markov model, where infections occur as an inhomogeneous Poisson process and recovery times depend on the infection times.

2.2.2 Homogeneous SE(n_1)I(n_2)R model

The SEIR model is an extension of the SIR model where there is a delay between the individual contracting the disease and being able to spread the disease; individuals in this class are referred to as exposed (E). Here we describe a more general model which allows the exposed period to be Erlang-n_1 distributed, the SE(n_1)I(n_2)R model. In this section we describe only the labelled SE(n_1)I(n_2)R model, but note that the standard SE(n_1)I(n_2)R model can be expressed in a similar way to the standard SI(n)R model.

The labelled SE(n_1)I(n_2)R model, with a population of size N, tracks which individuals are susceptible (s), exposed (e), infectious (i) or recovered (r). This is a semi-Markov model as times of infections and recoveries depend on the times of exposure and infection respectively. The state at time t is given by $X_t^k \in \{s, e, i, r\}$ for $k = 1, \dots, N$. The number of infectious and recovered individuals are as in the labelled SI(n)R model, in addition we define the number of exposed individuals as $E_t := \sum_{k=1}^{N} 1_{\{X_t^k = e\}}$. The state space of the labelled SI(n_1)E(n_2)R model is

$$\mathcal{S} = \{s, e, i, r\}^N.$$

The labelled SE(n_1)I(n_2)R model is defined in terms of three transitions, being transmission (or exposure), infection and recovery which are governed by parameters for transmission rate, β, infectious rate, σ, and recovery rate, γ respectively. As in the labelled S(n_2)R model, susceptible individuals become exposed at time t according to instantaneous exposure rate $\frac{\beta I_t}{N-1}$ and infectious individuals have a Gamma$(n_2$,$n_2\gamma)$ infectious period. This model also has exposed individuals, which are exposed for a Gamma$(n_1$,$n_1\sigma)$ distributed time before becoming infectious. The latent and infectious periods are defined such that their expected values are $1/\sigma$ and $1/\gamma$, respectively.

Transmissibility, or how easily the disease spreads through a population, for the homogeneous SI(n)R and SE(n_1)I(n_2)R models is most commonly quantified in terms of the *basic reproduction number*, $R_0 = \beta/\gamma$. This is the expected number of secondary infections of a primary infected individual in a large population of susceptible individuals. Another measure of transmissibility is the *early growth rate*, which is the exponential rate at which the number of infectious individuals increases in the early stages of an epidemic, this is calculated as the solution to

$$r(r + \sigma n_1)^{n_1} - R_0\gamma(\sigma n_1)^{n_1}\left(1 - \left(\frac{r}{\gamma n_2} + 1\right)^{-n_2}\right) = 0, \tag{2.5}$$

as described in [15]. Note that the above equation holds for the SI(n)R model where $n_1 = 0$ and $n_2 = n$. The reproduction number is a quantity that indicates on average how many people will get infected, whereas the growth rate indicates how quickly the epidemic spreads.

2.2.3 Household models

Epidemic models with two levels of mixing are models where individuals make effective contact according to two different rates [43]. In this thesis we consider specific epidemics with two levels of mixing called household epidemic models. These models consider individuals grouped into M mutually exclusive households, where the kth household has some known size, N_k. Individuals make effective contact at a high rate within households and at a low rate between households. Household epidemic models can be SIR-type or SEIR-type; here we describe the general household model for either type of epidemic model. Household epidemic models have a much more complicated state space than their homogeneous counterparts, as each household acts as a homogeneous epidemic in which infection can also be imported from other households. The simplest way to describe household epidemic models is by considering a model with labelled households, that is, we consider an epidemic model which tracks the state of each household as opposed to a model which aggregates the total number of households of the same size which are in the same state. Let h_k be a vector which tracks the state of the kth household, so a state is given as a list of states of each household, $\mathbf{h} := \{h_1, \ldots, h_M\}$, and the state space is given by $\hat{\mathcal{S}} = \{\mathbf{h} : h_k \in \mathcal{S}_k, \text{ for } k = 1, \ldots, M\}$, where $\mathcal{S}_k$ is the state space of a household of size N_k. Here $\mathcal{S}_k$ is either the state space of a standard SI(n)R or SE(n_1)I(n_2)R model for a population of size N_k.

Within-household transmission, recovery and, for a SEIR-type model, infection, occur within a household as in the homogeneous model. Households interact via between-household transmission events, which are governed by a between-household effective contact parameter, α. Specifically, let α be the rate at which individuals make effective contact with other individuals outside of their household. Let I_t, s_t^k and i_t^k be the total number of infectious individuals in the population, the number of susceptibles in household k and the number of infectious individuals in household k at time t respectively. Note that between-household effective contact only occurs between members of different households, so household k has s_t^k individuals that the disease can be transmitted to, between-household transmission can occur from $I_t - i_t^k$ individuals, and susceptible individuals could make between-household contact with $\sum_{j \neq k} N_k$ individuals. Hence, the rate at which between-household transmission occurs and is transmitted to an individual in household k is $\frac{\alpha s_t^k (I_t - i_t^k)}{\sum_{j \neq k} N_k}$.

To quantify the overall transmissibility in household epidemic models we calculate the *household reproduction number*, R_* [43]. This is the expected number of households infected by a primary infectious household in a population of susceptible households; where a household is considered infectious while it contains at least one infectious or exposed individual and a household is considered susceptible if it contains only susceptible individuals. It is one of at least five reproductive numbers that might be used when assessing the controllability of a disease in a community of households [44–46]. We consider R_* in this thesis as it is relatively easy to calculate and interpret. Let $\{X_t\}_{t \in \mathbb{R}^+}$ be the Markov chain that describes the state of an individual household from the time of the first exposure. Let $I(k)$ be the function which returns the number of infectious individuals corresponding to state k. Then

$$R_* = E\left[\int_0^\infty \alpha I(X_t)\ dt\right],$$

where X_0, the initial state of the process, corresponds to a household with a single infectious individual and all other individuals susceptible [43,47]. This can be calculated by solving a system of linear equations depending on the parameters of the epidemic model; the household growth rate can be calculated similarly [48,49].

2.3 Bayesian Inference

This section outlines the statistical quantities and methods that will be considered throughout this thesis for estimating model parameters. Once model parameters are inferred, epidemiological parameters of interest such as reproduction numbers and early growth rate, may be inferred as functions of the model parameters. All inference methods considered in this thesis are Bayesian, that is, we consider parameters to have distributions, rather than a fixed value, to be inferred.

Given a set of observations, y, and model parameters, θ, it is natural to consider the *likelihood* function, $L(\theta) = p(y|\theta)$, the probability of observing data y given parameters θ. Finding the θ that maximises this function is a sensible target for inference. However, in a Bayesian paradigm the objective is reversed and the quantity of interest is the *posterior* distribution, $p(\theta|y)$, that is, the distribution of the parameters, θ, given observations, y. Not only is the argument that maximises the posterior distribution of interest, but the variance and other quantities can be analysed once a posterior distribution is inferred. By Bayes theorem the posterior distribution is given by

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)\ d\theta} \\ &= \frac{L(\theta)p(\theta)}{\int_\Theta L(\theta)p(\theta)\ d\theta}, \end{aligned} \tag{2.6}$$

where $p(y) = \int_\Theta L(\theta)p(\theta)d\theta$ is a normalising constant known as the *evidence*; the distribution $p(\theta)$ reflects our prior beliefs about θ in the absence of data, this is known as a *prior distribution.* Note that the prior distribution is not inferred, it is specified before inference takes place. However, the choice of prior distribution may have a substantial impact on inference, so it is important that it is chosen carefully. Priors are usually chosen based on one of three reasons: they are chosen based on existing knowledge of their values; they are chosen to be flat with wide support, or to be objective *Jeffreys priors*, for when there is not existing knowledge of their values; or, they are chosen to improve computational efficiency for inference. The latter is possibly a trade-off against the former two. For example, some priors (known as conjugate priors) allow for the posterior distribution to have a closed form, which circumvents the need for integration and simplifies computation; they may have a mean and variance informed by existing knowledge or may be chosen to be relatively uninformative.

The model evidence can be thought of as the likelihood of a given model, it is the probability that the data was generated from the assumed model, so it can be used to perform Bayesian model selection. That is, with a set of candidate models, $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k\}$, and observed data, y, the evidence of model i is $p(y|\mathcal{M}_i)$. So if one chooses a prior of the candidate models, $p(\mathcal{M}_i)$, the *posterior model probability* can be calculated as

$$p(\mathcal{M}_i|y) = \frac{p(y|\mathcal{M}_i)p(\mathcal{M}_i)}{\sum_{j=1}^{k} p(y|\mathcal{M}_j)p(\mathcal{M}_j)},$$

and is interpreted as the probability that the process is described by model $\mathcal{M}_i$ given the observed data and the prior distributions. Alternatively, models can also be compared by Bayes Factors, which are a ratio of model evidences; these tell us how many times more likely observations are under one model compared to another model [50]. Though, choosing a model based on Bayes Factors is equivalent to choosing based on posterior model probabilities with a uniform prior over candidate models. One benefit of Bayes factors is that they do not need to be recomputed when another candidate model is added to the set of potential models; though the recalculation of posterior model probabilities are simple if model evidence has already been computed.

In many cases the posterior distribution is impossible to calculate analytically, due to the difficulty in calculating the evidence and/or the likelihood, so sampling based methods are typically employed. The remainder of this section discusses methods that can be used to sample from posterior distributions. Sampling methods can also be useful in estimating the evidence, and hence, performing Bayesian model selection.

2.3.1 Markov chain Monte Carlo

A classic method for Bayesian inference is Markov chain Monte Carlo (MCMC); this is a class of algorithms which sample from a Markov chain with a stationary distribution equal to a distribution of interest [51, 52]. In the context of model fitting, the Markov chain samples over the parameter space of the model being fitted. Sampling from the Markov chain's stationary distribution allows for estimation of the posterior distribution [51,52]. A common MCMC method is the Metropolis-Hastings algorithm, this allows the posterior distribution to be approximated without needing to evaluate the evidence [51,52]. Pseudocode for a Metropolis-Hastings algorithm is shown in Algorithm 2. The Metropolis-Hastings algorithm requires a choice of proposal density $q(\theta'|\theta)$, a distribution that samples the next candidate parameter set, θ', given the current parameter set, θ. As long as $q(\theta'|\theta)$ ensures irreducibility of the Markov chain, convergence to the posterior distribution is assured. Here irreducibility means that it is possible for the Markov chain to hit any state from any initial state. The choice of prior, $p(\theta)$, and proposal density, $q(\theta'|\theta)$, influence the rate of convergence of the Markov chain to its stationary distribution [51,52].

```
Initialization:
Set a prior distribution $p(\theta)$;
Set some initial set of parameter values $\theta^{(0)}$;
Set $n = 0$;
Set the number of iterations to some large value, $K$;
Set a proposal density of the form $q(\theta'|\theta)$;
Iterations: while $n \leq K$ do
    sample a candidate $\theta'$ from $q(\theta'|\theta^{(n)})$;
    sample a uniform [0,1] variable, $u$;
    calculate $\alpha(\theta^{(n)}, \theta') = \frac{L(\theta')p(\theta')q(\theta^{(n)}|\theta')}{L(\theta^{(n)})p(\theta^{(n)})q(\theta'|\theta^{(n)})}$;
    if $u < \alpha(\theta^{(n)}, \theta')$ then
        $\theta^{(n+1)} = \theta'$;
    else
        $\theta^{(n+1)} = \theta^{(n)}$;
    end
    $n = n + 1$;
end
```

Algorithm 2: Metrolpolis-Hastings algorithm

Algorithm 2 generates a Markov chain with stationary distribution equal to the the posterior distribution from Equation (2.6). Hence, we need to decide when the Markov chain has reached stationarity. The number of iterations until the Markov chain is thought to reach stationarity, b, is referred to as *burn in*; this is often chosen graphically. Specifically, we plot $\theta^{(n)}$ for $n = 0, ..., K$ and choose b, such that there is no trend in $\theta^{(b:K)}$; this plot is known as a *trace plot* [51, 52]. Examples of trace plots are shown in Figure 2.2. The top left panel shows a clear upwards trend, so stationarity has not been reached. The top right panel shows a large jump at about 1000 iterations, so perhaps 1000 is an appropriate burnin, but more iterations are needed to make an informed choice. The bottom left panel shows that samples are moving around the same region for many iterations, this indicates that the stationary distribution may have been reached, so little burn in is required. The last panel shows very slow movement, or *mixing*, around the state space; many more samples may be needed to assess stationarity. An alternative to checking a single trace plot is to run the algorithm multiple times with various initial parameter values, $\theta^{(0)}$, and compare trace plots to choose b such that each of the trace plots show similar behaviour. Once

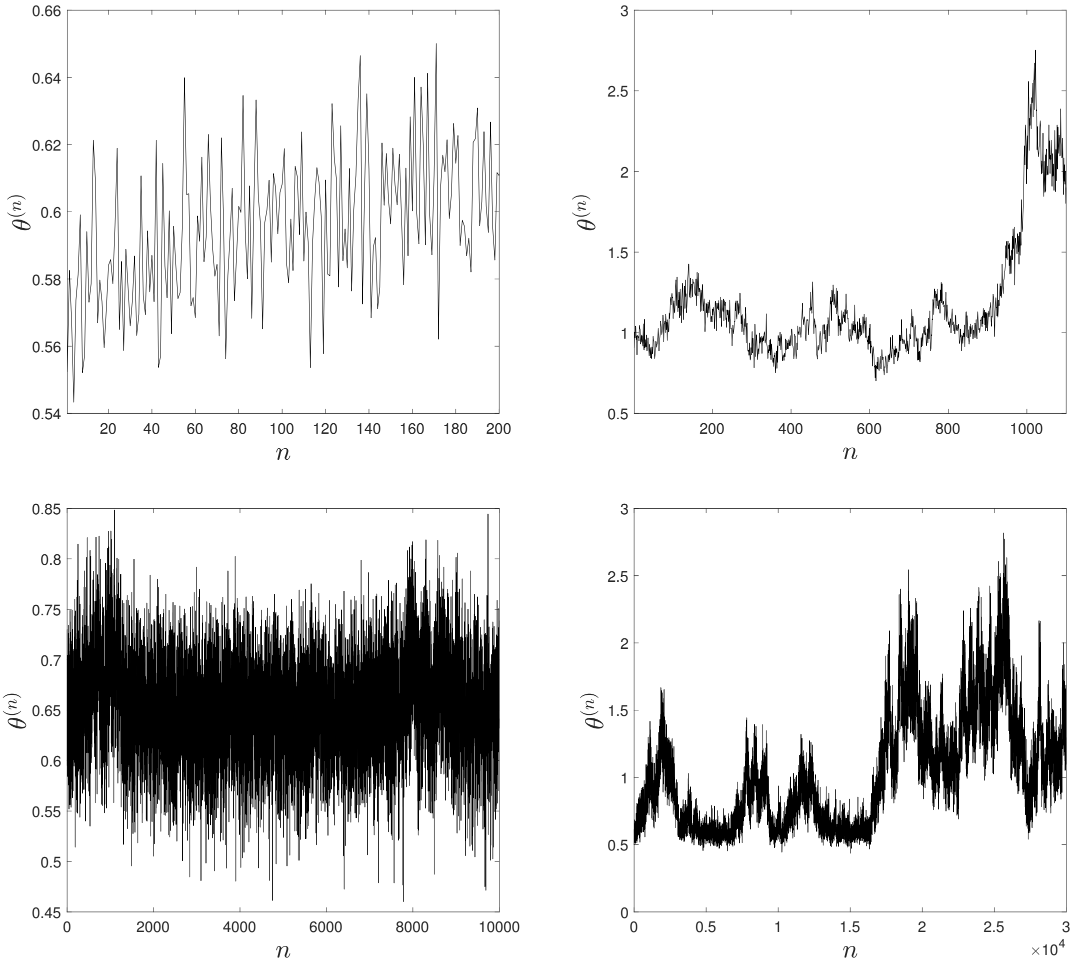

Figure 2.2: Trace plots from MCMC algorithms

the burn in has been chosen, samples from the Markov chain, $\theta^{(b:K)}$, can be used to calculate quantities of interest, such as expectations of functions of the parameters, which can be estimated by

$$E[f(\theta)] = \frac{1}{K-b+1}\sum_{i=b}^{K} f\left(\theta^{i}\right).$$

Alternatively, a kernel smoother can be applied to the samples to approximate the posterior distribution, $p(\theta|y)$; an example of this is given in Figure 2.3. In this thesis kernel densities are estimated by MATLAB functions ‘kde’, ‘kde2’ or ‘akde’ by Botev et al. [53].

The Metropolis-Hastings algorithm is simple and easy to implement, however, it assumes that the likelihood is computationally efficient to evaluate. In this thesis we consider data that arises from partially-observed Markov chains, so the likelihood can be written in terms of a matrix exponential calculation. However, these calculations become computationally infeasible for all but the simplest models with small state spaces. The remainder of this section addresses methods for Bayesian inference which

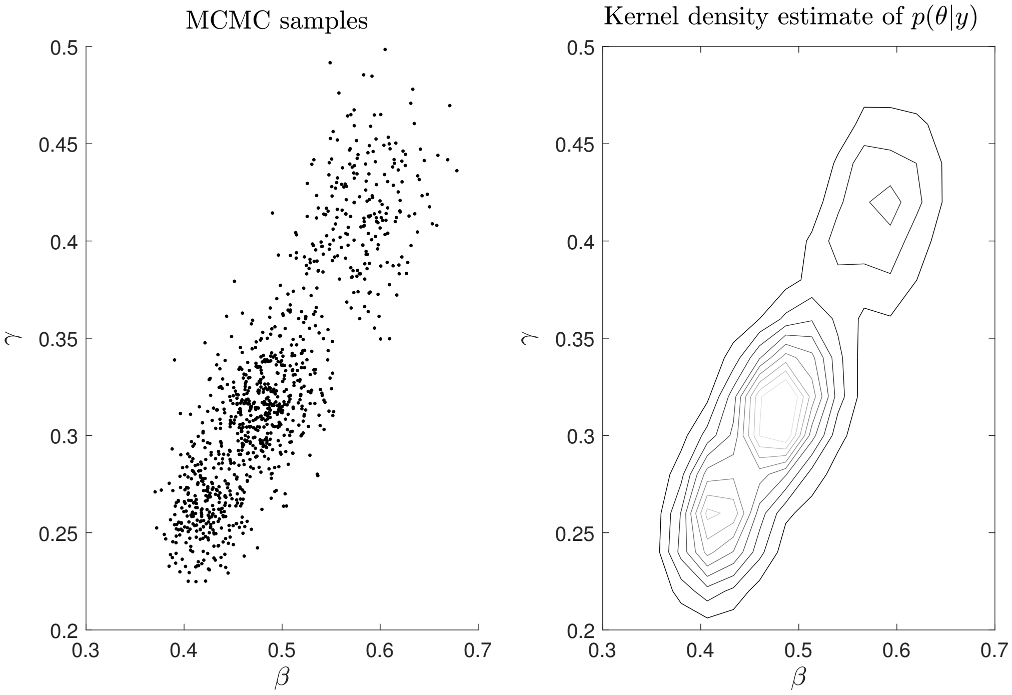

Figure 2.3: An example of using MCMC samples to estimate a posterior distribution via kernel smoothing. The left panel is a scatter plot of samples of parameters, β and γ, from an MCMC algorithm. The right panel is a contour plot of the posterior distribution which is estimated via kernel smoothing.

do not require calculation of the likelihood function; the two main tactics used to perform this kind of inference are either to simulate in a way to give unbiased estimates of the likelihood, or to infer unobserved transitions of the model via imputation methods which then allow calculation of a likelihood.

2.3.2 Gibbs sampling

Gibbs sampling is a tool used in Bayesian inference, and in this thesis it has particular utility in data-augmented MCMC (see Section 2.3.3) [17]. The Gibbs sampling scheme allows samples from a joint distribution to be obtained via sampling from conditional distributions. For example, consider joint density $p(\theta)$ where $\theta = \{\theta_1, \ldots, \theta_k\}$. If it is difficult to sample from $p(\theta)$, one can instead sample consecutively from conditional distributions, $p(\theta_j|\theta\backslash\{\theta_j\})$, a large number of times (as given in Algorithm 3). After a large number of iterations these samples are approximately distributed according to $p(\theta)$. These kinds of sampling schemes, however, scale poorly with the dimension of the parameter space and samples exhibit poor mixing when the target distribution is non-convex as can be the case for posterior distributions related to SEIR models (see Figure 2.8). The Gibbs sampler can be generalised, so that rather than sampling exactly from $p(\theta_j|\theta\backslash\{\theta_j\})$, one can sample a candidate θ_j and accept the candidate with the same acceptance-rejection probability as in a Metropolis-Hastings algorithm; we refer to this as a Hastings step.

```
Initialization:
Choose an initial parameter set, $\theta^{(0)} = \{\theta_1^{(0)}, \ldots, \theta_k^{(0)}\}$;
Choose a large number of samples, $K$;
Sampling:
for $n = 1 : K$ do
    for $j$=1:$k$ do
        sample $\theta_j^{(n)}$ from $p(\theta_j|\theta\backslash\{\theta_j\})$
    end
end
```

Algorithm 3: Gibbs sampling algorithm

Figure 2.4 shows an example of Gibbs sampling. The left panel shows the sample path over the parameter space alternates between jumps in the θ_1 and θ_2 directions. Level sets of the target density are shown under the sample path in the left panel of Figure 2.4, and the right panel shows a kernel density estimate of the target density based on 10,000 Gibbs samples.

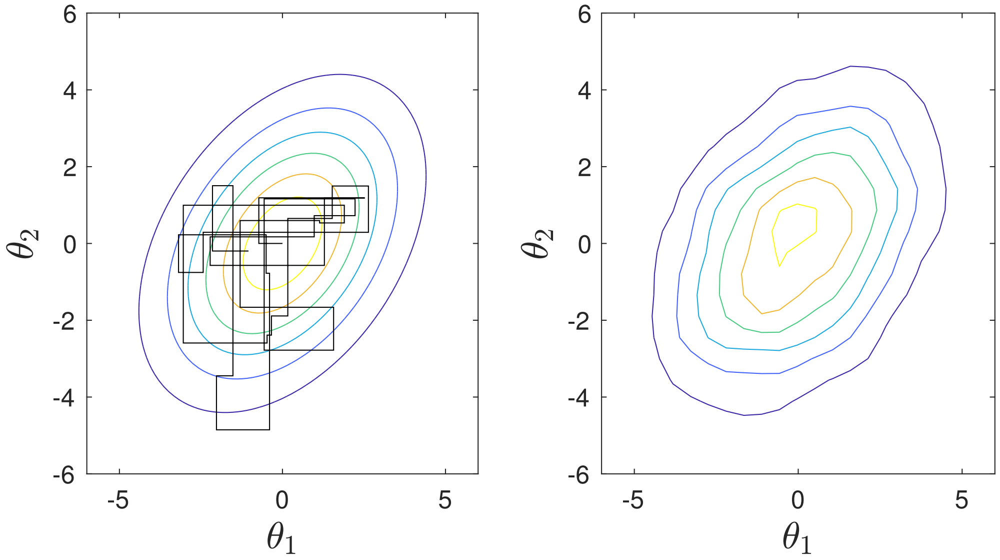

Figure 2.4: An example of Gibbs sampling with target density $p(\theta)$, where $\theta = \{\theta_1, \theta_2\}$. The left panel shows the sample path for the first 50 Gibbs samples plotted over the target density. For this example the target density is Gaussian and hence conditional densities used by the Gibbs sampler are univariate Gaussian. The left panel shows contours of a kernel density approximation of $p(\theta)$ based on 10,000 Gibbs samples.

2.3.3 Data-augmented MCMC

Data-augmented Markov chain Monte Carlo (DA-MCMC) is a powerful, exact, Bayesian inference method for partially-observed Markov chain inference problems in which the full likelihood is intractable. The general approach is to construct an augmented likelihood, the joint density of the data and the missing information given the model parameters, and use this to construct a single-component Metropolis-Hastings algorithm. Essentially it works by increasing the dimension of the inference problem in order to infer missing events as well as parameters via Gibbs sampling or Hastings steps [17]. That is, our observations, $y_{1:T}$, are augmented by the full hidden Markov chain, $x_{[0,T]}$. The idea is that if the likelihood, $p(y_{1:T}|\theta)$, is intractable then we can sample from the joint posterior,

$$\begin{aligned} p\left(\theta, x_{[0,T]}|y_{1:T}\right) &\propto p\left(y_{1:T}, x_{[0,T]}|\theta\right) p(\theta) \\ &\propto p\left(y_{1:T}|x_{[0,T]}, \theta\right) p\left(x_{[0,T]}|\theta\right) p(\theta), \end{aligned}$$

and integrate with respect to $x_{[0,T]}$. Note, $p(x_{[0,T]}|\theta)$ is given as a product of terms given by Equation (2.4), so if $p\left(y_{1:T}|x_{[0,T]}, \theta\right)$ is known this posterior has a tractable *augmented likelihood function*, $p\left(y_{1:T}, x_{[0,T]}|\theta\right)$. The remainder of this section gives some examples of DA-MCMC implemented for epidemic models. These examples are based on uniform priors, however, other sensible choices include: uninformative, exponential priors with low rate parameter; informative inverse uniform priors on parameters (this is equivalent to a uniform prior on the average time spent in a compartment); or, gamma distributed priors, which are conjugate to the augmented likelihood for many of the models in this thesis and hence allow for efficient computation.

Example 1: SIR model

The data-augmented MCMC implemented here is a modified version of that described in [17]. The main idea of the algorithm is that we augment the observations, $y_{1:T}$, with a complete SIR process, $x_{[0,T]}$, in which all transition times and events are given from time 0 up to time T. Suppose our data show the total number of infection events at a daily resolution; we must choose an $x_{[0,T]}$ that agrees with $y_{1:T}$ in that the number of infections each day in $x_{[0:T]}$ must match up with the entries of $y_{1:T}$ and must describe a feasible SIR process. Let $t_1 = 0, t_2 \ldots, t_{n-1}, t_n = T$ denote the transition times and boundaries of $[0, T]$. Let A be the set of transition time indices that correspond to infection events (excluding the initial infection) and B be the indices that correspond to recovery events. Note that, provided that $x_{[0,T]}$ agrees with $y_{1:T}$, we have that $p(y_{1:T}, x_{[0,T]}|\theta) = p(x_{[0,T]}|\theta)$. Hence, by Equation (2.4), we have an augmented likelihood of the form

$$f(y_{1:T}, x_{[0,T]}|\beta, \gamma) = \prod_{i \in A} \frac{\beta S_{t_{i-1}} I_{t_{i-1}}}{N-1} \prod_{j \in B} \gamma I_{t_{j-1}} \exp\left\{-\int_0^T \frac{\beta S_t I_t}{N-1} + \gamma I_t \, dt\right\},$$

where the integral can be expressed in terms of a sum in terms of $x_{[0,T]}$,

$$\int_0^T \frac{\beta S_t I_t}{N-1} + \gamma I_t \, dt = \sum_{j=1}^{n-1} \left(\frac{\beta S_{t_j} I_{t_j}}{N-1} + \gamma I_{t_j}\right)(t_{j+1} - t_j).$$

We proceed by taking Gibbs samples from $f(\beta|x_{[0,T]}, y_{[0:T]}, \gamma)$ and $f(\gamma|x_{[0,T]}, y_{[0:T]}, \beta)$, and Hastings samples from $f(x_{[0,T]}|\beta, \gamma, y_{[0,T]})$. Suppose we have priors $\beta/\gamma \sim$ Uniform(a,b) and $\gamma \sim$ Uniform(c,d), then

$$\begin{aligned} f(\beta|x_{[0,T]}, y_{1:T}, \gamma) &\underset{\beta}{\propto} f(y_{1:T}, x_{[0,T]}, \beta, \gamma) \\ &\underset{\beta}{\propto} f(y_{1:T}, x_{[0,T]}|\beta, \gamma)p(\beta|\gamma)p(\gamma) \\ &\underset{\beta}{\propto} \beta^{|A|} \exp\left\{ -\int_0^T \beta \frac{S_t I_t}{N-1}\, dt \right\} \quad \text{for } \beta \in [a\gamma, b\gamma]. \end{aligned}$$

where '$f \underset{\beta}{\propto} g$' is notation for '$f$ is proportionate to g with respect to β'. Hence,

$$\beta|x_{[0,T]}, y_{1:T}, \gamma \sim \text{Gamma}\left(|A|+1, \int_0^T \frac{S_t I_t}{N-1}\, dt \right)_{[a\gamma, b\gamma]},$$

where Gamma$(a,b)_x$ refers to a gamma distributed variable with shape parameter a and rate parameter b, truncated to the interval x. Similarly

$$f(\gamma|x_{[0,T]}, y_{[1:T]}, \beta) \underset{\gamma}{\propto} \gamma^{|B|-1} \exp\left\{ -\int_0^T \gamma I_t\, dt \right\} \quad \text{for } \gamma \in (c,d),$$

so,

$$\gamma|x_{[0,T]}, y_{[1:T]}, \beta \sim \text{Gamma}\left(|B|, \int_0^T I_t\, dt \right)_{[c,d]} \quad \text{for } |B| \neq 0.$$

Note that the shape parameter of $f(\gamma|x_{[0,T]}, y_{[0:T]}, \beta)$ is $|B|$, rather than $|B|+1$, because the conditional prior for β is given by $p(\beta|\gamma) = 1/(\gamma(b-a))$. As it is computationally efficient to sample from gamma distributed random variables, sampling from these conditional densities is highly efficient. Note, if $|B| = 0$, that is, there are no recoveries in $x_{[0,T]}$, Hastings steps are necessary to sample γ.

Lastly we sample from $f(x_{[0,T]}|y_{[0:T]}, \beta, \gamma)$ by performing Hastings steps. These steps need to ensure irreducibility of the proposed $x_{[0,T]}$ over the space of SIR epidemics which agree with the observed data. Hence, if given a feasible $x_{[0,T]}$ we need to be able to move infection times to over the day, move recovery events over $[0,T]$, and, as we do not know the number of recoveries, we need to be able to insert and remove recovery events. The Hastings steps consist of choosing one of these four kinds of proposals according to an arbitrary pmf $\{p_1, p_2, p_3, p_4\}$ and then proposing one the following moves accordingly.

(i) Uniformly randomly select an infection event and move it uniformly over the day to give candidate path $x'_{[0,T]}$. Accept this candidate path with probability

$$\min\left\{\frac{f(y_{1:T}, x'_{[0,T]}|\beta, \gamma)}{f(y_{1:T}, x_{[0,T]}|\beta, \gamma)}, 1\right\}.$$

(ii) Uniformly randomly select a recovery event and move it uniformly over $[0, T]$ to give candidate path $x'_{[0,T]}$. Accept this candidate path with probability

$$\min\left\{\frac{f(y_{1:T}, x'_{[0,T]}|\beta, \gamma)}{f(y_{1:T}, x_{[0,T]}|\beta, \gamma)}, 1\right\}.$$

(iii) Uniformly randomly select and remove a recovery event to give candidate path $x'_{[0,T]}$. Accept this candidate path with probability

$$\min\left\{\frac{f(y_{1:T}, x'_{[0,T]}|\beta, \gamma)|B|p_4}{f(y_{1:T}, x_{[0,T]}|\beta, \gamma)Tp_3}, 1\right\}.$$

(iv) Insert a recovery event at a Uniform$(0, T)$ distributed time to give candidate path $x'_{[0,T]}$. Accept this candidate path with probability

$$\min\left\{\frac{f(y_{1:T}, x'_{[0,T]}|\beta, \gamma)Tp_3}{f(y_{1:T}, x_{[0,T]}|\beta, \gamma)(|B| + 1)p_4}, 1\right\}.$$

So the DA-MCMC algorithm, in this case, begins by choosing a feasible initial hidden process, $x_{[0,T]}$, and model parameters. This can be done by sampling parameters from the prior distribution, uniformly generating infection times over each day, and Exponential(γ) infectious periods for every observed infectious case; this can be repeated until a feasible realisation from the SIR process is obtained. The algorithm proceeds by sampling from truncated gamma-distributed random variables and proposing simple Hastings steps for many iterations. After many iterations, the values of β and γ will be a sample from the posterior distribution of interest, $f(\beta, \gamma|y_{1:T})$. Note that this is a simple implementation of the DA-MCMC algorithm for demonstrative purposes; great efficiencies can be made by optimising the proposal density in the Hastings step or by using a non-centered reparameterisation of the model [54, 55].

Example 2: Labelled SIR model

An alternative approach to this inference problem is to assume a more complicated

Markov chain. In this new chain individuals are labelled, that is, a state in the chain gives information on which individuals are susceptible, infected and recovered, as described in the second half of Section 2.2.1. The reason we consider this slightly more complicated Markov chain is that, rather than uniformly randomly proposing recovery times over $[0, T]$, we can propose infectious periods for each individual from an appropriate distribution. Proposed changes to the underlying epidemic process are accepted more often, which improves mixing. Further, the method becomes simpler as we no longer need to change the dimension of the sample space; we can simply allow recoveries to occur after time T, as illustrated in Figure 2.5. While it is counter-intuitive to fix the dimension of the hidden process at its largest possible value, this allows the Hastings steps to be defined in terms of only two kinds of transitions, and, it simplifies that augmented likelihood function. Further, this approach easily extends to more complicated models.

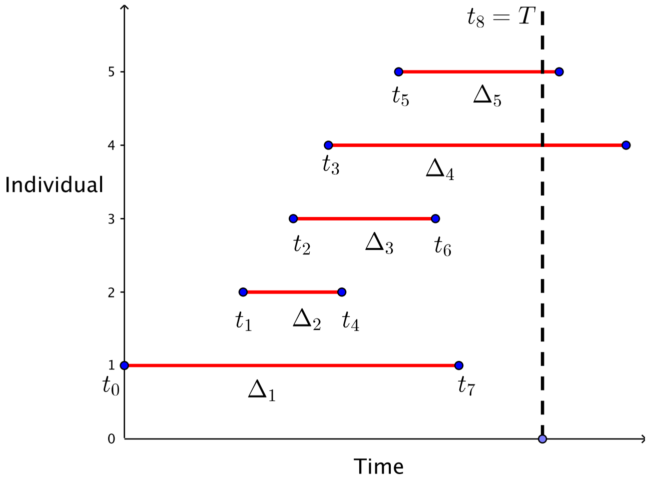

Figure 2.5: An Illustration of the labelled SIR process. The infectious periods of individuals are shown by red horizontal lines. Note that we have data up to the present time, T, but for the purposes of inference we impute recovery times of all infectious individuals, allowing recoveries to occur in the future.

The new likelihood for transitions up to time T is given by

$$f(y_{1:T}, x_{[0,T]}|\beta, \gamma) = \gamma^{|B|} \prod_{i \in A} \frac{\beta I_{t_{i-1}}}{N-1} \exp\left\{ - \int_0^T \frac{\beta S_t I_t}{N-1} + \gamma I_t \ \ dt \right\}. \tag{2.7}$$

However, as we allow recoveries to occur after time T, there is an extra term incorporated into the likelihood. After time T we observe no more infections, we merely need to account for recoveries. Hence, the process can simply be thought of as a death process, giving

$$f(x_{[T,\infty)}|\beta, \gamma) = \gamma^{|A|+1-|B|} \exp\left\{ \int_T^\infty \gamma I_t \ \ dt \right\}, \tag{2.8}$$

noting that absorbtion into $I_t = 0$ is guaranteed for some finite t. Combining Equations (2.7) and (2.8) gives

$$\begin{aligned} f(y_{1:T}, x_{[0,\infty)}|\beta, \gamma) =& \gamma^{|A|+1} \beta^{|A|} \prod_{i \in A} \frac{I_{t_{i-1}}}{N-1} \exp\left\{ - \int_0^T \frac{\beta S_t I_t}{N-1} \ \ dt \right\} \\ & \times \exp\left\{ - \int_0^\infty \gamma I_t \ \ dt \right\}. \end{aligned}$$

Equation (2.8) can be simplified further by noting that $\int_0^\infty I_t \ dt$ is simply the sum of the infectious periods of each individual. Let the infectious period of individual p be denoted by Δ_p. We can express the augmented likelihood as

$$\begin{aligned} f(y_{1:T}, x_{[0,\infty)}|\beta, \gamma) =& \gamma^{|A|+1} \beta^{|A|} \prod_{i \in A} \frac{I_{t_{i-1}}}{N-1} \exp\left\{ - \int_0^T \frac{\beta S_t I_t}{N-1} \ \ dt \right\} \\ & \times \exp\left\{ - \sum_{p=1}^{|A|+1} \gamma \Delta_p \right\}. \end{aligned} \tag{2.9}$$

Hence the marginal posterior for γ is

$$\gamma|\beta, x_{[0,\infty)}, y_{1:T} \sim \text{Gamma}\left(|A|+1, \sum_{p=1}^{|A|+1} \Delta_p \right)_{[c,d]},$$

and the β marginal posterior is as in the previous example.

The Hastings step now consists of choosing between shifting an infection event or an infectious period according to p and $1-p$, for $p \in (0,1)$. Specifically, the changes to $x_{[0,\infty)}$ are made as follows.

(i) Uniformly randomly select an infection time and move it uniformly over the day to get candidate path $x'_{[0,T]}$. Accept the new path with probability

$$\min\left\{ \frac{f(y_{1:T}, x'_{[0,T]}|\beta, \gamma)}{f(x_{[0,T]}|\beta, \gamma)}, 1 \right\}.$$

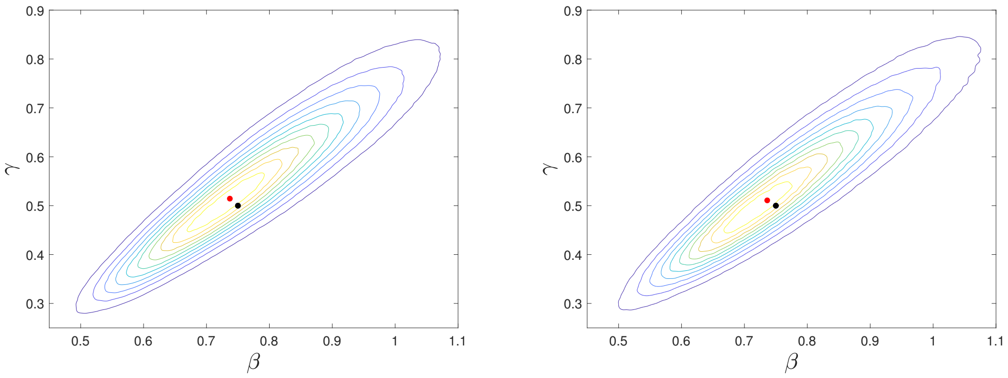

Figure 2.6: Kernal density estimates of the posterior distribution based on DA-MCMC samples from the labelled SIR model (left) and standard SIR model (right). The true parameters are $(\beta, \gamma) = (0.5, 0.75)$ with prior distribution given by $\beta/\gamma \sim$ Uniform(0.9,10) and $\gamma \sim$ Uniform(0.02,4). The population was of size N=100000 and inference is performed on a data set with 417 individuals infected over 33 days. The kernel density estimates are based upon 1.5×10^7 samples with 5×10^6 iterations of burn-in. The red dot is the MAP estimate and the black dot is the true parameter values.

(ii) Uniformly randomly select an individual that became infectious over [0,T], q, and sample a candidate Exponential(γ) distributed infection duration, Δ'_q, to get candidate path $x'_{[0,T]}$. The new point is accepted with probability

$$\min\left\{\frac{f(y_{1:T}, x'_{[0,T]}|\beta,\gamma)e^{-\gamma(\Delta_q-\Delta'_q)}}{f(x_{[0,T]}|\beta,\gamma)}, 1\right\}.$$

Figure 2.6 demonstrates that this method agrees with the unpaired version of the algorithm. The benefit of this method is that, although each iteration may take slightly longer, mixing is improved.

Example 3: Labelled SI(n)R model

Exponentially distributed infectious and latent periods are often assumed for computational reasons; in many cases it is more realistic to consider a less skewed distribution, the simplest such distribution to implement is an Erlang distribution. Here we present methodology for a homogeneous SIR model with Erlang(n, λ) infectious period; this is often referred to as a labelled SI(n)R model. Throughout this thesis we let $\lambda = n\gamma$, such that the mean infectious period is still $1/\gamma$. Inference for this model via DA-MCMC does not technically require n to be an integer, however, we only consider integers here to easily allow the model to be represented as a CTMC after inference.

We derive the augmented likelihood by considering the labelled approach from the previous example. Under the labelled approach we know if there are $|A|+1$ infections those individuals must each have $n-1$ phase changes and a recovery; so these individuals must have $n(|A|+1)$ transitions that occur at rate λ. Further, as the rates of this likelihood do not change from phase change events, the augmented likelihood may be expressed in terms of infectious periods without considering the phase change transitions. Hence, the augmented likelihood is given by

$$\begin{aligned} f(y_{1:T}, x_{[0,\infty)}|\beta,\lambda) &= \prod_{i\in A} \frac{\beta I_{t_{i-1}}}{N-1} \exp\left\{-\int_0^T \frac{\beta S_t I_t}{N-1}\ dt\right\} \\ &\times \lambda^{n(A+1)} \exp\left\{-\lambda \sum_{i=1}^{|A|+1} \Delta_p\right\}. \end{aligned}$$

Note that as this likelihood does not depend on the times of phase changes we can consider a likelihood which only accounts for infection events and recovery events, that is, we need not augment our data with all of the hidden transitions; this allows us to perform inference in a lower dimensional space. To get a new augmented likelihood in this lower dimensional space we integrate the augmented likelihood with respect to the inter-arrival times of all of the phase changes of each individual (where individual p has phase i duration Δ_p^i); that is we integrate over the region $A_p = \left\{\Delta_p^{(1:n-1)} : \sum_{k=1}^{n-1} \Delta_p^k \leq \Delta_p\right\}$ for all p. This yields a new augmented likelihood, $\hat{f}$, of the form

$$\begin{aligned} \hat{f}(y_{1:T}, x_{[0,\infty)}|\beta,\lambda) &= f(y_{1:T}, x_{[0,\infty)}|\beta,\lambda) \prod_{p=1}^{|A|+1} \frac{\Delta_p^{n-1}}{(n-1)!} \\ &= \prod_{i\in A} \frac{\beta I_{t_{i-1}}}{N-1} \exp\left\{-\int_0^T \frac{\beta S_t I_t}{N-1}\ dt\right\} \\ &\times \prod_{p=1}^{|A|+1} \frac{\lambda^n \Delta_p^{n-1}}{(n-1)!} \exp\left\{-\lambda\Delta_p\right\}. \end{aligned} \tag{2.10}$$

Hence during the Hastings step we need not sample phase change events. Note the likelihood above is expressed in a way such that it highlights the relationship of the likelihood function with the Erlang distributed recoveries. Further, it can be shown that for labelled models the augmented likelihood can be split into a term related to the inhomogeneous Poisson transmission process and a term of the product of the infectious period pdf's; this is the result of all individuals acting independently after transmission. That is, for, a SIR model with a general infectious period that depends on parameters ψ, the augmented likelihood becomes

$$f(y_{1:T}, x_{[0,\infty)}|\beta, \psi) = \prod_{i \in A} \frac{\beta I_{t_{i-1}}}{N-1} \exp\left\{ -\int_0^T \frac{\beta S_t I_t}{N-1} \; dt \right\} \prod_{p=1}^{|A|+1} f(\Delta_p|\psi);$$

a similar statement can be made with respect to the labelled SEIR-type model.

Let λ have a Uniform$(0.1n, 4n)$ prior distribution. From Equation (2.10) it follows that the marginal posterior distribution for λ is

$$\lambda|\beta, x_{[0,\infty)}, y_{1:T} \sim \text{Gamma}\left(n(|A|+1), \sum_{p=1}^{|A|+1} \Delta_p \right)_{[0.1n,4n]} ;$$

the marginal posterior distribution for β is as described in the previous sections. Lastly the Hastings steps consists, again, of either shifting the infection times or recovery times of the hidden process according to probabilities p and $1-p$ for $p \in (0,1)$. Specifically, the changes are made to $x_{[0,\infty)}$ as follows.

(i) Uniformly randomly select an infection time, and uniformly choose a candidate infection time over the day. The new point is accepted with probability

$$\min\left\{ \frac{\hat{f}(y_{1:T}, x'_{[0,\infty)}|\beta, \lambda)}{\hat{f}(y_{1:T}, x_{[0,\infty)}|\beta, \lambda)}, 1 \right\}.$$

(ii) Uniformly randomly select an individual that became infectious, say individual p, and sample a candidate Erlang(n, λ) distributed infectious period, Δ'_p. The new point is accepted with probability

$$\min\left\{ \frac{\hat{f}(y_{1:T}, x'_{[0,\infty)}|\beta, \lambda)}{\hat{f}(y_{1:T}, x_{[0,\infty)}|\beta, \lambda)} \times \left(\frac{\Delta_p}{\Delta'_p}\right)^{n-1} e^{-\lambda(\Delta_p - \Delta'_p)}, 1 \right\}.$$

The interesting feature of this scheme is that there is no need to keep track of the time of phase changes and there is no need to consider the number of individuals in each infectious phase, as there are no parts of the augmented likelihood that depend on these. Hence, the paired approach for the SI(n)R model should be computationally similar for all $n \in \mathbb{N}$.

Results from the inference using a single simulation of a SI(3)R epidemic are shown in Figure 2.7. From this we observe that the high kernel density values appear quite linear with very little variance about the line. This shows that, due to the low variance in the infectious period, we are able to accurately estimate R_0; however there is still uncertainty around individual parameter estimates.

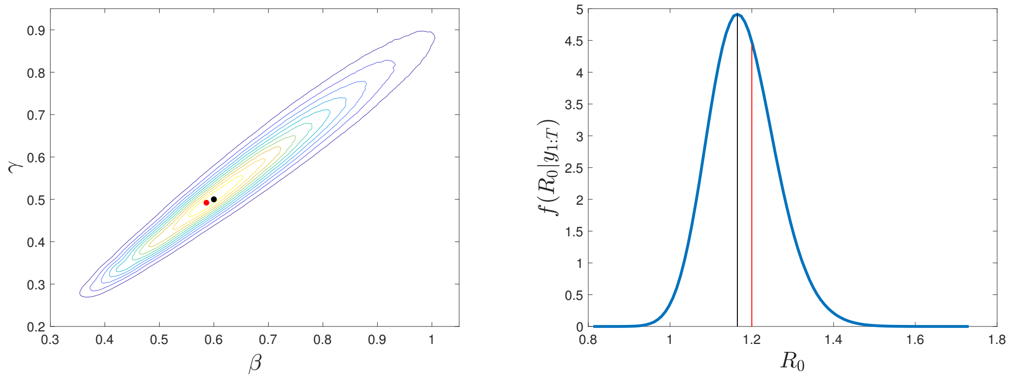

Figure 2.7: A Kernal density estimate of the parameters from an SI(3)R model (left) and kernel density estimate of R_0 (right) from DA-MCMC. The true parameters are $(\beta, \gamma) = (0.5, 0.6)$ with prior distribution given by $\beta/\gamma \sim$ Uniform(0.9,10) and $\gamma \sim$ Uniform(0.02,4). The population was of size $N = 10^5$ and inference is performed on a data set with 438 individuals infected over 21 days. The kernel density estimates estimates are based upon 2×10^7 samples with 5×10^6 iterations of burn-in. The black points are the MAP estimate and the red points are the true parameter values.

An extension of this approach is to also infer the shape parameter of the model; the extension simply requires sampling from

$$f(n|\gamma, \beta, x, \theta) \underset{n}{\propto} \frac{(\prod_{p=1}^{|A|+1} \lambda \Delta_p)^n p(n)}{n!^{|A|+1}}; \tag{2.11}$$

note that for $p(n)$ on finite support, sampling is a matter of calculating $f(n|\gamma, \beta, x, \theta)$ and normalising. In practice these quantities can be difficult to work with, as they require division by a factorial number to a possibly large power. For all but small data sets the mixing of the MCMC is slow in n, this is one motivation for a sequential DA-MCMC algorithm which is discussed further in Section 3.3.

Example 4: SEIR model with stochastic seeding time

In this section we present the methodology for data augmented inference on a homogeneous SEIR model. Suppose that the process is still initialised by a single infectious individual seeded in the population at a Uniform$(0, 1)$ distributed time. The only difference between the SEIR and the SIR model is that there is a Exponential(σ) distributed time between the exposure of an individual to the disease and the time at which they become infectious.

Similar to before, with C as the set of indicies corresponding to exposures from time t_1, the joint density conditional on β, σ, γ and t_1 can then be expressed as

$$\begin{aligned} f(y_{1:T}, x_{(t_1,T]}|\beta, \sigma, \gamma, t_1) &= \prod_{i\in A} \frac{\beta S_{t_{i-1}} I_{t_{i-1}}}{N-1} \prod_{j\in B} \gamma I_{t_{i-1}} \prod_{k\in C} \sigma E_{t_{i-1}} \\ &\times \exp\left\{ - \int_{t_1}^{T} \frac{\beta S_t I_t}{N-1} + \sigma E_t + \gamma I_t \;\; dt \right\}. \end{aligned}$$

The marginal distributions of β and γ are as for the homogeneous SIR model, except now the rates are integrated from t_1. We are left to sample from the marginal posterior distributions of σ and t_1. Suppose the prior of σ is Uniform(e, f). The marginal distribution of σ, is similar to the other parameters,

$$\sigma|\beta, \gamma, x_{(t_1,T]}, t_1 \sim \text{Gamma}\left(|C| + 1, \int_{t_1}^{T} E_t \; dt \right)_{[e,f]}.$$

The marginal posterior distribution t_1 is given by

$$\begin{aligned} f(t_1|\beta, \sigma, \gamma, x_{[t_2,T]}) &\underset{t_1}{\propto} f(y_{1:T}, x_{(t_1,T]}|\beta, \sigma, \gamma, t_1)p(t_1) \\ &= e^{(\beta+\gamma)t_1} \left[\int_0^{\min\{1,t_2\}} e^{(\beta+\gamma)t} \;\; dt \right]^{-1} \\ &= \frac{(\beta+\gamma)e^{(\beta+\gamma)t_1}}{e^{(\beta+\gamma)\min\{1,t_2\}} - 1} \qquad \text{for } t_1 \in (0, \min\{1, t_2\}), \end{aligned}$$

as the model assumption is that t_1 is seeded uniformly over the day, though conditionally t_1 must occur before t_2. As the cdf of the marginal posterior of t_1 is analytically invertible, t_1 may be sampled via inverse transform sampling. Note that $|B| \leq |A| + 1 \leq |C| + 1$, as the number of exposures and recoveries are bounded below and above by the number of observations, respectively. As the latent process needs to be altered by shifting removing or inserting exposure events, there are three new kinds of Hastings steps compared to those for the standard SIR model. We make one of the seven moves according to distribution $\{p_1, \ldots, p_7\}$. The moves related to exposures are as follows.

(v) Uniformly randomly select an exposure time event and choose a candidate Uniform(t_1, T) distributed recovery time to get new sample path $x'_{(t_1,T]}$. The new path is accepted with probability

$$\min\left\{\frac{f(y_{1:T}, x'_{(t_1,T]}|\beta, \lambda, t_1, \sigma)}{f(y_{1:T}, x_{(t_1,T]}|\beta, \lambda, t_1, \sigma)}, 1\right\}.$$

(vi) If $|C| > |A|$ uniformly randomly select an exposure time and remove it to get new sample path $x'_{(t_1,T]}$. The new path is accepted with probability

$$\min\left\{\frac{f(y_{1:T}, x'_{(t_1,T]}|\beta, \lambda, t_1, \sigma)|C|p_7}{f(y_{1:T}, x_{(t_1,T]}|\beta, \lambda, t_1, \sigma)(T - t_1)p_6}, 1\right\}.$$

(vii) Insert a Uniform(t_1, T) distributed exposure time to get new sample path $x'_{(t_1,T]}$. The new path is accepted with probability

$$\min\left\{\frac{f(y_{1:T}, x'_{(t_1,T]}|\beta, \lambda, t_1, \sigma)(T - t_1)p_6}{f(y_{1:T}, x_{(t_1,T]}|\beta, \lambda, t_1, \sigma)(|C| + 1)p_7}, 1\right\}.$$

Results for a simulated data set for the SEIR model are shown in Figure 2.8. Note that there appears to be poor identifiability in individual parameters, as the posterior shows relatively high variance, however the posterior appears to peak along the line where the reproduction number and growth rate take their input values. That is, while the data may not be rich enough to distinguish all model parameters, it may still be used to accurately infer important epidemiological quantities.

2.3.4 Particle filters

Particle filters, a kind of Sequential Monte Carlo (SMC) method, are a Bayesian inference method which targets a sequence of distributions [56]. This allows new data to be incorporated into the inference procedure without needing to perform inference anew on the whole series of observations. Further, these methods are embarrassingly parallelisable, which allows for efficient inference on a cluster. In this section we describe particle filters in their simplest form, in which a set of model parameters is fixed and the hidden process, $x_{[0:T]}$, is to be inferred.

Given some initial state density $p(x_0)$, the Q-matrix of the hidden process and a set of observations $y_{1:T}$, we wish to estimate the sequence of distributions of the hidden

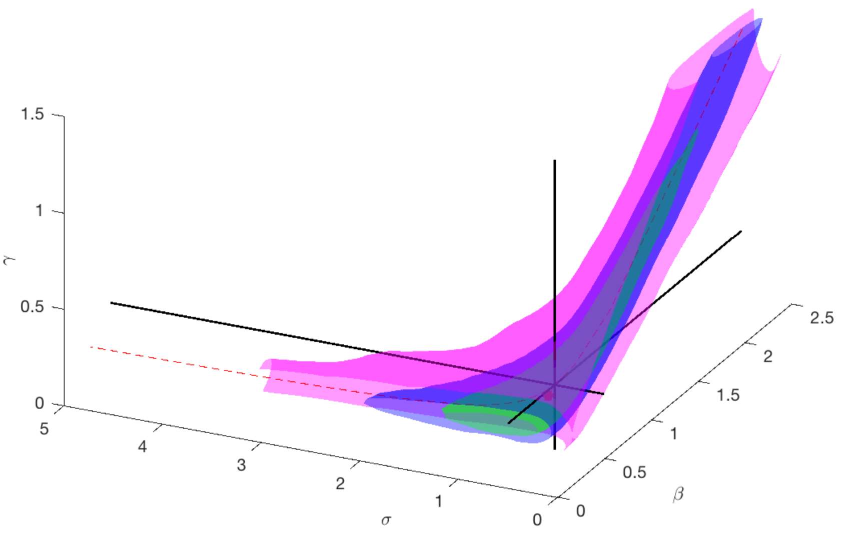

Figure 2.8: Kernel density estimate of the parameters from a SEIR model. These are based upon 2×10^7 samples with 5×10^6 iterations of burn-in. The true parameter values are given by the intersection of the black lines, the red dot gives the MAP estimate and the red dotted line is the line which gives the intersection of the true value of R and the early growth rate.

process, $p(x_{[0,T]}|y_{1:T})$, and the likelihood function, $p(y_{1:T})$. SMC typically makes use of the recursive formula

$$\begin{aligned} p(x_{[0,t+1]}|y_{1:t+1}) &= \frac{p(x_{[0,t+1]}, y_{t+1}|y_{1:t})}{p(y_{t+1}|y_{1:t})} \\ &= p(x_{[0,t]}|y_{1:t})\frac{p(x_{(t,t+1]}, y_{t+1}|y_{1:t}, x_{[0,t]})}{p(y_{t+1}|y_{1:t})} \\ &= p(x_{[0,t]}|y_{1:t})\frac{p(x_{(t,t+1]}, y_{t+1}|x_t)}{p(y_{t+1}|y_{1:t})} \\ &= p(x_{[0,t]}|y_{1:t})\frac{p(y_{t+1}|x_{[t,t+1]})p(x_{(t,t+1]}|x_t)}{p(y_{t+1}|y_{1:t})}, \end{aligned} \tag{2.12}$$

where the likelihood function is given by a product of likelihood increments,

$$p(y_{t+1}|y_{1:t}) = \int p(y_{t+1}|x_{[t,t+1]})p(x_{(t,t+1]}|x_t) \quad dx_{(t,t+1]}. \tag{2.13}$$

Typically when the state space is large it is difficult to evaluate Equation (2.13), and hence Equation (2.12). Thus, many methods aim to sample from the sequence of target distributions $p(x_{[0,t]}|y_{1:t})$ and use the samples to estimate the likelihood function.

Here we outline a common particle filter method for estimating solutions to Equations (2.12) and (2.13), the sequential importance sampling filter [22]. This is a method which obtains unbiased samples of $x_{[0,t]}$ by sequentially sampling $x_{(t,t+1]}$ from arbitrary densities and accounting for the sampling procedure. These could be obtained via a Doob-Gillespie algorithm, or if available, they could be obtained by sampling exactly from $p(x_{[t,t+1)}|x_t, y_{t+1})$, or by some other method. The algorithm is as follows: Take N samples, or particles, $x_0^{(1:N)}$, from some density $q_0(x_0)$. For each i calculate particle *weights*,

$$w_0^i = \frac{p(x_0^i)}{q_0(x_0^i)}$$

and *normalised weights*

$$W_t^i = \frac{w_t^i}{\sum_{j=1}^N w_t^j}. \tag{2.14}$$

From $t = 1$ we proceed by recursively sampling $x^i_{(t-1,t]}$ from arbitrary density $q(\cdot|x_{t-1})$ and estimate $p(x^i_{[0,t]}|y_{1:t})$ up to proportionality by calculating weights

$$w_t^i = W_{t-1}^i \frac{p(y_t|x^i_{[t-1,t]})p(x^i_{(t-1,t]}|x^i_{t-1})}{q(x^i_{(t-1,t]}|x_{t-1})},$$

which is a particle approximation of Equation (2.12) up to proportionality. The particle filter approximation of $p(x_{[0,t]}|y_{1:t})$ is given by

$$\hat{p}(x_{[0,t]}|y_{1:t}) = \sum_{i=1}^N \delta_{x^i_{[0,t]}}(x_{[0,t]})W^i,$$

where δ_x denotes the Dirac delta function with mass at x. The particle filter approximation of the likelihood increments, from Equation (2.13), is given by

$$\hat{p}(y_t|y_{1:t-1}) = \sum_{i=1}^N w_t^i.$$

The product of likelihood increments provide an unbiased estimate of the likelihood [57]. One issue with the particle filter is that some sampling densities can lead to weights with infinite variance and, more generally, the variance becomes large when applying the sequential importance sampling filter to a long series of observations. Variance can be decreased by resampling particles, where particle i is resampled with probability W_t^i. Once resampling is performed the particle weights are set to $W_t^i = \frac{1}{N}$ for $i = 1, \ldots, N$. The resampling procedure removes particles associated with low weights, which correspond to unlikely realisations of $x_{[0,t]}$, and increases the number of particles that correspond to likely realisations of $x_{[0,t]}$ given $y_{1:t}$. The sequential importance sampling algorithm with resampling is aptly referred to as the Sequential Importance Resampling algorithm; see Algorithm 4. Algorithm 4 only gives output for inferring the current state of the system, x_T; in practice one can infer $x_{[0,T]}$ via saving full trajectories $x_{[0,t]}$ at each iteration. Resampling need not be performed at every iteration, a common strategy is to only resample if the *effective sample size* (ESS) drops below a given threshold, say ζN for $\zeta \in (0, 1]$ [22]. The effective sample size is given by

$$\text{ESS} = \frac{1}{\sum_{i=1}^{N} (W^i)^2}.$$

The ESS represents the number of approximately independent samples in a set of correlated samples.

Resampling particles many times often leads to particles x_T coming from few initial states x_0; this issue is referred to as *particle degeneracy.* That is, if N is not sufficiently large and we start with a sample $x_0^{(1:N)}$ and sequentially resample the particles many times, the number of distinct x_0 values can only decrease. More generally, earlier states of the system will be degenerate if T is large and N is not sufficiently large. Particle degeneracy implies that we have few distinct samples from the posterior distribution, so estimates have high error. The method of resampling has an effect on how quickly particle degeneracy occurs. Algorithm 4 uses multinomial resampling for simplicity, but generally systematic resampling can improve performance [22, 58]. A solution to particle degeneracy is to implement a *smoothing* method, for example, forwards-filtering backwards sampling [59]. As the particle filter gives unbiased estimates of the likelihood function, the particle filter admits joint parameter and state inference, this is explained further in Section 2.3.5.

```
Inputs: The number of particles, $N$ and observed data, $y_{1:T}$.
Outputs: The likelihood, $\hat{p}(y_{1:T})$ and weighted particles, $\{x_T^i, W^i\}_{i=1}^N$.
Initialization:
for $i = 1, \dots, N$ do
    Sample an initial state $x_0^i \sim q_0(\cdot)$;
    Calculate initial weight $w^i = \frac{p(x_0^i)}{q(x_0^i)}$;
end
For $i = 1, \dots, N$ calculate normalised weights $W^i = w^i / \sum_{j=1}^N w^j$;
Iterations:
for $t = 1 : T$ do
    Importance sampling step:
    for $i = 1, \dots, N$ do
        Sample $x_{(t-1,t]}^i$ from arbitrary density $q(\cdot|x_{t-1})$;
        Update weights;
            $w^i = W^i \frac{p(y_t|x_{[t-1,t]}^i)p(x_{(t-1,t]}^i|x_{t-1}^i)}{q(x_{(t-1,t]}^i|x_{t-1}^i)}$;
    end
    For $i = 1, \dots, N$ calculate normalised weights $W^i = w^i / \sum_{j=1}^N w^j$;
    $\hat{p}(y_{1:t}) = \left(\sum_{i=1}^N w^i\right) \hat{p}(y_{1:t-1})$;
    Resampling step:
    if some condition holds then
        Sample $N$ indicies, $r_1, \dots, r_N$, from $1, \dots, N$ with probabilities $W^1, \dots, W^N$;
        For $i = 1, \dots, N$ set the new particle according to index $r_i$, that is, set
          $\{x_t^i, W^i\} = \{x_t^{r_i}, 1/N\}$;
    end
end
```

Algorithm 4: Psuedocode for a sequential importance resampling algorithm

2.3.5 Particle-marginal MCMC

Particle-marginal MCMC (PM-MCMC) is a method which allows parameter inference to be performed when the likelihood function is intractable [60]. It uses positive unbiased estimates of the likelihood function, which can be obtained from a particle filter, in an otherwise standard MCMC scheme. This new MCMC scheme still converges to the target distribution; however this new scheme exhibits slower mixing and requires a longer burn in period than the ideal MCMC algorithm [60].

Here we outline and justify the particle-marginal Metropolis-Hastings algorithm by expanding upon the description given in [60]. The algorithm works by creating a Markov chain that samples over the parameter space and a space of positive, unbiased, likelihood estimates. Once the Markov chain reaches stationarity, the samples from the parameter space are distributed according to the posterior distribution. Specifically, given some initial point θ and likelihood estimate $\hat{p}(y|\theta)$, we propose the next parameter value and likelihood estimate according to $\theta^* \sim q_1(\cdot|\theta)$ and $\hat{p}(y|\theta^*) \sim q_2(\cdot|\theta^*)$; where q_1 is an arbitrary sampling density and q_2 is such that $E\left[\hat{p}(y|\theta)\right] = p(y|\theta)$. The proposal is accepted with probability

$$\min\left\{1, \frac{\hat{p}(y|\theta^*)p(\theta^*)q_1(\theta|\theta^*)}{\hat{p}(y|\theta)p(\theta)q_1(\theta^*|\theta)}\right\},$$

which is the acceptance probability used in a standard Metropolis-Hastings algorithm where the likelihood is replaced by the unbiased estimate, $\hat{p}(y|\theta)$. Note that the acceptance probability above is equal to

$$\min\left\{1, \frac{\hat{p}(y|\theta^*)p(\theta^*)q_1(\theta|\theta^*)q_2(\hat{p}(y|\theta^*)|\theta^*)q_2(\hat{p}(y|\theta)|\theta)}{\hat{p}(y|\theta)p(\theta)q_1(\theta^*|\theta)q_2(\hat{p}(y|\theta^*)|\theta^*)q_2(\hat{p}(y|\theta)|\theta)}\right\}.$$

We can see that this is equivalent to an accepting a proposal of $\hat{p}(y|\theta^*)$ and θ^* in a MCMC scheme that targets $\hat{p}(y|\theta)q_2(\hat{p}(y|\theta)|\theta)p(\theta)$ up to proportionality. Marginalising over $\hat{p}(y|\theta)$ gives

$$p(\theta)\int \hat{p}(y|\theta)q_2(\hat{p}(y|\theta)) \;\; d\hat{p}(y|\theta) = p(y|\theta)p(\theta),$$

therefore the stationary distribution of the Markov chain marginalised with respect to $\hat{p}(y|\theta)$, is equal to the posterior distribution. As particle filters give unbiased estimates of the likelihood one can choose sampler q_2 to be a particle filter that samples unbiased estimates of the likelihood. Hence, PM-MCMC can be used to obtain samples from the parameter space, and for each of these parameter samples the particle filter gives samples from the state space.

2.3.6 Sequential Monte Carlo squared

Sequential Monte Carlo squared (SMC-squared) is an algorithm for both inferring parameters and states of a partially-observed stochastic process [24–26]. The algorithm uses a sequential importance resampling algorithm in the parameter space where the weight of each parameter value is obtained via a particle filter in the state space. This allows parameter and state estimates to be updated sequentially as new data are obtained. This provides an alternative to PM-MCMC, where the sequential nature of the algorithm allows samples to be made in important parts of the parameter space (samples with reasonable posterior support). During the inference, SMC-squared also offers an estimate of the model evidence, which allows one to perform Bayesian model selection. However, the error in the model evidence estimate can be too large to effectively select between models [25].

Algorithm 5 is an example of an SMC-squared algorithm. The algorithm considers N_θ weighted particles in the parameter space, $\{\theta^i, W^i\}_{i=1}^{N_\theta}$, where each parameter set, θ^i, is associated with N_x state particles, $[\{x_t\}_{j=1}^{N_x}]^i$. For simplicity Algorithm 5 assumes that each state particle has the same weight; though in practice one can run an SMC-squared algorithm while keeping track of both state particle weights and parameter particle weights. The state particles will have equal weights if resampling is performed after the particle filter step, or if the states are sampled such that the weights are equal, such as via the Alive particle filter [25]. Algorithm 5 avoids particle degeneracy (discussed in Section 2.3.4) by performing a resampling step (to remove particles with low posterior support) and a rejuvenation step (to smooth particles in the parameter space). Just as in the sequential importance resampling algorithm, this step is implemented when the ESS drops below a threshold, ζN_θ, for $\zeta \in (0, 1]$. Here, as the aim is to avoid degeneracy in the parameter space, as opposed to the state space, one can use a PM-MCMC kernel to rejuvenate particles. The PM-MCMC rejuvenation aims to shift particles, which already correspond to values with posterior support, around the parameter space such that there are many unique samples from the parameter space which have posterior support. The proposals of the PM-MCMC steps can also be informed by the particles; for example, fitting a Gaussian mixture to the current set of particles, which allows parameters to be sampled in parts of the parameter space which are known to have reasonable posterior support. Although the rejuvenation step avoids degeneracy of parameters and state particles, it does not decrease error in estimates of the evidence. Further, it is difficult to determine how many PM-MCMC steps are necessary to shift parameter particles around the state space sufficiently. A main contribution of this thesis is a modified SMC-squared algorithm with a rejuvenation step that allows evidence estimates to be updated to a given accuracy and incorporates stopping criterion that allows for sufficient iterations to be made to avoid particle degeneracy.

```
Inputs: The number of particles, $N_\theta$, the number of state particles, $N_x$, the number of PM-MCMC steps for each particle in rejuvenation step, $R$, and observed data, $y_{1:T}$.
Outputs: The evidence, $\hat{p}(y_{1:T})$ and weighted particles, $\left\{\theta^i, W^i\right\}_{i=1}^{N_\theta}$.
Initialization:
for $i = 1, \ldots, N_\theta$ do
    Generate an initial parameter set $\theta^i \sim p(\theta)$;
    Set initial weight $W^i = 1/N_\theta$;
    For $j = 1, \ldots, N_x$ generate initial state $x_0^j \sim q_0(\cdot)$;
    Resample initial states, where index $j$ is sampled in proportion to $\frac{p(x_0^j)}{q_0(x_0^j)}$;
end
for $t = 1 : T$ do
    for $i = 1 : N_\theta$ do
        Particle filter step:
        Run an iteration of the particle filter with inputs: $\theta^i$, $\left[\{x_{t-1}\}_{j=1}^{N_x}\right]^i$ and $y_t$;
        Return outputs: $\left[\{x_t\}_{j=1}^{N_x}\right]^i$ and $\hat{p}\left(y_t|y_{1:t-1}, \theta^i\right)$;
        Update likelihood estimates $\hat{p}(y_{1:t}|\theta^i) = \hat{p}(y_{1:t-1}|\theta^i)\hat{p}\left(y_t|y_{1:t-1}, \theta^i\right)$;
        Update unnormalised weights $w^i = \hat{p}(y_t|y_{1:t-1}, \theta^i)W^i$;
    end
    Calculate normalised weights $\left\{W^i\right\}_{i=1}^{N_\theta}$;
    $\hat{p}(y_{1:t}) = \left(\sum_{i=1}^{N_\theta} w^i\right) \hat{p}(y_{1:t-1})$;
    ESS $= 1/\sum_{i=1}^{N_\theta} \left(W^i\right)^2$;
    if $ESS < \zeta N_\theta$ then
        Resampling:
        for $i = 1 : N_\theta$ do
            Sample index $r$ with probability $W^r$;
            Set $\left\{\theta^i, \left[\{x_t\}_{j=1}^{N_x}\right]^i, W^i\right\} = \left\{\theta^r, \left[\{x_t\}_{j=1}^{N_x}\right]^r, 1/N_\theta\right\}$
        end
        Rejuvenation:
        for $i = 1 : N_\theta$ do
            Run R iterations of PM-MCMC with inputs: $\left[\{x_t\}_{j=1}^{N_x}\right]^i$, $\theta^i$ and $\hat{p}(y_{1:t}|\theta^i)$;
            Return outputs: updated samples of $\left[\{x_t\}_{j=1}^{N_x}\right]^i$, $\theta^i$ and $\hat{p}(y_{1:t}|\theta^i)$.
        end
    end
end
```

Algorithm 5: An SMC-squared algorithm

Chapter 3

Efficient Bayesian Inference for Epidemics

This chapter discusses applications of efficient data-augmented Markov chain Monte Carlo (DA-MCMC) for inferring parameters of epidemic models, given First Few Hundred (FF100) study data. Section 3.1 introduces an efficient implementation of DA-MCMC for inferring parameters of a household SIR model and compares this with an MCMC scheme that uses an approximation of the likelihood based on branching processes. Section 3.2 gives an implementation of DA-MCMC applied to a household SE(2)I(2)R model with a realistic household size distribution. Lastly, Section 3.3 describes a new kind of method dubbed sequential data-augmented Markov chain Monte Carlo (SDA-MCMC), which sequentially performs DA-MCMC in order to improve mixing. This is applied to a homogeneous SIR model and multiple independent outbreaks from an SI(n)R model with unknown shape parameter.

3.1 Data-Augmentation for an SIR Household Model

This section discusses some results from [6], which considers the problem of inferring epidemic parameters of an SIR household model supposing we only observe the total number of infections in each household at a daily resolution. This paper compares a branching process method developed in [61], with a novel, efficient data-augmented MCMC scheme. As such, this section highlights the data-augmented MCMC scheme and shows a comparison of the two methods.

3.1.1 Model and data assumptions

For this study we consider a SIR household model, as described in Section 2.2.3, in a population of M households each of size N. Here we let households be the same size for simplicity, however, these methods could easily be applied to a population of households of various sizes; the next section addresses a model with inhomogeneous household sizes.

We suppose all individuals in the population are initially susceptible, that is, the state of household j is $(s_t, i_t, r_t) = (N, 0, 0)$ for $j = 1, \ldots, M$. At some Uniform$(0, 1)$ distributed time, t_1, an infectious individual is seeded in the population. Without loss of generality the infection can be assumed to be seeded in the first household, that is, the state of the system becomes $(s_t, i_t, r_t) = (N-1, 1, 0)$ for household 1 and $(s_t, i_t, r_t) = (N, 0, 0)$, for households $j = 2, ..., M$. Once infection is seeded the SIR household dynamics progress the spread of the epidemic.

We consider the SIR process as a partially-observed Markov chain where we only observe the cumulative number of infections in each household on each day up to some time, T. That is, we observe $y = \{y_1, \ldots, y_T\}$, where y_t is a vector of length M which counts the total number of infections that have occurred over time $(t-1, t]$ in each household. Figure 3.1 gives a visual representation of the data, red dots correspond to observed cases in each household on each day.

3.1.2 Inference

We augment our data with the transition times and states to give $x_{(t_1,T]}$. Let $m \in \left\{\sum_{t=1}^{T} y_t, ..., 2\sum_{t=1}^{T} y_t\right\}$ be the unknown number of transitions over time $(t_1, T]$. Additionally we augment the data with a classification of missing events, that is, we augment the data by transition labels $\phi \in \{\text{within, between, recovery}\}^m$. This is such that we can construct sets of transition indices, A, B and C, which correspond to within-household infection, between-household infection and recovery events respectively (excluding the first infection event). Let t_j denote the time of the jth transition and $h(j)$ denote the household whose state changes at time t_j. In writing down the

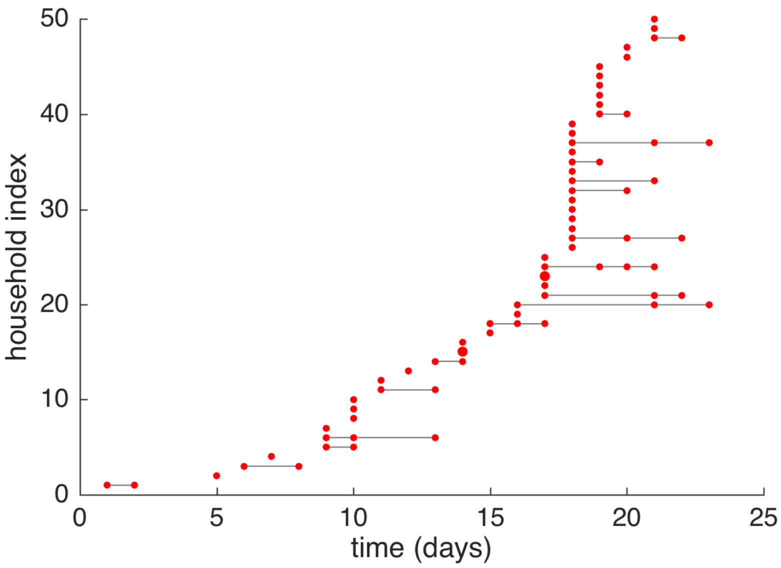

Figure 3.1: A single realisation showing the times of symptom onset, binned into days, in the first 50 infected households at the beginning of an epidemic outbreak. The size of points corresponds to the number of infections on that day. The lines provide a visual reference to link infections within the same household.

expression for the augmented likelihood function we adopt the convention that all quantities s_t^h, i_t^h and I_t are the number of susceptible individuals in household h, the number of infectious individuals in household h and the number of infectious individuals in the population at time t respectively. The augmented likelihood is the joint density of y and $x_{(t_1,T]}$ conditional on the between household infection rate, α, within household infection rate, β, recovery rate, γ, and seeding time, t_1. This is expressed as

$$f\left(y, x_{(t_1,T]}, \phi | \alpha, \beta, \gamma, t_1\right) = 1_{\{y, x_{(t_1,T]}, \phi\}} \prod_{j\in A} \frac{\beta s^{h(j)}_{t_{j-1}} i^{h(j)}_{t_{j-1}}}{N-1} \prod_{k\in B} \frac{\alpha s^{h(k)}_{t_{k-1}} \left(I_{t_{k-1}} - i^{h(k)}_{t_{k-1}}\right)}{N(M-1)} \prod_{l\in C} \gamma i^{h(l)}_{t_{l-1}}$$

$$\times \exp\left\{ -\int_{t_1}^{T} \gamma I_t + \sum_{h=1}^{M} \frac{\beta s_t^h i_t^h}{N-1} + \frac{\alpha s_t^h \left(I_t - i_t^h\right)}{N(M-1)}\, dt \right\}, \quad (3.1)$$

where $1_{\{y, x_{(t_1,T]}, \phi\}}$ denotes the indicator function which takes value 1 if $x_{(t_1,T]}$ with transitions according to labels ϕ could give rise to data set y. Note that the household model makes no distinction between the state change associated with a within or between household infection, however, augmenting the data with labels ϕ allows us to distinguish the sets A, B and C. This allows the augmented likelihood to have a product form which permits the use of conjugate priors. The use of conjugate priors allows the algorithm to run efficiently via Gibbs sampling in the α, β and γ dimensions.

We suppose that α, β/γ and $1/\gamma$ have independent Uniform(0.05,1), Uniform(0.25,4) and Uniform(0.25,7) priors respectively. This implies that γ has prior $p(\gamma) = 1/6.75\gamma^2$ and β has conditional prior $p(\beta|\gamma) = 1/3.75\gamma$. For simplicity let $t_{m+1} := T$. From Equation (3.1), the conditional distributions of α, β and γ are

$$\alpha|\beta, \gamma, x_{(t_1,T]}, \phi \sim \text{Gamma}\left(|B| + 1, \sum_{j=1}^{m}\sum_{h=1}^{M} \frac{s_{t_j}^h \left(I_{t_j} - i_{t_j}^h\right)(t_{j+1} - t_j)}{N(M-1)}\right)_{[0.05,1]},$$

$$\beta|\alpha, \gamma, x_{(t_1,T]}, \phi \sim \text{Gamma}\left(|A| + 1, \sum_{j=1}^{m}\sum_{h=1}^{M} \frac{s_{t_j}^h i_{t_j}^h (t_{j+1} - t_j)}{N-1}\right)_{[0.25\gamma,4\gamma]}$$

and

$$\gamma|\alpha, \beta, x_{(t_1,T]}, \phi \sim \text{Gamma}\left(|C| - 2, \sum_{j=1}^{m} I_{t_j}(t_{j+1} - t_j)\right)_{[1/7,4]},$$

where the subscripts denote the support of the distributions. The conditional distribution of γ has shape parameter $|C| - 2$ due to the γ terms in the denominators of the prior distributions; this means that if $|C| \leq 2$ Hastings steps may be needed to sample γ.

As the first event, an infection at time t_1, is generated at some Uniform$(0, 1)$ distributed time, the prior distribution of t_1 is Uniform$(0, 1)$ and hence the conditional distribution is given by

$$f\left(t_1|\alpha, \beta, \gamma, x_{[t_2,T]}, \phi\right) = \frac{(\alpha + \beta + \gamma)e^{(\alpha+\beta+\gamma)t_1}}{e^{(\alpha+\beta+\gamma)\min\{1,t_2\}} - 1}, \qquad \text{for } t_1 \in (0, \min\{1, t_2\})$$

which can be sampled efficiently by inverse transform sampling [62].

Lastly we are left to find a way of sampling from $f\left(x_{(t_1,T]}, \phi|\beta, \alpha, \gamma, t_1, y\right)$. We do this using a Hastings algorithm with five possible moves, described below. We abbreviate $f\left(y, x_{(t_1,T]}, \phi|\beta, \alpha, \gamma, t_1\right)$ by $f\left(x_{(t_1,T]}, \phi\right)$ for simplicity and denote our candidate values for $x_{(t_1,T]}$ and ϕ by $x^*_{(t_1,T]}$ and ϕ^* respectively. To sample from $f\left(x_{(t_1,T]}|\beta, \alpha, \gamma, t_1, y\right)$ we randomly choose from the following five kinds of moves according to an arbitrary probability mass function with non-zero components, $\{q_1, \dots, q_5\}$:

(i) Uniformly randomly select an infection time, t_j, and choose a candidate Uniform$(\lfloor t_j \rfloor, \lceil t_j \rceil)$

distributed infection time. The candidate is accepted with probability

$$\min\left\{\frac{f\left(x^*_{(t_1,T]},\phi^*\right)}{f\left(x_{(t_1,T]},\phi\right)},1\right\}.$$

(ii) Uniformly randomly select an infection event and change its type, ϕ^j, from between to within household infection or vice versa. The candidate is accepted with probability

$$\min\left\{\frac{f\left(x_{(t_1,T]},\phi^*\right)}{f\left(x_{(t_1,T]},\phi\right)},1\right\}.$$

(iii) Uniformly randomly select a recovery time, t_j, and choose a candidate Uniform(t_k,T) distributed recovery time, where t_k is the time of the first infection within the household. The candidate is accepted with probability

$$\min\left\{\frac{f\left(x^*_{(t_1,T]},\phi^*\right)}{f\left(x_{(t_1,T]},\phi\right)},1\right\}.$$

(iv) Insert a Uniform(t_k,T) distributed recovery time in a randomly chosen household. Let $\hat{M}$ be the number of households infected by time T. The candidate is accepted with probability

$$\min\left\{\frac{f\left(x^*_{(t_1,T]},\phi^*\right)\hat{M}(T-t_k)q_5}{f\left(x_{(t_1,T]},\phi\right)(|C|+1)q_4},1\right\}.$$

(v) Uniformly randomly select and remove a recovery event with probability

$$\min\left\{\frac{f\left(x^*_{(t_1,T]},\phi^*\right)|C|q_4}{f\left(x_{(t_1,T]},\phi\right)\hat{M}(T-t_k)q_5},1\right\}.$$

For the study presented here we implemented DA-MCMC by proposing moves (i)-(v) with probabilities $q_1 = q_2 = 0.05$ and $q_3 = q_4 = q_5 = 0.3$. Each iteration of the DA-MCMC algorithm is comprised of Gibbs samples of α, β, γ and t_1 followed by a Hastings step for $x_{(t_1,T]}$ and ϕ as per (i)-(v). The stationary distribution of the samples is the joint posterior distribution of $x_{(t_1,T]}, \phi, \alpha, \beta$ and γ, where consecutive samples are highly correlated. The marginal over the parameters is simply obtained by ignoring the samples of $x_{(t_1,T]}$ and ϕ. That is, one can run the DA-MCMC algorithm for many iterations and consider only samples of α, β, γ to estimate the posterior distribution, $f(\alpha, \beta, \gamma|y)$.

We also implement an approximate algorithm called the branching process algorithm (BPA), so that we can compare the accuracy and run time to the DA-MCMC method. The BPA is a Metropolis-Hastings algorithm where the likelihood function is approximated under branching process assumptions. That is, the assumption that infected households act independently after their initial infection, and hence infection at a household level occurs as a branching process. This assumption is reasonable if the number of households in the population is large and the proportion of susceptible individuals in the population is close to one, this is because the between rate at which new households are infected becomes $\frac{\alpha S_t I_t}{NM-1} \approx \alpha I_t$ and the rate at which infection within an already infectious household occurs is $\frac{\alpha s_t(I_t - i_t)}{MN-1} \approx 0$, as $I_t, s_t << MN$. The likelihood function for BPA is made up of components related to the number of newly infected households over each day and components related to within-household infections over each day. The distribution of the number of newly infected households over day t is approximated by Poisson($E[\int_{t-1}^{t} \alpha I_s ds]$), where this expectation is calculated via a modified forwards-backwards algorithm and a convolution method. The within-household infection component is made up of simple matrix exponential calculations, which are calculated in the process of computing $E[\int_{t-1}^{t} \alpha I_s ds]$. For more details on the BPA see [6].

3.1.3 Results

We compare the DA-MCMC algorithms performance with the branching process algorithm (BPA) on 50 simulated data sets with true parameter values $(\alpha, \beta, \gamma) = (0.32, 0.4, 1/3)$ and $50,000$ households of size $N = 3$ (the average household size in Australia is estimated to be 2.6 [63]). We only include data sets where at least 400 households became infected. The parameters are chosen such that the average infectious period, $1/\gamma$, is three days (this is a typical infectious period for influenza), the average number of secondary infections from a primary infectious individual in a large household of susceptibles, $R_0 = \beta/\gamma$, is 1.2 and the average number of secondary households infected by a primary infectious household in an otherwise susceptible population, R_*, is approximately 1.8. The simulated data sets are from the full stochastic household model, whereas BPA performs inference under a branching process assumption, that is, it assumes that households are conditionally independent after their initial infection.

Each algorithm is run at various stages of the epidemic in order to show how the posterior distributions converge as more households become infected; the inference for each simulation is run after 50, 100, 200, 300 and 400 households become infected. For the BPA, for each simulation, at each stage of the epidemic, 10^5 MCMC samples are obtained with a burn-in of 1000 iterations. For the DA-MCMC algorithm, for each simulation, at each stage of the epidemic, 2.5×10^6 iterations are run with an additional burn-in of 10^6 iterations and results are thinned to a sample of size 2.5×10^5. These numbers of iterations were chosen so that each sample had approximately the same multivariate effective sample size (mESS), where mESS is the number of approximately independent samples in a set of correlated samples from a multivariate distribution [64]. More iterations are needed for the DA-MCMC as the mixing is slower, the samples were thinned for data storage reasons.

Our results are presented in terms of: maximum a posteriori (MAP) estimates of model parameters in Figure 3.2; MAP estimates of the household reproduction number, R^*, and early growth rate, r, in Figure 3.3; and, posterior distributions of R^* and r for a single simulated data set over time in Figure 3.4. Means and standard deviations for MAP estimates are given explicitly in Table 3.1.

	α	β	γ	R_*	r
True Values	0.32	0.4	0.333	1.803	0.190
BPA	0.345(0.028)	0.387(0.027)	0.310(0.034)	2.091(0.210)	0.237(0.028)
DA-MCMC	0.312(0.022)	0.392(0.027)	0.314(0.032)	1.839(0.168)	0.191(0.024)
(BPA)-(DA-MCMC)	0.033(0.012)	-0.005(0.012)	-0.004(0.020)	0.2519(0.066)	0.046(0.008)

Table 3.1: Means and standard deviations of maximum a posteriori (MAP) estimates of $(\alpha, \beta, \gamma, R_*, r)$. The means and standard deviations of the 50 MAP estimates based upon data with 400 infected households for each parameter is shown in the form mean(standard deviation) for the BPA and DA-MCMC methods. The last row shows the difference in the mean and standard deviation between the two methods.

From Figure 3.2, for both methods, we observe that MAP estimates begin nega-

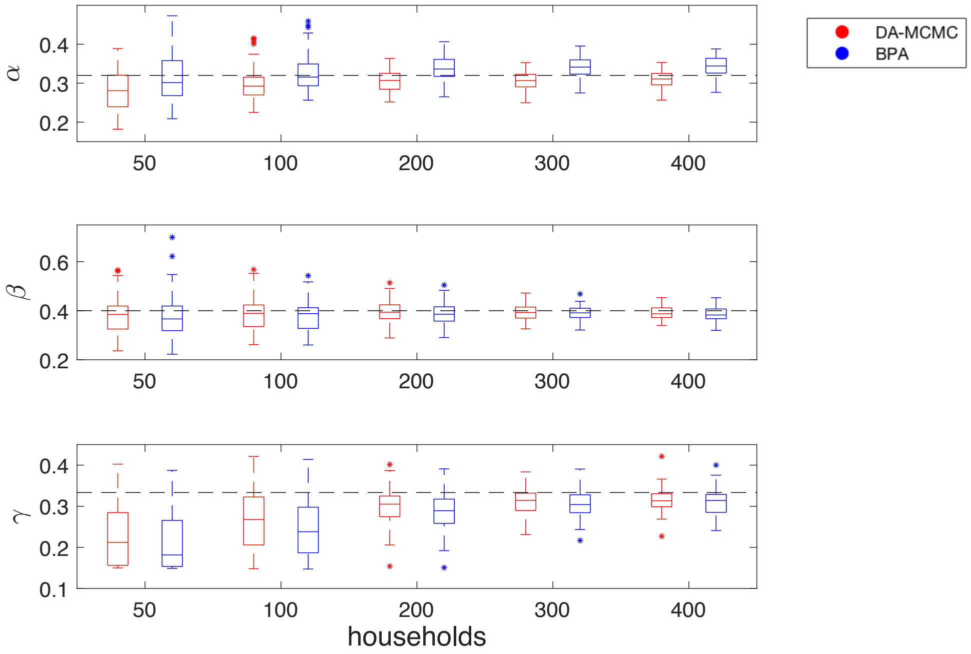

Figure 3.2: Boxplots of maximum a posteriori (MAP) estimates of (α, β, γ) from 50 simulations. Red and Blue boxes correspond to results from 2.5×10^6 iterations, thinned to 2.5×10^5 samples, of the DA-MCMC algorithm and 10^5 iterations of the BPA, respectively. MAP's are calculated from 3-dimensional kernel density estimates. The pairs of boxes from left to right are MAP's from inference based upon data with 50, 100, 200, 300 and 400 infected households. Black dashed lines indicate the true parameter values at $(\alpha, \beta, \gamma) = (0.32, 0.4, 1/3)$.

tively biased for all parameters and converge towards fixed points as more data are obtained. The median of the MAPs of the BPA method for β and γ are lower than that of the DA-MCMC method, whereas the median of the MAPs of α are higher. The boxes associated with β and γ for each method are overlapping, whereas the boxes associated with α are biased higher for the BPA method when data are based upon 300 and 400 infected households. In Figure 3.3, we observe that the boxes of the MAP estimates converge to the true values of R_* and r for the DA-MCMC method, whereas they are biased above the true value for the BPA method. The positive bias in these quantities is due to the overestimation of α by the BPA method. The box plots indicate a general trend that the variability of the MAP estimates decrease as more data are obtained. It should be noted that these box plots do not show the correlation structure

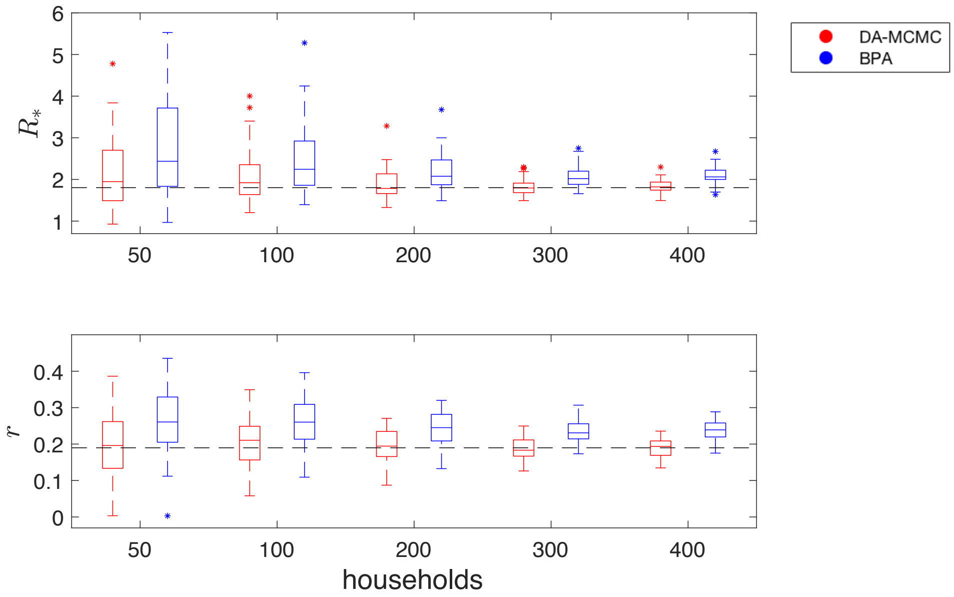

Figure 3.3: Boxplots of maximum a posteriori (MAP) estimates of (R_*, r) from 50 simulations. Red and Blue boxes correspond to results from 2.5×10^6 iterations (thinned to 2.5×10^5) of the DA-MCMC algorithm and 10^5 iterations of the BPA and respectively. MAP's are calculated from 2 dimensional kernel density estimates. The pairs of boxes from left to right are MAP's from inference based upon data with 50, 100, 200, 300 and 400 infected households. Black dotted lines indicate the true parameter values at $(R_*, r) \approx (1.803, 0.190)$.

of the parameters; this is not presented here as the dimension of the parameter space makes the correlation structure difficult to display. In Figure 3.4 the posterior densities of R_* and r appear similar between the two methods, although the bias of the BPA is clear.

For both methods the variability in the posterior distribution is observed to decrease in a similar way as more households are infected. Table 3.1 shows that when inference is run after 400 households are infected, the mean of the MAPs of both methods lie within a standard deviation from the true values of (α, β, γ). We also find that the average MAP estimates of β/γ is found to be 1.2484 in both methods; this excellent agreement at the household level indicates that the branching process is an appropriate approximation for the full household epidemic process. Out of the two methods, only the means of the MAPs from the DA-MCMC method lie within a standard deviation

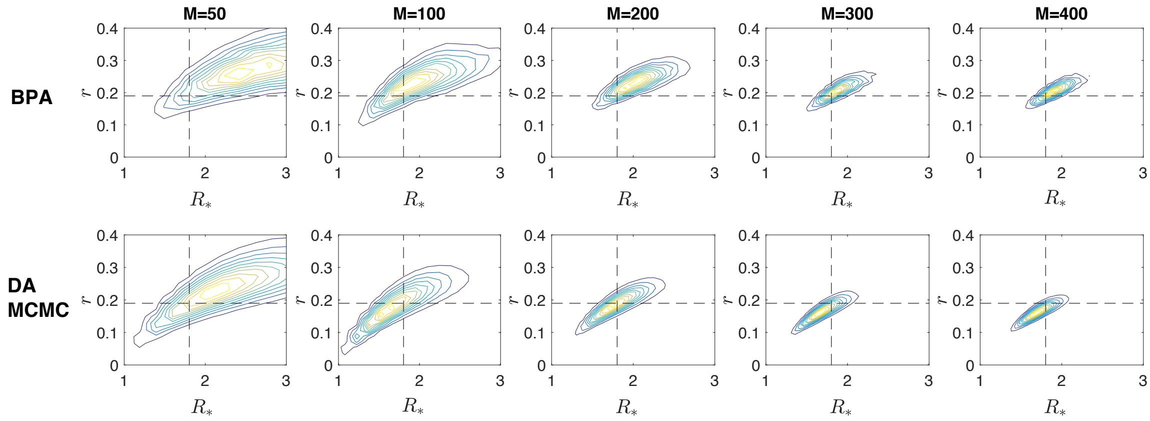

Figure 3.4: Kernel density estimates of R_* and r from BPA (top) and from DA-MCMC (bottom). The BPA results are from 10^5 MCMC samples with 500 iterations of burn-in. The DA-MCMC results are based upon 5×10^6 samples with 10^6 iterations of burn-in. True parameter values, $(R_*, r) \approx (1.803, 0.190)$, are shown at the intersection of the black dashed lines. Inference was run after 50, 100, 200 and 300 households were observed to be infectious.

of the true values of R_* and r. As both methods were run on the same simulations we can compare the difference of MAPs from the two methods, this is given in the final row of Table 3.1. The difference of the MAPs for β and γ lie within a standard deviation of 0, the difference for α, R_* and r are in excess of 2.5 standard deviations from 0. This indicates that the BPA method leads to a significantly different answer, in terms of α, R_* and r compared to exact methods. On average we saw a 7.8%, 16.0% and 24.7% positive error in α, R_* and r respectively.

The efficiency of the two algorithms cannot be compared directly in terms of iterations per time, as samples from the DA-MCMC are more highly correlated than samples from the BPA [65]. Hence, the algorithms are compared in terms of their mESS per hour. Figure 3.5 shows that the DA-MCMC is initially much more efficient than the BPA algorithm, however it scales poorly as more data are obtained and is less efficient than the BPA after 200 households are infected. The efficiency of the BPA algorithm appears to be highly left skewed, as there were some outlying simulations that were much less efficient than the others. These outliers were still more efficient when using the BPA method when inference is based on 400 infected households. Note, the mESS of the BPA and DA-MCMC had an average of 3366 and 4138 when inference is based on 400 infected households, so even though the results are based upon different sample sizes, the mESS are comparable and sufficiently large. On average the DA-MCMC algorithm with 50, 100, 200, 300 and 400 infected households will take 0.06, 0.19, 1.45, 5.14 and 13.72 hours respectively to obtain an effective sample size of 3000, whereas the BPA algorithm can do the same in 0.49, 0.53, 0.70, 0.94 and 1.24 hours respectively. The BPA method is twice as efficient as the DA-MCMC algorithm by the time 200 households are infected and it is more than 11 times as efficient when 400 households are infected.

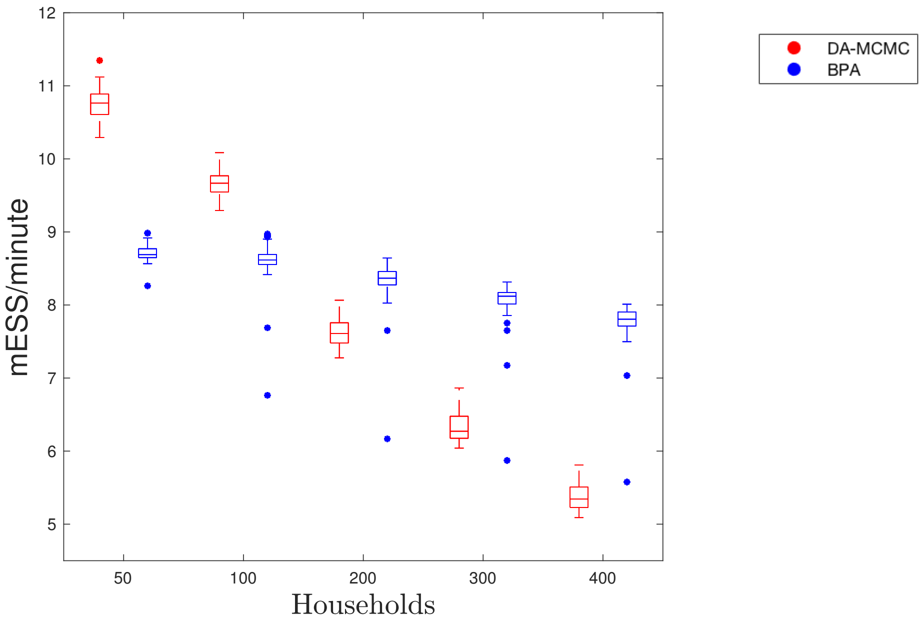

Figure 3.5: Boxplots of the efficiency of each method against the number of infected households. Here efficiency is presented in terms of log mESS per hour. These estimates are based upon running each algorithm for 50 simulations with 50, 100, 200, 300 and 400 infected households.

In summary, we find that DA-MCMC, as an exact algorithm, leads to estimates with lower bias than the approximate BPA. Further, these estimates have lower variance than the BPA algorithm. For small data sets the DA-MCMC algorithm is also more efficient, but it scales poorly as the data size increases. This indicates that for larger data sets, if given limited computational budget, the asymptotically-exact DA-MCMC method may need to be swapped for the approximate BPA.

3.2 Inference for an SE(n_1)I(n_2)R Household Model

This section shows an efficient way to implement DA-MCMC to infer parameters of an SE(n_1)I(n_2)R model with a distribution of households from daily resolution case data. The inference method is based on ideas from the previous section, but also uses a labelled representation of the model to make efficient proposals and allows events to be proposed after the time of the last observation (as discussed in Example 2, Section 2.3.3).

3.2.1 Model

We consider an SE(n_1)I(n_2)R household model, as described in Section 2.2.3, with Erlang(n_1, λ_1) latent periods and Erlang(n_2, λ_2) infectious periods such that $\sigma = \lambda_1/n_1$ and $\gamma = \lambda_2/n_2$. For this study we consider a population of M households of various sizes $N_1, \dots, N_M$, with a total population size $P := \sum_{h=1}^{M} N_h$. In our example we apply DA-MCMC to a simulated outbreak in a population of households of up to six individuals with a household size distribution similar to that of Adelaide, informed by the Australian Bureau of Statistics 2016 report [63]. As in Example 2, from Section 2.3.3, we suppose each individual is labelled, such that we can propose changes to the hidden process based upon latent and infectious periods, as opposed to exposure and infection times.

We suppose all individuals in the population are initially susceptible, that is, the state of household j is $(s_t, e_t, i_t, r_t) = (N_j, 0, 0, 0)$ for $j = 1, \dots, M$. At some Uniform$(0, 1)$ distributed time, t_1, an infectious individual is seeded in the population uniformly randomly. That is, household k moves to state $(s_t, e_t, i_t, r_t) = (N_k - 1, 0, 1, 0)_k$ with probability N_k/P and all other households remain in state $(s_t, e_t, i_t, r_t) = (N_j, 0, 0, 0)$. Unlike in the previous section, the infection cannot be seeded in the first household; households have different sizes, so infection is more likely to be seeded in larger households. Once infection is seeded the SE(n_1)I(n_2)R household dynamics progress the spread of the epidemic.

We consider the SE(n_1)I(n_2)R process as a partially-observed Markov chain where we observe only the cumulative number of infections in each household on each day up to some time, T. That is, as in the previous section, we observe $y = \{y_1, \dots, y_T\}$, where y_t is a vector of length M which counts the total number of infections that have occurred over time $(t-1, t]$ in each household.

3.2.2 Inference

As in the previous section we augment our data with the transition times and states to give $x_{(t_1,T]}$. Let $m \in \{\sum_{t=1}^{T} y_t, ..., 2\sum_{t=1}^{T} y_t\}$ be the unknown number of transitions over time $(t_1, T]$. Additionally we augment the data with a classification of missing events, that is, we augment the data by transition labels $\phi \in \{\text{within, between, infection, recovery}\}^m$. This is such that we can construct sets of transition indices, A, B, C and D which correspond to within-household exposure, between-household exposure, infection and recovery events respectively (excluding the first infection event). Let t_j denote the time of the jth transition and $h(j)$ denote the household whose state changes at time t_j. We denote the number of infectious individuals in household h and the number of infectious individuals in the population at time t by i_t^h and I_t respectively. The augmented likelihood is the joint density of y, ϕ and $x_{[t_1,T]}$ conditional on α, β, γ, σ and t_1. This is expressed as

$$\begin{aligned}
f(y, x_{[t_1,T]}, \phi|\alpha, \beta, \sigma, \gamma, t_1) = & \mathbb{1}_{\{y, x_{[t_1,T]}, \phi\}} \prod_{j\in A} \frac{\beta i^{h(j)}_{t_{j-1}}}{N_{h(j)}-1} \prod_{k\in B} \frac{\alpha\left(I^{h(k)}_{t_{k-1}} - i^{h(k)}_{t_{k-1}}\right)}{P - N_{h(k)}} \\
& \times \exp\left\{-\int_{t_1}^{T} \sum_{h=1}^{M} \frac{\beta s_t^h i_t^h}{N_h - 1} + \frac{\alpha s_t^h \left(I_t - i_t^h\right)}{P - N_h}\, dt\right\} \\
& \times \left(\frac{\lambda_1^{n_1}}{(n_1-1)!}\right)^{|A|+|B|+1} \left(\prod_{l=1}^{|A|+|B|+1} \delta_l^{n_1-1}\right) \exp\left\{-\lambda_1 \sum_{l=1}^{|A|+|B|} \delta_l\right\} \\
& \times \left(\frac{\lambda_2^{n_2}}{(n_2-1)!}\right)^{|D|} \left(\prod_{r=1}^{|D|} \Delta_r^{n_2-1}\right) \exp\left\{-\lambda_2 \sum_{r=1}^{|D|} \Delta_r\right\} \qquad (3.2)
\end{aligned}$$

where I_t is the number of infectious individuals in the population, and s_t^h and i_t^h are the number of susceptible and infectious individuals in household h, at time t. We suppose that β and α have Uniform$(0, 10)$ priors and γ and σ have InverseUniform(0.25,7) priors. Hence, from Equation (3.2), their conditional distributions are

$$\alpha|\beta, \sigma, \gamma, x_{[t_1,T]}, \phi, y \sim \text{Gamma}\left(|B|+1, \sum_{j=1}^{m}\sum_{h=1}^{M} \frac{s_{t_j}^h \left(I_{t_j} - i_{t_j}^h\right)(t_{j+1} - t_j)}{P - N_h}\right)_{[0,10]},$$

$$\beta|\alpha, \sigma, \gamma, x_{[t_1,T]}, \phi, y \sim \text{Gamma}\left(|A|+1, \sum_{j=1}^{m}\sum_{h=1}^{\hat{M}} \frac{s_{t_j}^h i_{t_j}^h (t_{j+1} - t_j)}{N_h - 1}\right)_{[0,10]},$$

$$\lambda_1|\alpha,\gamma,\beta,x_{[t_1,T]},\phi,y \sim \text{Gamma}\left(n_1(|A|+|B|+1)+1-2,\sum_{l=1}^{|A|+|B|+1}\delta_l\right)_{[n_1/7,4n_1]},$$

$$\lambda_2|\alpha,\sigma,\beta,x_{[t_1,T]},\phi,y \sim \text{Gamma}\left(n_2|D|+1-2,\sum_{r=1}^{|D|}\Delta_r\right)_{[n_2/7,4n_2]},$$

and as infection is seeded in the population at a Uniform$(0,1)$ distributed time the conditional distribution is given by

$$f(t_1|\gamma,\beta,\alpha,x_{[t_2,T]},\phi)=\frac{(\alpha+\beta+\lambda_2)e^{(\alpha+\beta+\lambda_2)t_1}}{e^{(\alpha+\beta+\gamma)\min\{1,t_2\}}-1}, \qquad \text{for } t_1\in(0,\min\{1,t_2\})$$

which can be sampled efficiently by inverse transform sampling.

We use a Hastings step to sample from $f(y,x_{[t_1,T]},\phi|\beta,\sigma,\alpha,\gamma,t_1)$. For this model we choose one of the following seven moves according to distribution $\{p_1,\ldots,p_7\}$:

(i) Uniformly randomly select an individual that became infectious at time t_j and choose a candidate Uniform$(\lfloor t_j\rfloor,\lceil t_j\rceil)$ distributed infectiousness time, t_j^*. Randomly choose a Gamma$(n_1,\lambda_1)_{[0,t_j^*-t_1]}$ distributed candidate exposed period δ_l^*; denote the pdf of the candidate distribution by $g_1(\delta_l^*,t_j^*)$. Randomly choose a Gamma(n_2,λ_2) infectious period Δ_r^*; denote the probability density function of the candidate distribution by $g_2(\Delta_r^*)$. The new point is accepted with probability

$$\min\left\{\frac{f(x^*_{[t_1,T]},\phi^*)g_1(\delta_l,t_j)g_2(\Delta_r)}{f(x_{[t_1,T]},\phi)g_1(\delta_l^*,t_j^*)g_2(\Delta_r^*)},1\right\}.$$

(ii) Uniformly randomly select an exposure event and change its type, ϕ^j, from between-household to within-household exposure or vice versa. The new point is accepted with probability

$$\min\left\{\frac{f(x_{[t_1,T]},\phi^*)}{f(x_{[t_1,T]},\phi)},1\right\}.$$

(iii) Uniformly randomly select an individual that is exposed at time T. Randomly choose a candidate exposure time with distribution $T-\text{Exponential}(\frac{\lambda_1}{n_1})_{[0,T-t_1]}$, and randomly choose a candidate infection time with distribution $T+\text{Exponential}(\frac{\lambda_1}{n_1})$ to give candidate latent period δ_j^*. The new point is accepted with probability

$$\min\left\{\frac{f(x^*_{[t_1,T]},\phi^*)}{f(x_{[t_1,T]},\phi)}e^{-\frac{\lambda_1}{n_1}(\delta_j-\delta_j^*)},1\right\}.$$

(iv) Uniformly randomly select an individual who is susceptible at time T in an infectious household and choose a candidate $T-\text{Exponential}(\frac{\lambda_1}{n_1})_{[0,T-t_1]}$ distributed exposure time and corresponding $T+\text{Exponential}(\frac{\lambda_1}{n_1})$ distributed infection time to give candidate exposed period, δ_j^*. Let the exposure be of either type with probability 1/2. Denote the number of susceptible individuals in infectious households by $\hat{S}$ and note that the number of exposed individuals in the population at time T is $|A|+|B|-|D|+1$. The new point is accepted with probability

$$\min\left\{\frac{f(x^*_{[t_1,T]},\phi^*)}{f(x_{[t_1,T]},\phi)}\times\frac{2\hat{S}\left(1-e^{-\frac{\lambda_1}{n_1}(T-t_1)}\right)}{p_4\left(\frac{\lambda_1}{n_1}\right)^2 e^{-\frac{\lambda_1}{n_1}\delta_j^*}}\times\frac{p_6}{|A|+|B|-|D|+2},1\right\},$$

this probability is expressed in the form: likelihood ratio × the reciprocal of the proposal density × the proposal density associated with moving from the candidate state back to the current state.

(v) Uniformly randomly select an individual who is susceptible at time T in non-infectious household and choose a candidate $T-\text{Exponential}(\frac{\lambda_1}{n_1})_{[0,T-t_1]}$ distributed exposure time and corresponding $T+\text{Exponential}(\frac{\lambda_1}{n_1})$ distributed infection time to give candidate exposed period, δ_j^*. Note that the number of susceptible individuals in the population at time T is $P-|A|-|B|-1$. The exposure type must be a between household exposure. The new point is accepted with probability

$$\min\left\{\frac{f(x^*_{[t_1,T]},\phi^*)}{f(x_{[t_1,T]},\phi)}\times\frac{(P-|A|-|B|-1-\hat{S})\left(1-e^{-\frac{\lambda_1}{n_1}(T-t_1)}\right)}{p_5\left(\frac{\lambda_1}{n_1}\right)^2 e^{-\frac{\lambda_1}{n_1}\delta_j^*}}\times\frac{p_6}{|A|+|B|-|D|+2},1\right\}.$$

(vi) Uniformly randomly select an individual who is exposed at time T and remove the corresponding exposure and infection time. If the individual is in an infectious household accept the proposal with probability

$$\min\left\{\frac{f(x^*_{[t_1,T]},\phi^*)}{f(x_{[t_1,T]},\phi)}\times\frac{|A|+|B|-|D|+1}{p_6}\times\frac{p_4\left(\frac{\lambda_1}{n_1}\right)^2 e^{-\frac{\lambda_1}{n_1}\delta_j^*}}{2(\hat{S}+1)\left(1-e^{-\frac{\lambda_1}{n_1}(T-t_1)}\right)},1\right\},$$

else accept with probability

$$\min\left\{\frac{f(x^*_{[t_1,T]},\phi^*)}{f(x_{[t_1,T]},\phi)}\times\frac{|A|+|B|-|D|+1}{p_6}\times\frac{p_5\left(\frac{\lambda_1}{n_1}\right)^2 e^{-\frac{\lambda_1}{n_1}\delta_j^*}}{(P-|A|-|B|-\hat{S})\left(1-e^{-\frac{\lambda_1}{n_1}(T-t_1)}\right)},1\right\}.$$

(vii) For the first infected individual choose a Gamma(n_2,λ_2) infectious period, Δ_r^* and accept the move with probability

$$\min\left\{\frac{f(x^*_{[t_1,T]},\phi^*)g_2(\Delta_r)}{f(x_{[t_1,T]},\phi)g_2(\Delta^*_r)},1\right\}.$$

3.2.3 Results

This method was applied to a population of households with a household distribution informed by the household distribution of Adelaide, Australia, as per the Australian Bureau of Statistics 2016 report [63]. This report estimates the number of households in Adelaide, Australia, with one to five individuals and the number of houses of more than five individuals, this is shown in Figure 3.6. Hence, we apply the method to a population with the same household distribution, however, as we have no further information on households larger than five individuals, we suppose these all contain six individuals. The shape parameters of the exposed and infectious periods were fixed

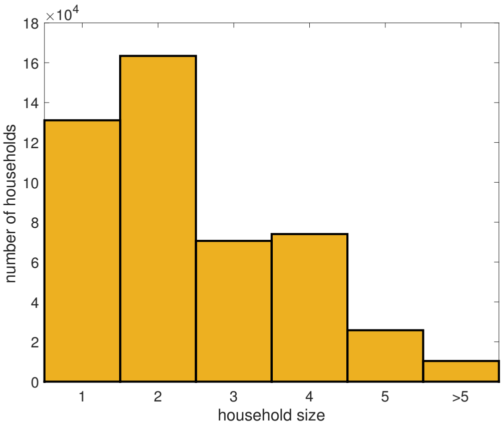

Figure 3.6: The household size distribution of Adelaide, Australia, according to the Australian Bureau of Statistics 2016 report.

at $n_1 = n_2 = 2$, this allows for these periods to have a mode that is not centred at 0 without being too restrictive on their variance. Samples are obtained by running 20 independent chains of DA-MCMC for three days with 10^6 iterations of burn-in, these are thinned by a factor of 10 for data storage reasons. The priors considered here were less informative than in Section 3.1, however, the data seem to have been able to infer each parameter to reasonable accuracy (see Figure 3.7). Further, we are able to infer epidemiological parameters of interest to reasonable accuracy, in particular, the household reproduction number and early growth rate.

A major limitation of this approach is that all cases are assumed to be observed, whereas in reality there may be multiple unobserved households and possibly unobserved cases in observed households. Further, a long burn-in period meant that DA-MCMC had to run for days with multiple independent chains to obtain a reasonable sample from the posterior distribution. The scale of days can make the approach difficult in terms of being practically useful for FF100 studies. Further, the mixing scales poorly with the length of the time series and number of observations, so while it may be possible to apply this with 200 infected households, it may be prohibitively slow for 300 infected households.

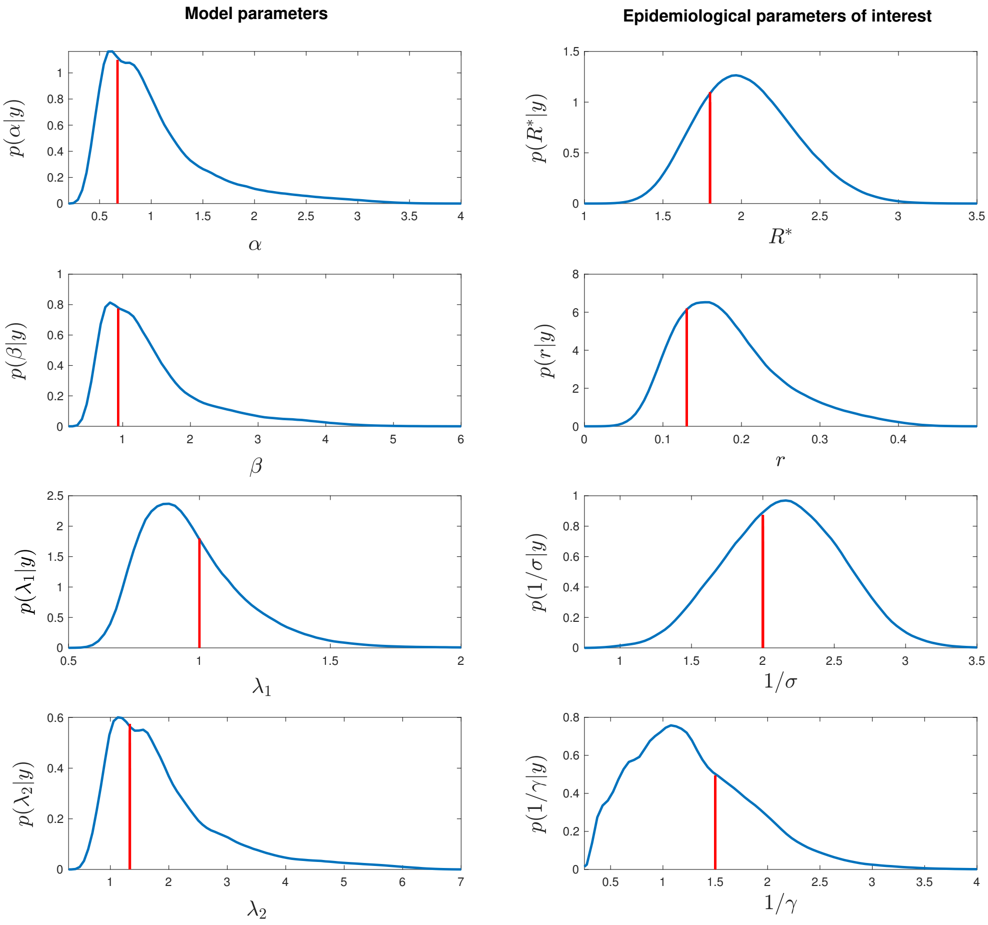

Figure 3.7: Marginal kernel density estimates from DA-MCMC inference based on a simulated data set of 200 infected households from an SE(2)I(2)R model with the household distribution of Adelaide, Australia. The data set contained 322 infectious cases over 30 days. Panels on the left give marginal kernel density estimates of model parameters, whereas panels on the left give marginal kernel density estimates of epidemiological parameters of interest. The DA-MCMC is based on 20 independent chains with 10^6 iterations of burn-in run over three days. The samples were thinned by a factor of 10 to give a total of 1.73×10^6 samples. True parameter values, $(1/\gamma, 1/\sigma, \beta, \alpha) = (1.5, 2, 0.933, 0.6)$.

3.3 Sequential DA-MCMC

Sequential Monte Carlo methods are able to update a posterior distribution iteratively as more data are obtained, rather than beginning the inference anew when new observations are made. Methods such as data-augmented MCMC do not naturally allow for sequential updates, as they rely on imputing a whole sequence of transition times and events. This section presents an approximate Bayesian inference algorithm, dubbed sequential data-augmented Markov chain Monte Carlo (SDA-MCMC), that allows for iterative updates on Markov chains with missing data. This is an approximate inference method as some distributions are approximated by kernel density estimates. The sequential updates of SDA-MCMC allow for faster mixing of the chain than DA-MCMC in some cases, as fewer hidden variables are integrated over in each step. This improved mixing allows SDA-MCMC to perform inference in cases that DA-MCMC cannot, such as for inferring the shape parameter of an SI(n)R model from multiple outbreaks data. Due to limitations of the method, which are discussed in Section 3.3.6, this section gives a only a brief outline of SDA-MCMC and applications where it has been applied effectively.

3.3.1 The aim of SDA-MCMC

Suppose we are to infer parameters, θ, of some Markov chain which is partially observed over time. Assume that if we have data set $y_{1:t}$, observed at times $1, \dots, t$, where each data point, y_s, depends on the state of the underlying Markov chain, x_t, over time interval $(s-1, s]$ such that $\Pr(y_{t+1:T} | x_{[0,t]}, y_{1:t}) = \Pr(y_{t+1:T} | x_t)$. The algorithm described here does not require data to be observed at constant time steps, though it is presented as such here for simplicity.

We present an algorithm that allows inference to be performed sequentially as more data becomes available. More generally, the algorithm presented allows one to update the posterior given a new set of data points, that is, the algorithm need not run every time new data are made available, but rather, can be run in batches. This allows for greater flexibility and possibly allows for more efficient inference.

If we have performed DA-MCMC for the parameter set, θ, based on data $y_{1:t}$ we have stored samples of θ and can construct a kernel density estimate of the posterior distribution. Further, we would have sampled full sequences of $x_{[0,t]}$. The scheme proposed here does not require storage of the entire set of sequences, rather, we store the state of the process at time t, x_t. If we obtain more data such that we have a larger data set, $y_{1:T}$, we want to calculate the target distribution

$$f(\theta, x_T | y_{(1:T)}) = \int f(\theta, x_{[t,T]} | y_{1:T}) \; dx_{[t,T)}.$$

This target distribution can be approximated by using samples from $f(\theta, x_{[t,T]}|y_{1:T})$ and integrating over sample paths, $x_{[t,T)}$. To sample from $f(\theta, x_{[t,T]}|y_{1:T})$ we note that

$$\begin{aligned} f(\theta, x_{[t,T]}|y_{1:T}) &\propto f(y_{t+1:T}, x_{(t,T]}|x_t, \theta) f(\theta, x_t|y_{1:t}), \\ &\propto f(y_{t+1:T}, x_{(t,T]}|x_t, \theta) f(\theta|x_t, y_{1:t}) f(x_t|y_{1:t}). \end{aligned}$$

Here we have used the conditional independence of $y_{t+1:T}, x_T|x_t$ and $y_{1:t}$. Note we already have samples from $f(\theta, x_t|y_{1:t})$ from the DA-MCMC run at time t. So we can create an algorithm from taking samples from $f(\theta, x_t|y_{1:t})$ in order to sample from $f(y_{t+1:T}, x_{(t,T]}|x_t, \theta)$.

3.3.2 Algorithm description

The SDA-MCMC algorithm is comprised of the following four steps.

(i) Initialise by sampling (x_t, θ) from the initial DA-MCMC sample, then choose augmented data $x_{(t,T]}$ such that it agrees with data $y_{t+1:T}$. Denote the discrete and continuous elements in θ as θ^d and θ^c respectively. For each x_t and θ^d, calculate kernel density estimates of θ^c samples to approximate $f(\theta^c|x_t, \theta^d, y_{1:t})$. Calculate probability mass function $f(x_t, \theta^d|y_{1:t})$ as the proportion of samples from the initial DA-MCMC run in each state.

(ii) Sample θ^c by a Metropolis-Hastings step with target density

$$f(\theta^c|x_{[t,T]}, \theta^d, y_{1:T}) \propto f(x_{(t,T]}, y_{t+1:T}|x_t, \theta) f(\theta^c|x_t, \theta^d, y_{1:t}).$$

Often efficient proposals can be made by sampling θ^c with probability proportionate to augmented likelihood, $f(x_{(t,T]}, y_{t+1:T}|x_t, \theta)$; this is a reasonable proposal as it corresponds to a Gibbs step if x_t were known and $y_{t+1:T}$ were our only observations. The proposed parameters, $\bar{\theta}^c$, is accepted with probability

$$\frac{f(\bar{\theta}^c|x_t, \theta^d, y_{1:t})}{f(\theta^c|x_t, \theta^d, y_{1:t})}. \tag{3.3}$$

(iii) Propose new θ^d and changes to $x_{[t,T]}$ and accept via a Metropolis-Hastings step with target density

$$f(y_{t+1:T}, x_{(t,T]}|x_t, \theta) f(\theta^c|x_t, \theta^d, y_{1:t}) f(x_t, \theta^d|y_{1:t}).$$

(iv) Repeat step (ii)-(iii) for many iterations.

3.3.3 SIR model example

We first apply this method to the SIR model, as one of the simplest epidemic models that these methods could be applied to. We assume that data are the total number of infectious cases at a daily resolution up to some time T. The method requires the posterior distribution of the two model parameters and the number of infectious individuals at the end of each day; note we do not need to infer any other states as the number of susceptible individuals are known at the end of each day. That is, SDA-MCMC requires samples from $f(\beta, \gamma, I_t|y_{1:t})$, for some time $t < T$, in order to infer $f(\beta, \gamma|y_{1:T})$ sequentially.

Suppose we have already performed DA-MCMC up to day t, and hence have samples from $f(\beta, \gamma, I_t|y_{1:t})$. To calculate the posterior distribution based upon data up to time $t+1$, $f(\beta, \gamma, I_t|y_{1:t+1})$ we consider the augmented likelihood at time $t+1$. As per usual, assume that we augment the likelihood with hidden states $x_{(t,t+1]}$, where $x_t = (S_t, I_t)$. Denote the transition times by $\{t_1, t_2, \ldots, t_m\}$, let $t_0 = t$, and, sets A and B denote the sets of indices that correspond to infection and recovery events over $(t, t+1]$ respectively. This gives augmented likelihood

$$\begin{aligned} f(x_{(t,t+1]}, y_{t+1}|\beta, \gamma, x_t) &= 1_{\{x_{[t,t+1]}, y_{t+1}\}} f\left(x_{(t,t+1]}|\beta, \gamma, x_t\right) \\ &\propto 1_{\{x_{[t,t+1]}, y_{t+1}\}} \prod_{j\in A} \frac{\beta S_{t_{j-1}} I_{t_{j-1}}}{N-1} \prod_{k\in B} \gamma I_{t_{j-1}} \exp\left\{-\int_t^{t+1} \frac{\beta S_\tau I_\tau}{N-1} + \gamma I_\tau \, d\tau\right\}. \end{aligned}$$

Note that, with respect to β and γ, the augmented likelihood is proportional to $\text{Gamma}(|A|+1, \int_t^{t+1} \frac{S_\tau I_\tau}{N-1} \, d\tau)$ and $\text{Gamma}(|B|+1, \int_t^{t+1} I_\tau \, d\tau)$ respectively. Hence, if β and γ are proposed according to these distributions they cancel out the likelihood term in the acceptance probability, so the acceptance probability is simplified to Equation (3.3). The augmented likelihood and these proposal densities can be used to construct an efficient SDA-MCMC algorithm as given by Algorithm 6. We sample new initial conditions, x_t, by noting that the only unknown state at the end of day t is I_t, so if a new state I_t^* were chosen, $S_{[t,T]}$ is unchanged and the number of infectious individuals over $[t,T]$ becomes $I_{[t,T]}^* = I_{[t,T]} - I_t + I_t^*$, which is only feasible if $I_{[t,T]}^* > 0$. The acceptance probability in Equation (3.4) targets $f(x_{[t,t+1]}|y_{1:t+1}, \beta^{(l)}, \gamma^{(l)})$ as

$$\begin{aligned} f(x_{[t,t+1]}|y_{1:t+1}, \beta, \gamma) &= \frac{f(x_{[t,t+1]}, y_{t+1}|y_{1:t}, \beta, \gamma)}{f(y_{t+1}|y_{1:t}, \beta, \gamma)} \\ &= \frac{f(x_{(t,t+1]}|x_t, \beta, \gamma) f(x_t|y_{1:t}, \beta, \gamma)}{f(y_{t+1}|y_{1:t}, \beta, \gamma)} \\ &= \frac{f(x_{(t,t+1]}|x_t, \beta, \gamma) f(\beta, \gamma|x_t, y_{1:t}) f(x_t|y_{1:t})}{f(y_{t+1}|y_{1:t}, \beta, \gamma) f(\beta, \gamma|y_{1:t})}, \end{aligned}$$

so the Metropolis-Hastings acceptance probability with proposal density $f(x_t|y_{1:t})$ reduces to Equation (3.4).

Figure 3.8 shows the posterior densities calculated by the SDA-MCMC and by a standard DA-MCMC algorithm over time. Inference for the first 13 days was performed by DA-MCMC and inference for days 14, 15 and 16 was performed by both algorithms for comparison. We observe that the posterior distributions appear rather similar; though the kernel density estimates from the SDA-MCMC algorithm appear coarser, this is may be due to parameters being inferred by Hastings steps, rather than by Gibbs sampling. The SDA-MCMC algorithm also, in this case, appears to show a slight reduction in support around the boundary of the posterior distribution. For $t = 14$, 15 and 16, DA-MCMC and SDA-MCMC achieve a multivariate-effective sample size per minute of 1230, 856 and 796, and, 3752, 2021 and 1955, respectively. This shows that the SDA-MCMC in this case is much more efficient than the DA-MCMC algorithm, that is, there is a benefit to sequentially updating the posterior in this way, rather than running inference anew each time. Figure 3.9 shows the distributions $f(I_t|y_{1:t})$ for t=14, 15 and 16; they appear smooth, indicating that there were sufficient samples to estimate $p(x_t|y_{1:t})$ in each iteration.

```
Initialization:

Run DA-MCMC, or SDA-MCMC on data set $y_{1:t}$ to obtain samples from
$f\left(\beta, \gamma, x_t | y_{1:t}\right)$.
For all $x_t$ calculate kernel density estimates of $f\left(\beta, \gamma | x_t, y_{1:t}\right)$ from the samples of $\beta$
and $\gamma$ corresponding to each value of $x_t$.
Empirically estimate $f\left(x_t | y_{1:t}\right)$ as the proportion of DA-MCMC samples with state $x_t$.
Choose initial parameters $(\beta, \gamma)^{(0)}$ and augmented data $x_{[t,t+1]}$, set $x_{t+1}^{(0)} = x_{t+1}$.
Set the number of iterations to some large $K$.

Iterations:
for $l = 1 : K$ do
    Hastings Step for $\beta$ and $\gamma$:
    Propose $\beta^*$ according to Gamma$\Big(|A| + 1, \int_t^{t+1} \frac{S_\tau I_\tau}{N-1}\, d\tau\Big)$
    Propose $\gamma^*$ according to Gamma$\Big(|B| + 1, \int_t^{t+1} I_\tau\, d\tau\Big)$
    Let $(\beta, \gamma)^{(l)} = (\beta, \gamma)^*$ with probability

        $$\min\left\{1, \frac{f\left(\beta^*, \gamma^* | x_t, y_{1:t}\right)}{f\left(\beta^{(l-1)}, \gamma^{(l-1)} | x_t, y_{1:t}\right)}\right\},$$

    otherwise $(\beta, \gamma)^{(l)} = (\beta, \gamma)^{(l-1)}$.

    Hastings Step for $x_t$:

    Propose $x_t^*$ from $f(x_t | y_{1:t})$ and set $x_{[t,t+1]}^* = x_{[t,t+1]} - x_t + x_t^*$.
    If $x_{[t,t+1]}$ is feasible given $y_{1:t+1}$, let $x_{t+1}^{(l)} = x_{t+1}^*$ and $x_{[t,t+1]} = x_{[t,t+1]}^*$ with
    probability

        $$\min\left\{1, \frac{f\left(x_{(t,t+1]}^* | \beta^{(l)}, \gamma^{(l)}, x_t^*\right) f\left(\beta^{(l)}, \gamma^{(l)} | x_t^*, y_{1:t}\right)}{f\left(x_{(t,t+1]} | \beta^{(l)}, \gamma^{(l)}, x_t\right) f\left(\beta^{(l)}, \gamma^{(l)} | x_t, y_{1:t}\right)}\right\}, \qquad (3.4)$$

    otherwise $x_{t+1}^{(l)} = x_{t+1}^{(l-1)}$.

    Hastings Step for $x_{(t,t+1]}$:

    Move infection times and/or insert, remove or move recovery times as in the
    DA-MCMC algorithm
end
Save $(\beta, \gamma, x_{t+1})^{(1:K)}$ as samples from $f(\beta, \gamma, x_{t+1} | y_{1:t+1})$.
```

Algorithm 6: An example of the SDA-MCMC algorithm for an SIR model.

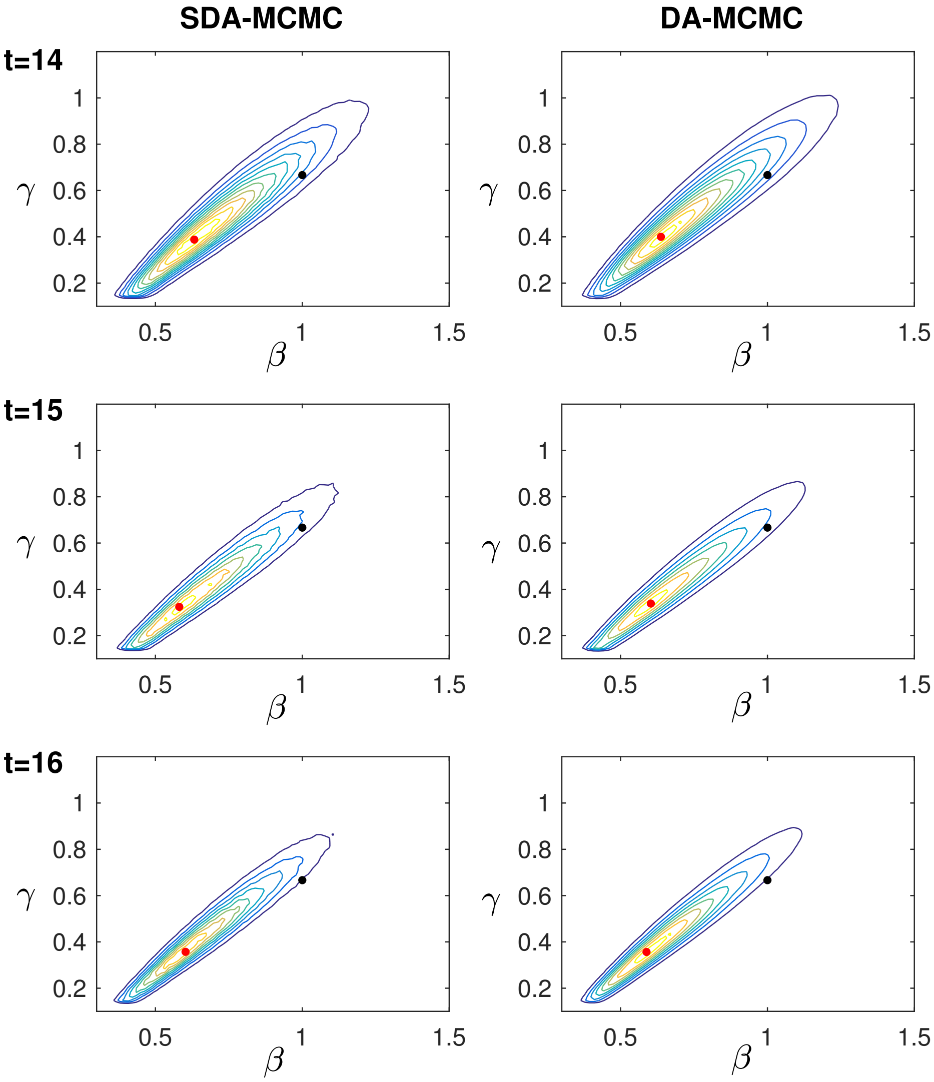

Figure 3.8: Posterior distributions of an SIR model calculated from the SDA-MCMC algorithm (left) and the DA-MCMC algorithm (right). The population size was $N = 100000$ and the true parameter values were $(R_0, \gamma) = (1.5, 1/3)$. The model is initialised at a Uniform$(0, 1)$ time in state $(N - 1, 1, 0)$. The algorithm progresses 1 day at a time from top to bottom, beginning from two weeks worth of data. The black dots corresponds to the true parameter values and red dots correspond to the MAP estimates.

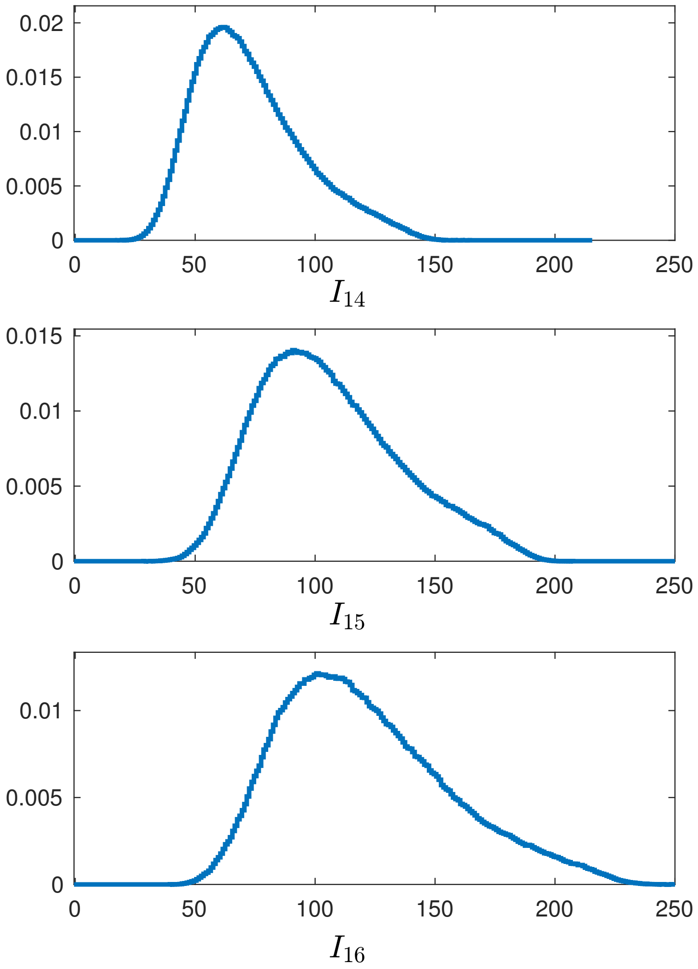

Figure 3.9: The posterior distribution of the number of infectious individuals, $f\left(I_t|y_{1:t}\right)$, for $t = 14$, 15 and 16.

3.3.4 Modifications for independent outbreak data

Suppose we observe p independent outbreaks of the same disease, we can express the augmented posterior distribution (up to proportionality) as

$$f\left(\theta, x^{1:p} | y^{1:p}\right) \propto f(\theta) \prod_{i=1}^{p} f\left(x^i, y^i | \theta\right),$$

where x^i is the entire path corresponding to outbreak i and y_i is the datum obtained over outbreak i. Note that here we use superscripts to denote the outbreak index, whereas previously we used subscripts to index time. Suppose we were to observe completed outbreaks over time, if we had performed inference on m completed outbreaks and then data on $p - m$ more outbreaks became available, we can express the augmented posterior distribution as

$$f\left(\theta, x^{m+1:n} | y_{1:n}\right) \propto f\left(\theta^c | \theta^d, y_{1:m}\right) f(\theta^d | y_{1:m}) \prod_{i=m+1}^{p} f(x^i, y_i | \theta).$$

Estimating the posterior distribution based on all p outbreaks now only requires imputation of hidden variables related to the $p - m$ new outbreaks, that is, we need not impute hidden variables related to the first m outbreaks. Inference is performed similar to in Algorithm 6, where $f\left(\theta^c | \theta^d, y_{1:m}\right)$ is analogous to $f\left(\beta, \gamma | x_t, y_{1:t}\right)$, $f(\theta^d | y_{1:m})$ is analogous to $f\left(x_t | y_{1:t}\right)$ and $\prod_{i=m+1}^{n} f(x^i, y_i | \theta)$ is analogous to the augmented likelihood $f(x_{(t,t+1]}, y_{t+1} | \beta, \gamma, x_t)$.

3.3.5 SI(n)R example

Suppose there is data on multiple completed outbreaks from an SI(n)R epidemic where n is an unknown parameter to be inferred. Once many infections occur, mixing becomes prohibitively slow for the DA-MCMC algorithm and some states corresponding to positive probability become virtually inaccessible (as discussed in Section 2.3.3, Example 3). A DA-MCMC scheme that can be applied to multiple completed outbreaks is outlined in Appendix A; we apply both an SDA-MCMC and a DA-MCMC scheme to a simulated data set from multiple completed SI(n)R outbreaks. Figure 3.10 shows that the DA-MCMC algorithm results in a posterior distribution with no support for $n = 1$, whereas the SDA-MCMC algorithm returns a posterior distribution with support for $n = 1$. Clearly there is non-zero posterior probability for $n = 1$, this highlights that the DA-MCMC algorithm mixes very slowly in the n dimension whereas the SDA-MCMC algorithm has successfully reduced the correlation between successive iterations by integrating out the latent parameters in batches.

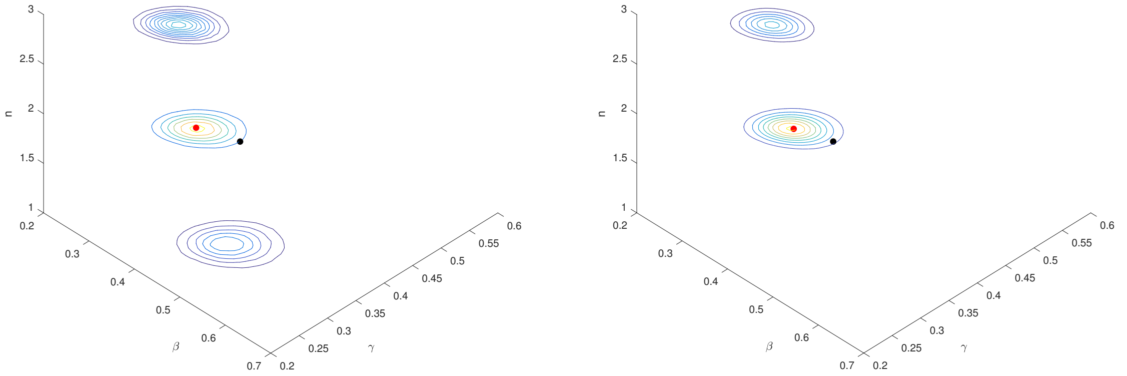

Figure 3.10: Posterior distributions calculated using the SDA-MCMC algorithm (left) and a DA-MCMC algorithm (right) based upon case data from 200 SI(2)R completed outbreaks in subpopulations of size 3. The true parameter values are $(R_0, \gamma) = (1.5, 1/3)$ and infection is seeded in each household at a Uniform$(0, 1)$ distributed time. The red dot corresponds to the MAP estimate and the black dot corresponds to the true parameter value. SDA-MCMC was based upon inference performed on batches of 67, 67 and 68 outbreaks with 1×10^7 iterations and 2×10^6 iterations of burn-in each. The DA-MCMC algorithm was performed based upon 1×10^7 iterations and 2×10^6 iterations of burn-in.

3.3.6 Discussion

SDA-MCMC targets the same sequence of distributions as SMC, however, SDA-MCMC may be preferable if it is difficult to simulate from $x_{(t,T]}$ in a way that agrees with the data set $y_{t+1:T}$. The new algorithm requires us to have a reasonable empirical approximation to $f(x_t, \theta^d | y_{1:t})$ from the initial DA-MCMC run; so if the sample size from the initial DA-MCMC is too low this distribution may be inaccurate. These inaccuracies are most likely to occur in Markov chains with high-dimensional state-spaces or if $f(x_t, \theta^d | y_{1:t})$ has large variance. The algorithm also relies on us having a reasonable approximation of $f(\theta^c|x_t, \theta^d, y_{1:t})$, which can also be inaccurate depending on mixing of the DA-MCMC chain. Further, kernel density estimation of $f(\theta^c|x_t, \theta^d, y_{1:t})$ is inefficient to compute for high-dimensional θ^c. For example, we attempted to implement this for the SEIR model but calculation and use of the kernel density estimate was too computationally expensive.

Some of the key differences from DA-MCMC is that SDA-MCMC is that $f(x_{(t,T]}, y_{t+1:T}|y_{t+1:T}, x_t, \theta)$ does not have an analytical form, as it depends on $f(x_t, \theta|y_{1:t})$, which is estimated via kernel density estimation, whereas, DA-MCMC considers augmented likelihood of the form $f(x_{[0,T]}, y_{1:T}|\theta)$, which typically can be expressed analytically. As a result we generally cannot Gibbs sample from the parameters, so Hastings steps are required.

A difficulty in constructing the Hastings step is that $x_{(t,T]}$ trivially depends on x_t so it could be difficult to sample paths with new starting points; that is, we may need to sample an x_t and based on the x_t alter $x_{(t,T]}$ to ensure feasibility of the new sample path. For compartmental models when x_t is sampled, states $x_{(t,T]}$ can be moved up or down accordingly, such that all events in $(t, T]$ are preserved (though feasibility does need to be checked). For example in the SIR case, we propose a new state I_t^* so the number of infectious individuals over $[t, T]$ becomes $I_{[t,T]}^* = I_{[t,T]} - I_t + I_t^*$, which is only feasible if $I_{[t,T]}^* > 0$.

The reduction of support in the tails of the posterior distribution in the SIR example indicate that some kind of rejuvenation step may be required (similar to SMC algorithms) [22, 24]. That is, if there is too much error in the posterior distribution estimate DA-MCMC may need to be run on the full chain. How one assesses error in the posterior distribution estimate is unclear, which makes it difficult to decide when to rejuvenate.

Chapter 4

Accuracy and Precision of Estimates Under Various Surveillance Schemes

In the early stages of an outbreak, FF100 studies are resourced to collect data on only the first few hundred symptomatic individuals and their contacts [3]. The most important contacts of infectious individuals are typically members within their household, colleagues from their place of work, or peers from their school, as they are easily surveilled and are more likely to become infected than other contacts. It is important to optimise data collection protocols in order to most accurately characterise the disease given limited resources to collect. This chapter aims to explore the accuracy and precision of estimates of measures of transmissibility (the reproduction number and exponential growth rate, as defined in Section 2.2.2) under various schemes for collecting temporal infectious case data. In particular, we aim to determine the best contacts to surveil: those in small subpopulations, such as households; those in large subpopulations, such as workplaces; or, a combination of small and large subpopulations. Supposing that there are resources available to surveil up to, say, 600 individuals; we wish to decide how to split these individuals into subpopulations to obtain precise estimates of the reproduction number and exponential growth rate. We first conduct an analysis on the sizes of subpopulations to surveil to obtain estimates with the lowest variance and bias under the model that generated the data. The variances of estimates under a known model are useful, but in practice we do not know the most appropriate model; so it is also important to choose to surveil subpopulations in a way that is robust with respect to model misspecification. Hence, we conduct an analysis of the robustness of estimates under model misspecification. The bias introduced via model misspecification shows that model selection is a necessary aspect of choosing an optimal surveillance scheme.

This chapter employs DA-MCMC and SDA-MCMC to solve inference problems; DA-MCMC is an exact Bayesian algorithm and SDA-MCMC is an approximate Bayesian algorithm (see Chapter 3). Given a data set, estimates from these algorithms are as precise as possible, however, some data sets are more informative than others. There is control over how the population is surveilled given fixed resources, for example, either households, schools or workplaces could be surveilled; differences in surveillance protocol may lead to differences in how informative a typical data set is. Note that these kinds of subpopulations are surveilled as they contain contacts of infectious individuals which are most easily surveilled, and, surveilling all contacts of every infectious individual is practically infeasible. We base inference on daily infectious case count in each of the surveilled subpopulations until the disease dies out in each subpopulation. Hence, our data has both final size and temporal information. If we surveil many small households, we get the benefit of observing many short sets of time series data and many samples from final epidemic size distributions. Conversely, outbreaks in a few large workplaces gives few samples from the final epidemic size distribution, but may lead to long time series data. It is not yet known in the literature which of these will lead to more informative data sets. Observing outbreaks in a mix of both small and large subpopulations may average out the benefits and drawbacks each of the dataset types.

Results from this chapter can be used to inform how to conduct epidemic surveillance in a way that gives as precise an assessment of the disease as possible under the models considered. Throughout this chapter there are three important assumptions: the first is that outbreaks in subpopulations act independently, the second is that transmission is frequency-dependent, the third is that all infectious cases are observed within a subpopulation. The first assumption is reasonable if members of surveilled subpopulations are non-overlapping and if the overall population is large. The second assumption is that the transmission rate in a subpopulation is of the form $\beta SI/(N-1)$, which depends on the proportion of the subpopulation infected, as opposed to βSI, which depends on the number of individuals infected in the subpopulation. We mention this assumption here to highlight that the conclusions may not carry over to density-dependent models. In the literature household models with frequency-dependent transmission [49, 66] and density-dependent transmission [67–69] have been considered. Frequency-dependence is generally accepted as suitable for sexually-transmitted infectious diseases and vector borne diseases [70]. A study, which considered both density and frequency dependent infection for influenza A in small households, found that frequency-dependent model was a better fit for influenza A [69]. There is some evidence for density-dependant infection of pneumococcal in households [67], but it is not known how the transmission term changes for lager households, and it is accepted that actual transmission is likely to be somewhere between density and frequency dependent [67, 68]. As there is limited knowledge of the appropriate transmission rate function for large households, and frequency-dependent transmission suits small households for influenza A, it may be an appropriate model for transmission for pandemic preparedness. The third assumption is reasonable if the disease results in few asymptomatic cases and if we either are likely to observe the first case in a subpopulation, or we are able to retrospectively determine the days on which individuals became symptomatic.

In general, inference with model misspecification, that is, inference on a model that did not give rise to the data, can lead to biased results. Through a simulation study, we assess how different kinds of model misspecification on outbreaks in subpopulations bias estimates. In particular we look at misspecification of exposed and infectious period distributions. Misspecification has been considered in the context of incorrectly assuming data came from a SIR model, where it is actually simulated from an SI model [32], whereas here we consider SEIR and SIR models with different infectious and exposed period distributions. Biases with respect to the basic reproduction number, R_0, under assumptions of misspecified mixing structure has been previously investigated [33]. In comparison, this chapter investigates bias with misspecification of the exposed and infectious period distributions on both the reproduction number and the exponential growth rate, r. This growth rate is a function of parameters related to within-subpopulation infection, so in the context of small households the term does not have a direct applicability. However, for a frequency-dependent transmission model, the estimated growth rate translates to the exponential growth rate in outbreaks in large subpopulations, which is of practical importance.

In cases where bias is large if the model is misspecified, it is advantageous to perform inference on both the model and parameters. This chapter aims to understand how different kinds of model misspecification biases estimates of transmissibility if model selection is not performed. To this end, we compare results from inference where data are generated and inferences are made on the model that generated the data, and inferences are made on a different model. We use DA-MCMC to infer the model parameters of SE(n_1)I(n_2)R models, and SDA-MCMC to infer both shape and model parameters for a SI(n)R model. It should be noted that we do not make inferences on the shape parameters of the SE(n_1)I(n_2)R model as the mixing for the DA-MCMC method is prohibitively slow; alternative methods such as SMC-squared or PM-MCMC could possibly be employed, but could be inefficient due to the relatively large data sets. We investigate inference on multiple SE(n_1)I(n_2)R outbreaks using these kinds of methods in Chapter 6.

4.1 Methods

Our data sets are simulated from SEIR and SE(2)I(2)R models and inference is conducted assuming the underlying model is a SEIR, SE(2)I(2)R or SI(n)R model (with unknown shape parameter), for specific details on the inference procedure see Appendix A. The data sets used give the total number of infectious cases at a daily resolution, though, we also investigate whether final epidemic size data are more robust under model misspecification. We assume that in each subpopulation a single infection is seeded at a uniformly distributed time on the first day and from there on the epidemic proceeds as per the continuous-time Markov chain model dynamics, until there are only susceptible or recovered individuals in the subpopulations. If there are no secondary infections in any of the subpopulations then the likelihood function is given by

$$\left(\frac{n_2}{R_0 + n_2}\right)^{n_2 M},$$

where n_2 is the shape parameter of the infectious period and M is the number of subpopulations. Hence, if there are no secondary infections, the data set only informs the reproduction number. Data sets without any secondary infections would infer R_0 as a low value; in this case as R_0 has a uniform prior and a monotonically decreasing likelihood function so the MAP of R_0 is at its lower boundary value, 0.25. Further, the exposed period will be entirely informed by the prior distribution and the infectious period will be negatively biased. As such, we only include simulated data sets in which at least one secondary infection occurs in one of the subpopulations.

We infer parameters given 600 individuals grouped into different subpopulations and assess the impact of the grouping of the 600 individuals on inference. The goal is to group the 600 individuals in a way such that estimates have as low variance and bias as possible. Note, the number 600 was chosen as it is roughly the number of people we may expect to be able to surveil, in a FF100 study and it can be divided evenly in many ways. We simulated 250 epidemics from an SE(2)I(2)R model and a SEIR model for both a moderately and highly transmissible disease, $(\beta, \sigma, \gamma) = (0.933, 0.5, 2/3)$ and $(\beta, \sigma, \gamma) = (0.933, 0.5, 0.45)$, respectively. The scenarios considered had the 600 individuals split into 200 subpopulations of 3 individuals, 150 subpopulations of 4 individuals, a combination of 100 subpopulations of 3 individuals and 3 subpopulations of 100 individuals, 6 subpopulations of 100 individuals and 2 subpopulations of 300 individuals, these scenarios are labelled 200:3, 150:4, (100:3,3:100), 6:100 and 2:300 respectively. Model parameters were estimated for each simulated data set using the DA-MCMC algorithm described in Appendix A, with a burn-in of 10^6 iterations with 2×10^6 samples (thinned to 2×10^5 for data storage reasons) and 10 Hastings steps proposed for the hidden process in each iteration. Once model parameters are inferred we calculate key epidemiological parameters R_0 and r as β/γ and the solution to Equation (2.5) (Section 2.2.2) respectively [15]. All results from the DA-MCMC samples are given as box plots of the median values and contour plots of kernel density estimates of the posterior distribution of R_0 and r.

In Section 4.2.1 we perform inference on the SEIR and SE(2)I(2)R data sets assuming the correct model is known. In Section 4.2.2 we use the same simulated data sets from the SE(2)I(2)R model used in Section 4.2.1 and perform inference as though the data were generated from a SEIR model and a SI(n)R model with an unknown shape parameter to be inferred. The results based on the SEIR model are used to show how simplifying the exposed and infectious period to be exponentially distributed biases our assessment of transmissibility. The results based on the SI(n)R model show how the lack of a exposed period affects results. This is important as the exposed period has an impact on temporal data but not final size distribution. The shape parameter is inferred here, as this is a tractable inference problem using SDA-MCMC (implemented as in Section 3.3.5). However, the high dimensionality of inference with the SE(n_1)I(n_2)R model makes joint inference of shape parameters and rate parameters a difficult problem. Lastly, we perform inference based on a SIR, or equivalently a SEIR model, using only the final epidemic size data. This allows us to assess the biases due to temporal information under model misspecification.

4.2 Results

4.2.1 Subpopulation size and outbreak repetitions

Figures 4.1 and 4.2 show boxplots of the median estimates of the reproduction number, R_0, and growth rate, r, for a SEIR and SE(2)I(2)R model, respectively (with parameters inferred under the models that generated the data). We observe that estimates of key epidemiological parameters tend to be unbiased, possibly apart from estimates related to scenario 2:300 with moderate transmissibility parameters. In this scenario for the SEIR model, on average, estimates of r were 35% lower than the true value and estimates of R_0 were 9% lower than the true value, and for the SE(2)I(2)R model, on average, estimates of r were 46% lower than the true value and estimates of R_0 were 17% lower than the true value. For the moderate transmissible parameters, bias from the 2:300 scenario correspond to simulations where each of the subpopulations experience early fade out. Figure 4.3 shows the total number of secondary infections under each model and scenario; we find that the moderately transmissible SEIR and SE(2)I(2)R models experience early fade out in every subpopulation roughly 40% and 30% of the time for scenario 2:300, respectively. Similarly, outliers from the moderately transmissible 6:100 scenario and the high transmissible 2:300 scenario correspond to simulations which experienced early fade out (which occurs in roughly 5%-15% of these simulations). The multiple modes in Figure 4.3 for scenarios 6:100 and 2:300 correspond to various numbers of subpopulations experiencing epidemic fade out. High transmissibility parameters led to unbiased estimates in all scenarios, as there is a lower chance of epidemic fade out in subpopulations. The variance in estimates in all cases was lowest for the 200:3 scenario. For the SEIR model with moderate and high transmissibility the variance of estimates of (R_0, r) were $(0.0221, 0.0011)$ and $(0.0500, 0.0012)$, respectively. For the SE(2)I(2)R model with moderate and high transmissibility the variance of estimates of (R_0, r) were $(0.0143, 0.0008)$ and $(0.0353, 0.0008)$, respectively. The variance in estimates in all cases was highest for the 2:300 scenario. For the SEIR model with moderate and high transmissibility the variance of estimates of (R_0, r) were $(0.0550, 0.0033)$ and $(0.1845, 0.0068)$, respectively. For the SE(2)I(2)R model with moderate and high transmissibility the variance of estimates of (R_0, r) were $(0.0922, 0.0035)$ and $(0.1175, 0.0043)$, respectively. Many outbreaks in small populations tend to led to lower variance estimates than few outbreaks on large subpopulations. This informs us that transmissibility can be more accurately inferred if data are collected on many small subpopulations rather than few large subpopulations. We see that in almost all cases the variance in the estimates of the SE(2)I(2)R model is slightly lower than that of the SEIR model, that is, we are more easily able to determine key model parameters for models with Erlang-2 distributed exposed and infectious periods; this is likely due to the lower variance in the infectious and exposed period distributions.

We conclude that in all cases data on many outbreaks in subpopulations leads to more accurately inferred epidemiological parameters, assuming we know the model that gives rise to the data. Out of the scenarios considered it was best to surveil 200 subpopulations of size 3 or 150 of size 4. These resulted in accurate estimates with low variance, this partially because the distribution of the number of secondary infections has low variance and no support near 0, and because the temporal data are informative enough to obtain low variance estimates of r. Hence, if we know which model is most appropriate, we would recommend surveilling many households as opposed to few schools or workplaces. Further, if the disease is highly transmissible and has Erlang distributed infectious and exposed periods we are able to more accurately infer parameters.

4.2.2 Latent and infectious period distribution

Figure 4.4 shows box plots of median parameter estimates of R_0 and r from data generated from an SE(2)I(2)R model and inference based on a SEIR model. For moderate transmissibility scenarios 200:3, 150:4, (100:3,3:100), 6:100 and 2:300 the average percentage difference between estimates of (R_0, r) and their true value is (18,20), (12,11), (12,9), (-2,-17) and (-9,-40), respectively. For high transmissibility scenarios 200:3, 150:4, (100:3,3:100), 6:100 and 2:300 the average percentage difference between estimates of (R_0, r) and their true value is (30,8), (21,6), (13,0.3), (1,-2) and (-2,-5), respectively. We conclude that, for both highly and moderately transmissible diseases, misspecifying an SE(2)I(2)R model as a SEIR model gives positively biased estimates of R_0 from scenarios with many outbreaks in small households. Estimates of R_0 tend

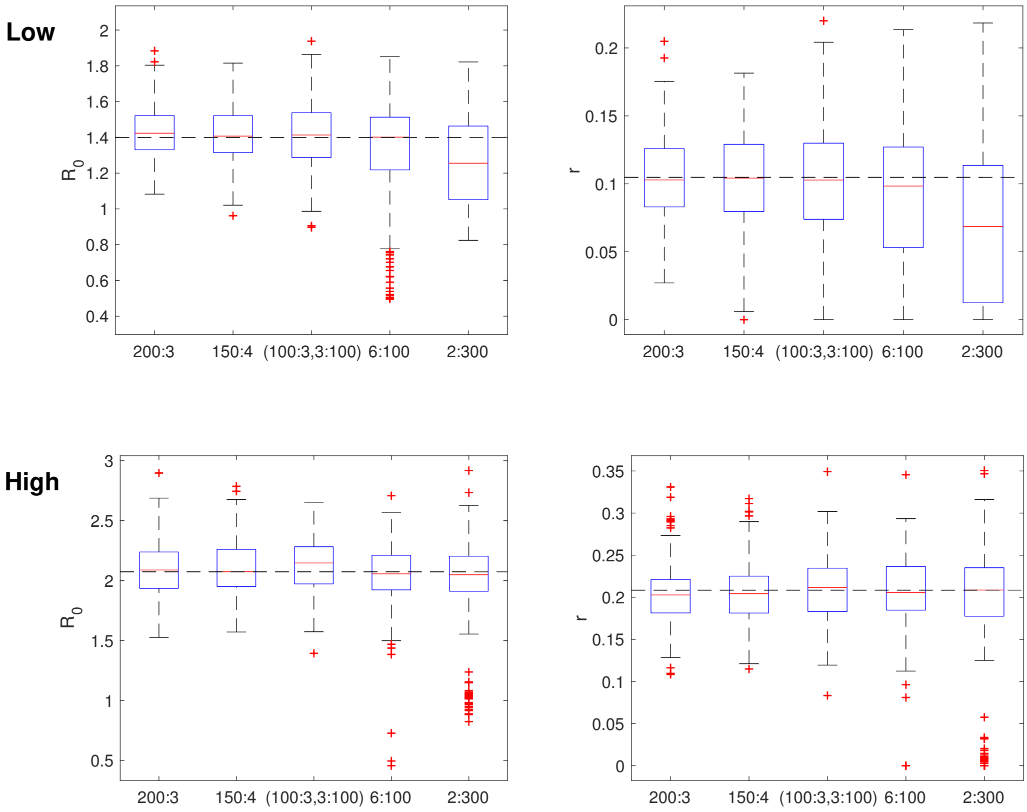

Figure 4.1: Boxplots of median estimates of R_0 and r from 250 completed simulations of SEIR epidemics. Infection in each subpopulation is seeded at a Uniform$(0, 1)$ distributed time. Subpopulation distribution 1-5 correspond to 200 households of size 3, 150 of size 4, 100 of size 3 and 3 of size 100, 6 of size 100 and 2 of size 300 respectively. Parameter sets were $(\beta, \sigma, \gamma) = (0.933, 0.5, 2/3)$ (top) and $(\beta, \sigma, \gamma) = (0.933, 0.5, 0.45)$ (bottom).

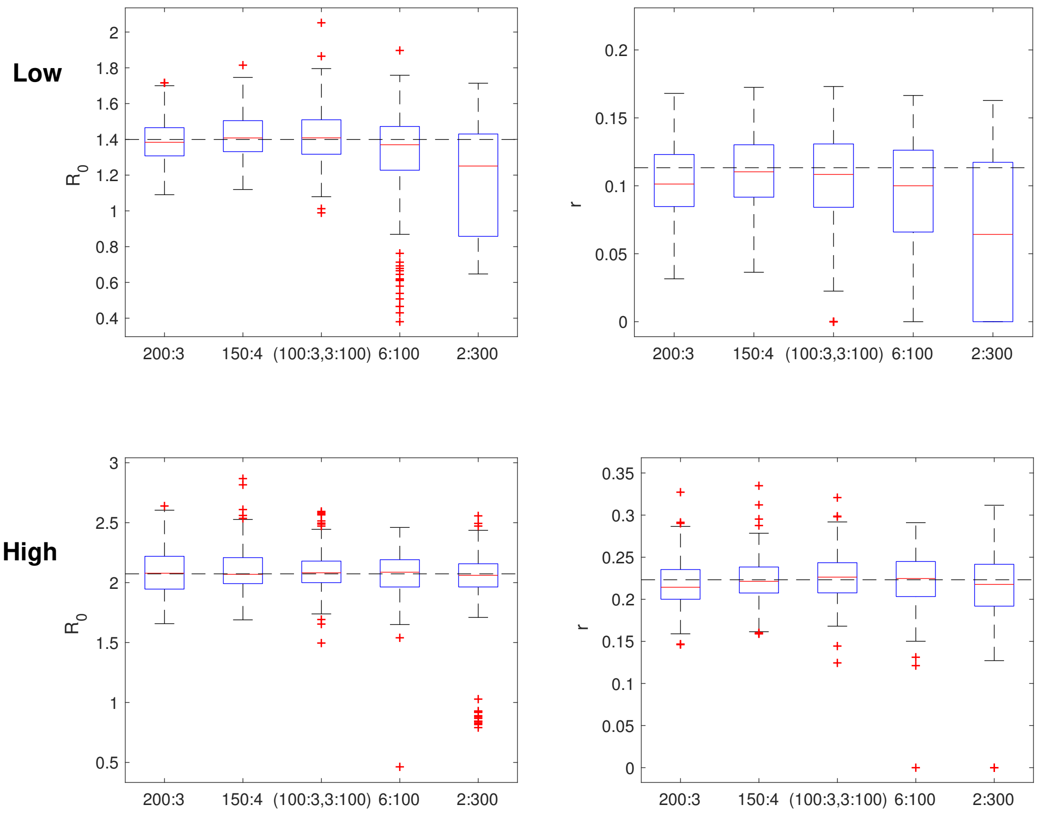

Figure 4.2: Boxplots of median estimates from 250 completed simulations of SE(2)I(2)R epidemics. Infection in each subpopulation is seeded at a Uniform$(0, 1)$ distributed time. Subpopulation distribution 1-5 correspond to 200 households of size 3, 150 of size 4, 100 of size 3 and 3 of size 100, 6 of size 100 and 2 of size 300 respectively. Parameter sets were $(\beta, \sigma, \gamma) = (0.933, 0.5, 2/3)$ (top) and $(\beta, \sigma, \gamma) = (0.933, 0.5, 0.45)$ (bottom).

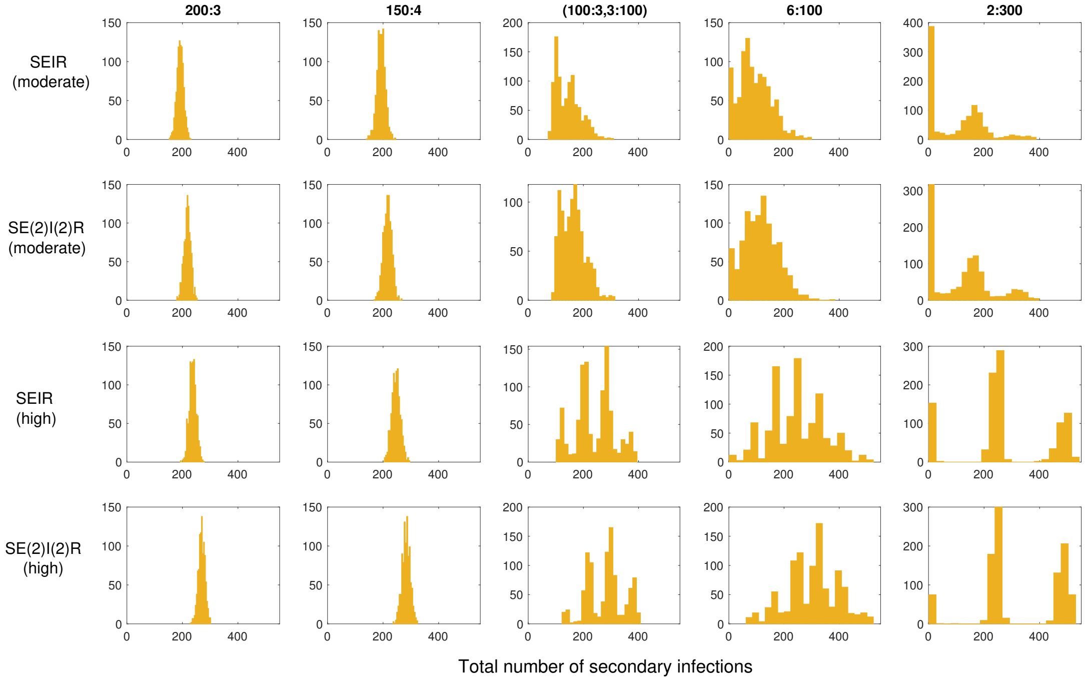

Figure 4.3: Histograms of the total number of secondary infections from moderate and high transmissibility SEIR and SE(2)I(2)R epidemics. Subpopulation distributions correspond to 200 households of size 3, 150 of size 4, 100 of size 3 and 3 of size 100, 6 of size 100 and 2 of size 300 respectively. Each histogram is generated from 1000 simulated data sets. Parameter sets for moderate and high transmissibility were $(\beta, \sigma, \gamma) = (0.933, 0.5, 2/3)$ and $(\beta, \sigma, \gamma) = (0.933, 0.5, 0.45)$, respectively.

to have low bias for scenarios 6:100 and 2:300 and for the high transmissibility scenario, estimates of r have low bias for scenarios 6:100 and 2:300. Understanding these biases is confounded by opposing sources of positive and negative bias. Some positive bias in R_0 occurs as the SE(2)I(2)R model generally leads to more secondary infections in total than the SEIR model (see Figure 4.3). However, the scenarios with larger subpopulations still have a reasonable probability of early epidemic fade out in all of the subpopulations, which leads to negative bias in R_0. Further, it is unclear as to how the temporal information affects results. In the moderate transmissibility scenario all of our estimates of r are biased, but the high transmissibility scenario has low bias, particularly in scenarios with large subpopulations. Note that as r is the positive solution to Equation (2.5) from Section 2.2.2, for the SEIR model r is given by

$$r = \frac{1}{2}\left(-\sigma - \gamma + \sqrt{(\sigma+\gamma)^2 - 4(1-R_0)\gamma\sigma}\right). \tag{4.1}$$

Hence, positive bias in R_0 may lead to positive bias in r. In comparison to Figure 4.2, the parameter estimates have comparable variance, but are highly biased, particularly in the scenarios with small subpopulations. The lower bias for larger subpopulations may be due to both a lower bias in R_0 and longer sets of temporal information, which are useful for estimating r. It is important to note that estimates of key epidemiological parameters were generally positively biased, so this kind of model misspecification generally leads to conclusions that more people will get infected and that the disease will spread faster than is actually the case.

The results of SE(2)I(2)R data analysed under a SI(n)R model (where n is an inferred parameter), given in Figure 4.5. These show that all epidemiological parameters tend to be underestimated for both high and moderate transmission epidemics across all subpopulation sizes. However, the bias in general tended to be lower in scenarios with outbreaks in larger subpopulations. In general we would expect to see this negative bias in r when we perform inference from a SI(n)R model, as infections are not delayed due to an exposed period.

Figure 4.6 shows individual posterior distributions from a single SE(2)I(2)R simulation under each scenario with inference based on the true model, a SEIR model and a SI(n)R model. We observe that under the true model the small subpopulation scenario

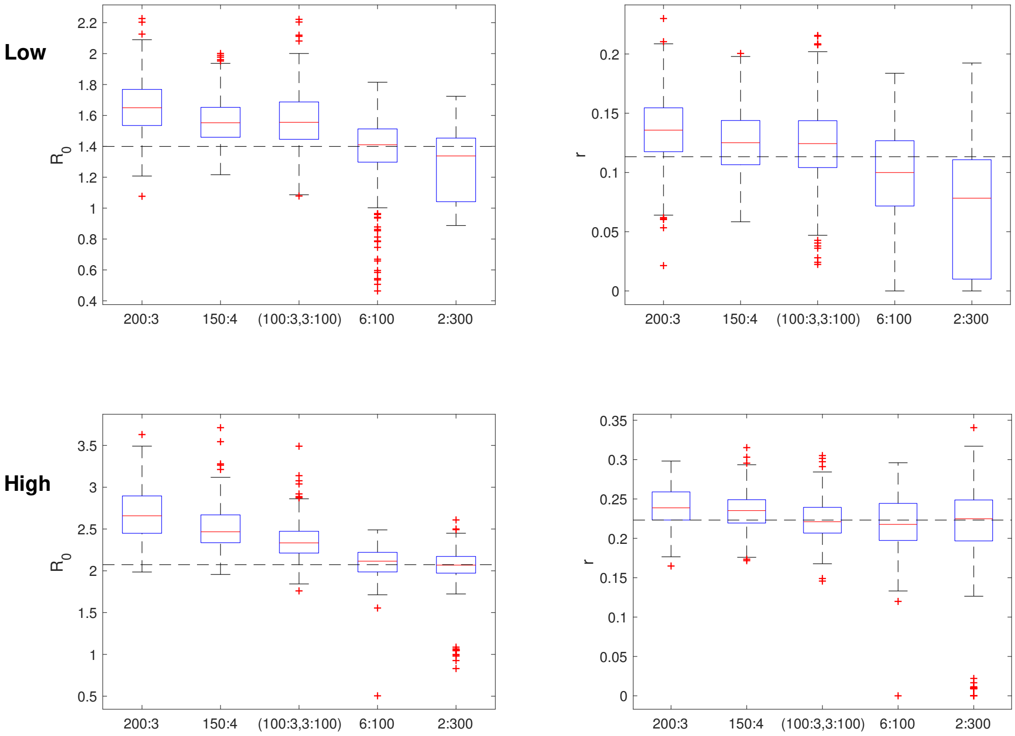

Figure 4.4: Boxplots of median estimates based on SEIR inference from 250 completed simulations of SE(2)I(2)R epidemics. Infection in each subpopulation is seeded at a Uniform$(0, 1)$ distributed time. Subpopulation distribution 1-5 correspond to 200 households of size 3, 150 of size 4, 100 of size 3 and 3 of size 100, 6 of size 100 and 2 of size 300 respectively. Parameters for the simulations were $(\beta, \sigma, \gamma) = (0.933, 0.5, 2/3)$ (top) and $(\beta, \sigma, \gamma) = (0.933, 0.5, 0.45)$ (bottom).

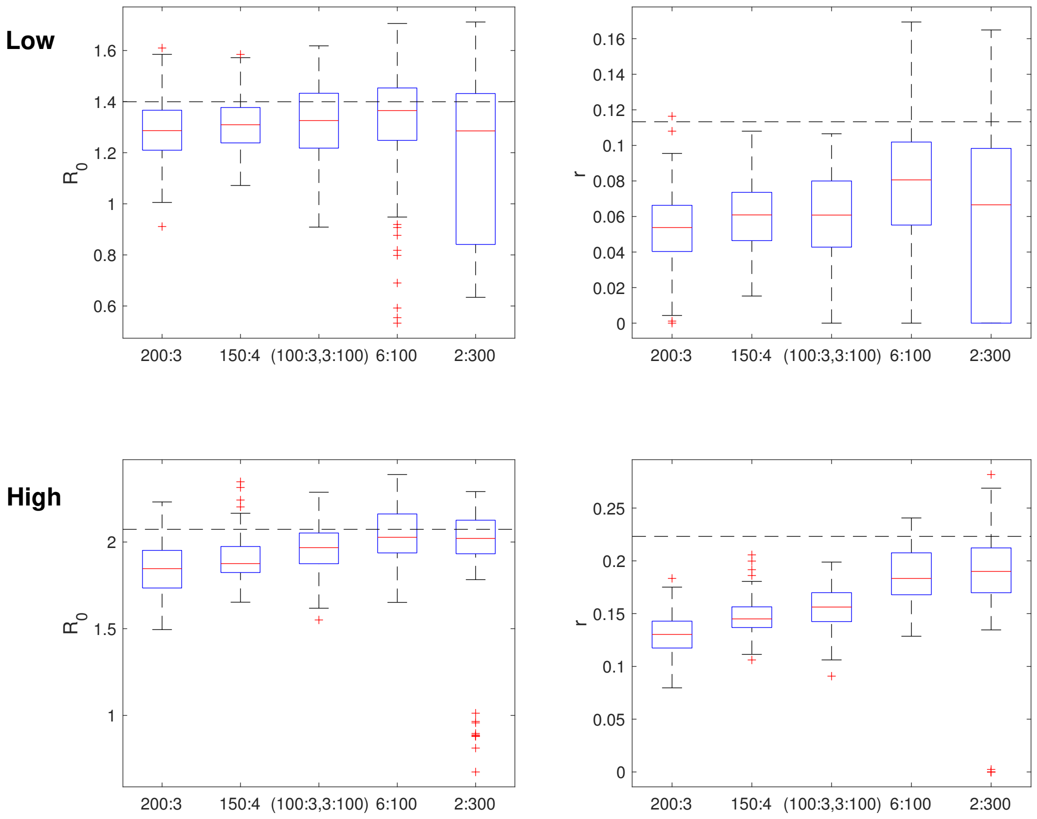

Figure 4.5: Boxplots of median estimates based on SI(n)R inference from 250 completed simulations of SE(2)I(2)R epidemics. Infection in each subpopulation is seeded at a Uniform$(0, 1)$ distributed time. Parameters for the simulations were $(\beta, \sigma, \gamma) = (0.933, 0.5, 2/3)$ (top) and $(\beta, \sigma, \gamma) = (0.933, 0.5, 0.45)$ (bottom).

leads to posterior distributions with low variance in R_0, whereas large subpopulations scenarios led to posterior distributions with larger variance in R_0 but comparable variance in r, and the mixed scenario had the lowest variance in r. Under the SEIR model the small subpopulation scenarios have a large positive bias in r and a slight positive bias in R_0 with an greater variance in both variables compared to results based on the true model. The mixed scenario saw a change in the location of the posterior distribution, with an increased variance in r. The large subpopulation scenarios had very little change in the location of the posteriors, but the variance in R_0 increased. Under the SI(n)R model all posteriors had a reduction in variance and a slight change in location. For the small and mixed scenarios the location moved lower with respect to R_0 and r; for the large scenarios the locations of the posteriors shifted up with respect to R_0.

Lastly, we assess whether final epidemic size data are preferable to temporal data under model misspecification when inferring R_0. Note, these results are only expressed in terms of R_0 as r cannot be inferred from final epidemic size data only. Results based on final epidemic size data are given in Figure 4.7. We see that estimates have very low variance and high bias for scenarios 200:3 and 150:4; the low variance is due to the large number of repetitions of final epidemic size data, and the high bias is due to the misspecification of the infectious period distribution. For scenarios 6:100 and 2:300 we note results look very similar to results from Figures 4.4 and 4.5; this indicates that R_0 is largely estimated by the final epidemic size data for outbreaks in large subpopulations under model misspecification.

It appears that scenarios 6:100 and 2:300 had the lowest bias under all model misspecifications considered. Scenario 6:100 has the least bias introduced from epidemic fadeout and has a lower variance than scenario 2:300, so this is preferred. However scenario (100:3,3:100) tended to have lower variance estimates and was not as biased as scenarios 200:3 and 150:4 in all cases. For the models considered, scenario (100:3,3:100) leads to unbiased estimates with lower variance under the correct model, and is less susceptible bias from initial fadeout than both scenario 6:100 and 2:300. The mix of many small subpopulations and a few large subpopulations allow for many sets of final epidemic size data and some longer sets of temporal data in large populations. So, although the mix is not the best choice under a correctly specified model, it appears

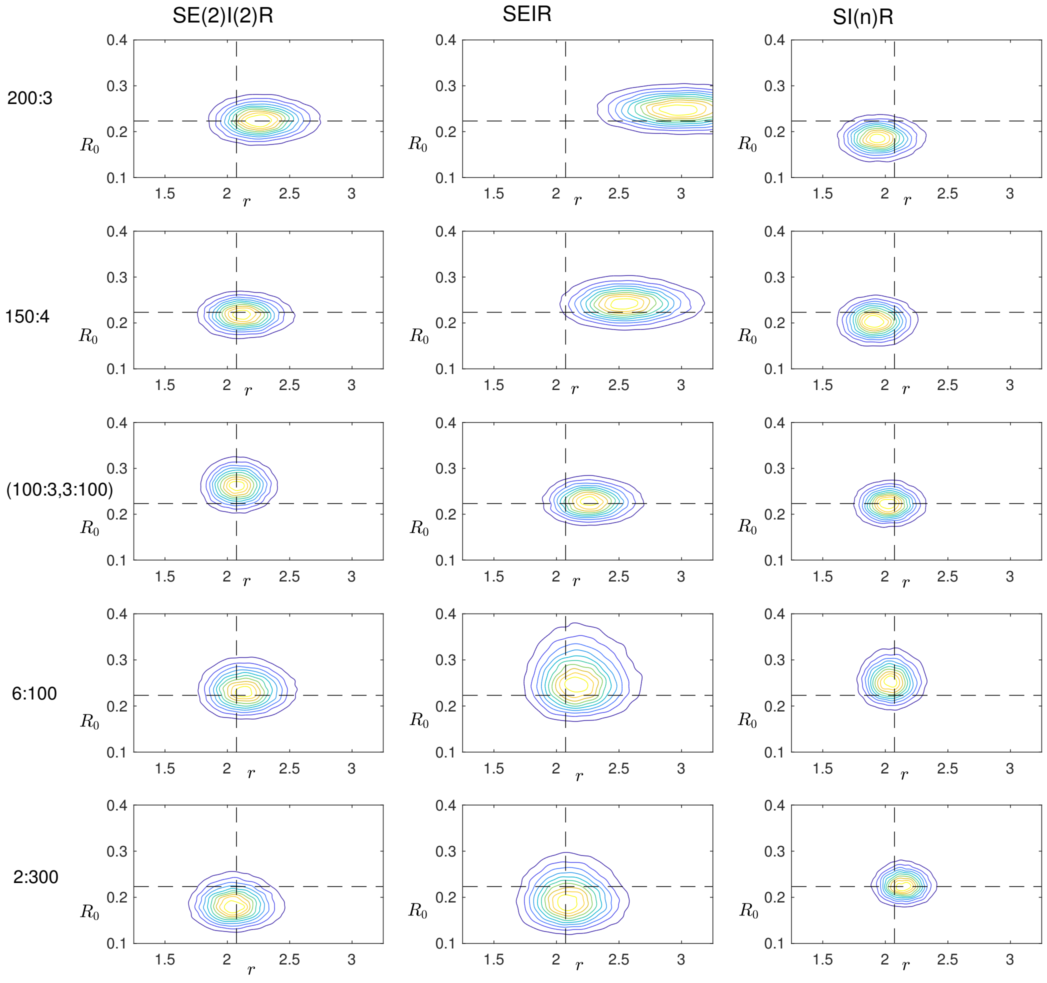

Figure 4.6: Posterior distributions of R_0 and r based on completed SE(2)I(2)R simulated data from each scenario inferred under each model. Infection in each subpopulation is seeded at a Uniform(0, 1) distributed time. Parameters for the simulations $(\beta, \sigma, \gamma) = (0.933, 0.5, 0.45)$.

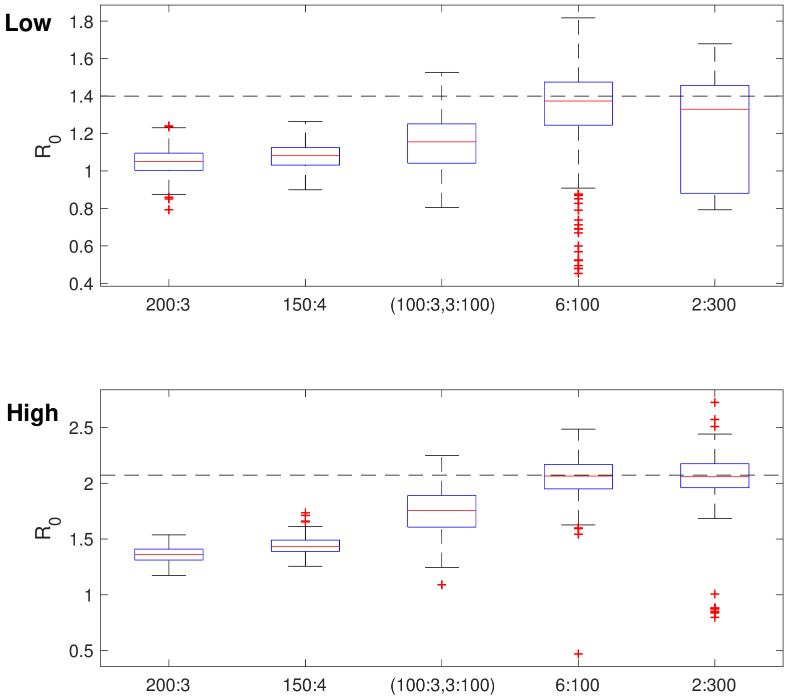

Figure 4.7: Boxplots of median estimates based on a SIR inference on final epidemic size data from 250 simulations of SE(2)I(2)R epidemics. Parameter sets were $(\beta, \sigma, \gamma) = (0.933, 0.5, 2/3)$ (top) and $(\beta, \sigma, \gamma) = (0.933, 0.5, 0.45)$ (bottom).

to be the most robust surveillance scheme under the kinds of model misspecification considered here.

4.3 Discussion

We investigated how to surveil a mix of subpopulation sizes, for example, corresponding to households, schools and workplaces, in a way that gives the most precise estimates of epidemiological parameters. These units are a natural setting for surveillance as they are less resource intensive to surveil than many randomly selected individuals in the population.

We used data-augmented MCMC methods to infer parameters of models from simulated data sets. We found that, if we knew the correct model, having many repetitions of small outbreaks led to more precise estimates than few outbreaks in large populations. However, under model misspecification the least biased results came from the scenario with six outbreaks in subpopulations of size 100, that is, the least biased estimates came from a scenario in which the subpopulations were large but the probability of initial epidemic fadeout in all subpopulations was relatively low. The scenario with a mix of many small subpopulations and few large subpopulations led to estimates with lower variance compared to those based on large subpopulations, and a lower bias under model misspecification compared to those with only small subpopulations. The 'mix' scenario is a trade-off between the many small, and few large subpopulation outbreaks. These results indicate that if a model is not selected in an informed way, we should not surveil only small subpopulations; a mix of small and large, such as households and schools would be more appropriate. Definitively claiming this as the most robust scheme in general is difficult, as there are confounding effects from biases arising from different probabilities of initial fade out, different lengths of data sets, different distributions for the number of secondary infections and different numbers of samples from final size distributions in subpopulations. If there is good evidence for a particular model, we should surveil many small subpopulations, such as households. In general, R_0 tended to be more robust under model misspecification than r, though r was estimated well if the disease is highly transmissible and the infectious and exposed periods are modelled but have misspecified distributions.

The limitations of our analysis include that we only considered models with frequency-dependent transmission, so results may not hold if density-dependent transmission is more appropriate. Although, there is some evidence that models with frequency-dependent transmission fit influenza A better than models with density-dependent transmission in small households [69], it is unknown as to how this behaves for larger households. In other cases, density-dependent models were found to be more suitable [67, 70]. In reality, the transmission for infectious diseases may be somewhere between the two [67,68]. The kind of transmission has an impact on the effect of vaccination measures [71], so the characterisation of the transmission may be of practical concern. This indicates that further study into the transmission type may be necessary, this could be done by via model selection, or by inferring an extra model parameter, ω, for transmission rate $\beta\frac{SI}{(N-1)^{\omega}}$ for $\omega \in [0,1]$, which represents a trade-off between density and frequency-dependence.

We also assumed that all cases in subpopulations are observed, which is only sensible if asymptomatic cases are rare and symptoms are easily detected within a surveilled subpopulation. We also only included data sets that contained at least one secondary infection in one of the subpopulations, this meant that some simulated data sets for scenarios with few outbreaks and moderate transmissibility were rejected. These data sets were entirely uninformative for the exposed period and led to negatively biased the estimates of R_0 and the infectious period.

All inference was based on transmission within subpopulations, however, we wish to use this to characterise transmission in the overall population. Hence the estimates are implicitly related to the spread of disease in a homogeneously mixing population and does not necessarily capture the spread of disease between subpopulations in a population with inhomogeneous transmission. We also assumed that data sets are from completed outbreaks in each subpopulation, however, in general it is unknown as to when the outbreak is completed. Our analysis could be extended to allow some outbreaks to be continuing, however this reduces the efficiency of data-augmented MCMC algorithms as the dimension of the hidden process is large; whereas data from full outbreaks give the exact number of exposed, infectious and recovered individuals. Further, we note that subpopulations are chosen as they are an efficient unit to surveil, however, our analysis does not account for any difference in efficiency of data collection. For example, it may be as easy to surveil 600 individuals in subpopulations of size 3 as it is to surveil 800 individuals in subpopulations of size 100. It is possible to redo the analysis by modelling the efficiency of surveillance, but it is unclear as to how the efficiency varies with the subpopulation size, so a fixed number of individuals was chosen.

The analysis of robustness of the inferred epidemiological parameters under model misspecification revealed that biases were unpredictable due to differences in the distributions of number of secondary infections, the number of samples from final size distributions, and, the lengths of data sets. This indicates that model selection is a necessary aspect of characterising emerging infectious diseases from FF100 study data. Our work in Chapters 5 and 6 aims to improve efficiency of current methods for Bayesian model selection, such as SMC-squared [24, 25], so that they can be applied to these kinds of data sets and models. To assess which surveillance scheme is optimal, we propose choosing a sensible subset of models, with frequency and density dependent transmission and various infectious and exposed period distributions, then simulating from different models and performing Bayesian Model Averaging under each surveillance scenario. The resulting estimates should be robust and lead to a more definitive optimal surveillance scheme.

Chapter 5

Model Selection via Importance Sampling

Bayesian model selection is a way of choosing between competing models in a way that incorporates prior knowledge of models and their parameters; this is naturally a desirable approach if there are few datum but reasonable prior knowledge of a process. It compares models using the probability that each model generated the data given some observations and prior distribution over models and parameters [72]. We consider the model to be inferred as a parameter (typically with a uniform prior) and hence multiplying the model evidences, that is, the likelihood of each model, by the prior distribution and normalising gives the probability of each model given the data [38]. While the interpretability and the consistency with the Bayesian paradigm is desirable, Bayesian model selection has some issues: many datum may be required before models can be distinguished; and the evidence is typically difficult to calculate as the likelihood is intractable. This is true for epidemic models, which for small populations are most naturally represented as partially-observed continuous-time Markov chains (CTMCs) [6, 73, 74].

This chapter describes an efficient way of performing Bayesian model selection for partially-observed CTMCs and applies our method to the two important problems in mathematical epidemiology: inferring the shape of the infectious period distribution and identifying the onset of symptoms relative to infectiousness. Our method works by calculating unbiased estimates of the likelihood via sequential importance resampling, which in turn is used in another importance sampling algorithm to estimate posterior model probabilities. A novel feature of this method is that the likelihood is estimated via a scheme which is ideally suited for partially-observed CTMCs, where one component of the state is observed exactly [74]. This works by sampling realisations of the partially-observed process in a way that realisations always match with observations. Our combined method is both computationally efficient and embarrassingly parallelisable and hence well suited to implementation on modern computing hardware. Further, the ability to estimate error bounds and use stopping criterion ensures the accuracy of evidence estimates. Similar CTMC models see wide use in areas of biology such as phylogenetics [75], ecology [76] and cell biology [77], and hence our method may be applied to a broad range of biological model selection problems.

We first discuss the methods for importance sampling over the parameter space, and the sequential importance resampling algorithm used to estimate the likelihood in Section 5.1. We then describe the two case studies, their implementation and results in Sections 5.2 and 5.3. Lastly we conclude in Section 5.4 and discuss how our method relates to those already in the literature.

5.1 Methods

Importance sampling is a method for estimating properties of one distribution by sampling from another distribution; the bias is corrected to obtain unbiased estimates related to the distribution of interest [78]. This technique is useful when the target distribution is difficult to sample from, such as the distribution of latent variables from a partially-observed Markov chain. This method has been effectively applied for parameter inference, for example in particle marginal MCMC schemes [60, 79] and has also been applied to model selection [23, 25, 72, 80]. In this chapter we use importance sampling in the space of latent variables to estimate likelihoods via sequential importance resampling and within the parameter space to calculate the model evidence. The novelty of this approach is that the likelihood is estimated via the most suitable kind of importance sampling scheme for these kinds of models, as described in [74].

Here we introduce importance sampling for model selection, detail how likelihood estimates are used in model selection and describe how sequential importance resampling is used for likelihood estimation.

5.1.1 Importance sampling for evidence estimation

Importance sampling for model selection has been discussed in [23, 25, 72], and in this chapter we adopt a similar approach. However, our implementation uses an efficient particle filter to obtain unbiased estimates of the likelihood, which allows us to obtain unbiased estimates of the evidence.

Suppose we have data set y of observations from a model with parameter set θ in parameter space Θ. Let $p(\theta)$ denote the prior distribution and $p(y|\theta)$ denote the likelihood function, then the model evidence is given by

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta.$$

To estimate $p(y)$ we may sample m parameter sets $\theta_1, \ldots, \theta_m$ from some arbitrary, importance sampling, density, $q(\cdot|y)$, with support Ψ where $\Theta \subseteq \Psi$. We then compute random variables

$$Z_i = \frac{p(y|\theta_i)p(\theta_i)}{q(\theta_i|y)}, \tag{5.1}$$

for $i = 1, \ldots m$, where $p(\theta)$ is the prior distribution and $p(y|\theta)$ is the likelihood function. These random variables are chosen so that their expectation is given by

$$\begin{aligned} E[Z_i] &= \int_{\Psi} \frac{p(y|\theta)p(\theta)}{q(\theta|y)} q(\theta|y) \;\; d\theta \\ &= \int_{\Theta} p(y|\theta)p(\theta) \;\; d\theta \\ &= p(y). \end{aligned}$$

Hence, the mean of the Z_i's provides an unbiased estimate of the evidence, $\hat{p}(y)$. If the likelihood is intractable we can use unbiased estimates, $\hat{p}(y|\theta)$, instead of $p(y|\theta)$ in Equation (5.1), by the law of total expectation. Note that using importance sampling to estimate $p(y)$ involves obtaining independent identically distributed samples of Z_i and taking their mean. Therefore the central limit theorem can be applied once a large number of samples are obtained, so we can estimate error bounds on $\hat{p}(y)$ which shrink like $1/\sqrt{m}$. Hence, one can sample until estimates of the evidence are within a specified tolerance. Further, the independence of the sampling procedure allows these computations to be run in parallel and so modern computation hardware can be easily utilised. Of course, having to estimate the likelihood inflates the variance of $\hat{p}(y)$, so a larger number of samples are required for $\hat{p}(y)$ to converge to a given accuracy.

The variance of this estimate is lowest if the sampling density, $q(\cdot|y)$, is similar to the posterior distribution. Our method requires no parameter inference, but may be made more efficient, especially in higher dimensions, by choosing a sampling distribution similar to the posterior distribution if parameter inference has been performed; for example, the sampling distribution could be a Gaussian with the mean and covariance matrix estimated from MCMC samples [23]. In this chapter we do not use MCMC samples to inform the choice of $q(\cdot|y)$, but this could be done in practice to improve the convergence of estimates.

5.1.2 Sequential importance resampling algorithm for likelihood estimation

Consider a time series, $y = (y_1, \ldots, y_T)$, from a partially-observed CTMC. We can express the likelihood at θ as

$$p(y|\theta) = \prod_{t=1}^{T} p(y_t|y_{1:t-1}, \theta),$$

where we adopt the convention that $p(y_1|y_{1:0}, \theta)$ is taken to mean $p(y_1|\theta)$. Hence, the likelihood can be calculated via a product of the likelihood increments, $p(y_t|y_{1:t-1}, \theta)$, for $t = 1, \ldots, T$. We apply a sequential importance resampling algorithm, which uses importance sampling to estimate the likelihood increments [81, 82]. This algorithm uses resampling at the end of each time increment to lower the variance of likelihood estimates, and remove realisations that were unlikely to have occurred. Throughout this section all probabilities are calculated with respect to some parameter set, θ, but this notation is suppressed to make statements concise.

Let x_t denote the state of the partially-observed process at time t. The sequential importance resampling algorithm begins with a set of n particles, each particle is associated with an initial state and a weight, $\{x_0, w\}^{1:n}$. These initial states are distributed according to $p(x_0)$ and the weights begin as 1. Suppose at each iteration of the algorithm we have states $x_{t-1}^{1:n}$ distributed according to particle filter density $\hat{p}(x_{t-1}|y_{1:t-1}) \approx p(x_{t-1}|y_{1:t-1})$, we evolve these over a day according to the importance sampling density, $q(\cdot|y_t, x_{t-1})$, to obtain realisations $\tilde{x}_{(t-1,t]}^{1:n}$ with weights given by

$$w^i = \frac{p\left(\tilde{x}_{(t-1,t]}^i, y_t|x_{t-1}^i\right)}{q\left(\tilde{x}_{(t-1,t]}^i|y_t, x_{t-1}^i\right)}, \qquad \text{for } i = 1, \ldots, n.$$

These are computed such that the expected value of the weights are

$$\begin{aligned} E[w] &= \int\int \frac{p\left(\tilde{x}_{(t-1,t]}, y_t|x_{t-1}\right)}{q\left(\tilde{x}_{(t-1,t]}|y_t, x_{t-1}\right)} q\left(\tilde{x}_{(t-1,t]}|y_t, x_{t-1}\right) \\ &\quad \times \hat{p}(x_{t-1}|y_{1:t-1}) \;\; d\tilde{x}_{(t-1,t]} \;\; dx_{t-1} \\ &\approx \int\int p\left(\tilde{x}_{(t-1,t]}, y_t|x_{t-1}\right) p(x_{t-1}|y_{1:t-1}) \;\; d\tilde{x}_{(t-1,t]} \;\; dx_{t-1} \\ &= p(y_t|y_{1:t-1}). \end{aligned}$$

Hence, the mean of these weights provides an approximation of the likelihood increment, $p(y_t|y_{1:t-1})$; the product of these increments gives an unbiased estimate of the likelihood [57]. Further, resampling states from $\tilde{x}_t^{1:n}$ according to normalised weights, $w_i / \sum_j w_j$, gives a sample, $x_t^{1:n}$, from $\hat{p}(x_t|y_{1:t})$. Hence, from the initial conditions we can iterate forwards and recursively estimate the likelihood. Performance of this algorithm can be improved by updating weights over multiple steps and only resampling once the effective sample size drops below a specified threshold [83].

Thus far we have described the general approach of using importance sampling and sequential importance resampling to estimate posterior model probabilities, however, an important aspect of the sequential importance sampling process is the evolution of particles over each day. The difficulty in evolving particles in the cases considered here is that precise observations are made from models with many latent variables, which can make data-augmentation or rejection-sampling approaches slow. In Sections 5.2.2 and 5.3.3 we describe the method from [74], which builds upon [84], for generating realisations $\tilde{x}_{(t-1,t]}$ from initial state x_{t-1} for each of the case studies. The main benefit of this approach is that we use an efficient sampling distribution $q(\cdot|y_t, x_{t-1})$ which is tractable and generates $\tilde{x}_{(t-1,t]}$ which match observations almost surely, and, are similar to the partially-observed process. This method makes particles match observations by first generating observation times, then generating events between observations and occasionally forcing events to occur or blocking events to ensure feasibility of the process. For example, if we are yet to observe an infection but a recovery would lead to the epidemic ending, the recovery rate is set to 0. Importance sampling in this way allows for estimation of likelihoods associated with rare events – for example it can be easily used to estimate the tails of the likelihood, or just used to sample when observations are unlikely – whereas rejection sampling methods tend to perform poorly as simulations are unlikely to match the data.

The overall approach to model selection here is to: (i) sample from the parameter space using some importance sampling distribution; (ii) for each of those samples estimate the likelihood by sequential importance sampling; (iii) plug this likelihood estimate into Equation (5.1) to obtain importance weights; and, (iv) use the mean of these weights as an unbiased estimate of the model evidence. We can keep sampling and obtaining estimates of the model evidence until they satisfy some stopping criteria;

for example, once credible intervals are of a specified width. Once the model evidence is computed for all candidate models we can multiply these by the prior distribution over the models and normalise to obtain the posterior model probabilities.

5.2 Case study I: Inferring the shape of the infectious period distribution

Our first case study uses the importance sampling Bayesian model selection method to infer the appropriate infectious period distribution for an SI(k)R model [85–87], where symptoms are assumed to coincide with a transition into the infectious class. This study is motivated by influenza outbreaks in which only symptom onset (which is correlated with infection) is observed, and recoveries are not. Here we consider the infectious period to be either exponential, Erlang-2 or Erlang-5 distributed; these represent high, medium and low variance infectious periods respectively. This study aims to answer whether case data at a daily resolution are sufficient for discriminating between these models, how well parameters need to be known in order to discriminate between models, and how much data are required to effectively discriminate between models.

It is known that with final size data the SI(k)R model has a tractable likelihood function [87,88], which allows for efficient Bayesian model selection. But, it is currently an open question as to whether full temporal data are more effective for model selection as, although there is more information in the data, the parameter space is larger (effectively going from 2 to 3 dimensions to also infer γ). Hence, we compare model selection results from the full temporal data with results from final outbreak size data.

The temporal data considered in this chapter consists of daily symptom onset counts from completed outbreaks in small populations, which we refer to as households. Final size data are derived from the temporal data by summing over the total number of cases in each household. We let all households be of size 4 for simplicity, though this can easily be extended to allow for a distribution of household sizes. Each outbreak is modelled as a compartmental CTMC, however we only observe a small portion of the epidemic process, so this is a partially-observed CTMC. We assume that all events of symptom onset are observed until the epidemic fades out in the household and set the time of the first observation within each household as 0.

We describe the epidemic model for households in Section 5.2.1, the importance sampling method for estimating likelihood increments in Section 5.2.2, the parameters used for the simulation study in Section 5.2.3, and show results in Section 5.2.4.

5.2.1 SI(k)R model

For a population of size N, the SI(k)R model is a compartmental model that allows each individual to be in one of $k+2$ compartments: they are either susceptible; infectious in phase j, for $j = 1, \ldots, k$; or, recovered. Here the different phases of infectiousness have no physical interpretation; they are introduced in order to allow the overall infectious period to be Erlang-k distributed [85–87]. Let S, I_j and R denote the number of susceptible, infectious phase j and recovered individuals respectively. As these numbers must always be non-negative and sum to N, we have the state space

$$\mathcal{S} = \left\{ (S, I_1, \ldots, I_k, R) \in \mathbb{N}^{k+2} : S + \sum_{j=1}^{k} I_j + R = N \right\},$$

where we take $\mathbb{N}$ to contain 0. There are three kinds of transitions for this model: infection; phase change; and, recovery. Infectious individuals of all phases make effective contact with other individuals in the population at rate β, and if this contact is with a susceptible individual then that individual becomes infectious, corresponding to a transition into the infectious phase 1 compartment. An infectious phase j individual for $j = 1, \ldots, k-1$, moves into the next phase at rate $k\gamma$; this rate is chosen such that the infectious period is Erlang-k distributed with mean $1/\gamma$. Similarly, an infectious phase k individual recovers at rate $k\gamma$ and is no longer able to spread the disease. The transitions and rates associated with changes in the number of individuals in each compartment are given in Table 5.1.

For this study we assume that symptom onset corresponds to the infection transition, and we assume that we observe the number of these transitions at a daily resolution. The model is initialised at the time of the first observations, hence the initial state is $(S, I_1, I_2, \ldots, R) = (N-1, 1, 0, \ldots, 0)$.

5.2.2 Importance sampling for SI(k)R outbreaks

Suppose we have observations from M outbreaks, $y^{1:M}$. As these outbreaks occur independently we have a likelihood function which is of the form

$$p\left(y^{1:M} | \theta\right) = \prod_{i=1}^{M} p\left(y^{i} | \theta\right).$$

Table 5.1: Transitions and rates for the SI(k)R model. Only compartments that change are shown, all other compartments remain the same.

Transition Type	State Change	Rate
Infection	$(S, I_1) \to (S-1, I_1+1)$	$\frac{\beta S \sum_{j=1}^{k} I_j}{N-1}$
Phase change type j	$(I_j, I_{j+1}) \to (I_j - 1, I_{j+1}+1)$	$k\gamma I_j$
Recovery	$(I_k, R) \to (I_k - 1, R+1)$	$k\gamma I_k$

Hence, the likelihood can be estimated via multiplying estimates of likelihoods for each outbreak, $p(y^i|\theta)$. The sequential importance resampling scheme allows us to calculate each of these iteratively, so we are left to describe how to generate realisations over the day from an initial state, x_{t-1}, in a way that ensures consistency with the observations, and, how to evaluate the importance sampling weights. We do this as per the method of [74].

Suppose for each outbreak we have a dataset $y = (y_1, \ldots, y_T)$ where y_t gives the cumulative number of infection events over $(t-1, t]$ for $t = 1, \ldots, T$; note that this does not include the initial infectious individual in the population. To estimate $p(y_t|y_{1:t-1})$, we begin by uniformly generating y_t observation times over $(t-1, t]$. These times are ordered, so the joint density is that of y_t order statistics of Uniform$(t-1, t)$ random variables, so we initialise the importance weights by $w = 1/y_t!$. Then, beginning from time $\tau = t - 1$ in state,

$$\tilde{x}_\tau = (S(\tau), I_1(\tau), \ldots, R(\tau)),$$

determined by the final state of the previous iteration, we generate events between observation times. However, we only allow phase change or recovery transitions to occur, as the observations (and hence infections) have already been generated. Let a be a vector of rates,

$$a = \left(\frac{\beta S(\tau) \sum_{j=1}^{k} I_j(\tau)}{N-1}, k\gamma I_1(\tau), \ldots, k\gamma I_k(\tau) \right),$$

as per Table 5.1, associated with state $\tilde{x}_\tau$. We consider a process with modified rates b, given by

$$\begin{aligned} b_1 &= 0 \\ b_{2:k} &= a_{2:k} \\ b_{k+1} &= 1_{\{\sum_j I_j > 1\}} a_{k+1}; \end{aligned}$$

where $1_{\{A\}}$ is an indicator function which takes the value 1 if logical statement A holds or 0 otherwise. The modified rates are constructed so that no further observations (infections) can occur, as $b_1 = 0$, and no recoveries can occur if it would lead to epidemic fade-out. Let τ' denote either the next observation (or if there are no further observations over the day let $\tau' = t$). We sample an Exponential($\sum_j b_j$) candidate time increment, $\Delta\tau$. Then one of three kinds of updates occur:

(i) if $\tau + \Delta\tau < \tau'$ we generate an event at time $\tau + \Delta\tau$ and let the event be of type i with probability $b_i / \sum_j b_j$, the importance weight is updated to

$$w \leftarrow w \times \frac{a_i e^{-\sum_j a_j \Delta\tau}}{b_i e^{-\sum_j b_j \Delta\tau}},$$

we update the time $\tau \leftarrow \tau + \Delta\tau$, and we update the state to $\tilde{x}_\tau$ according to transition type i;

(ii) if $\tau + \Delta\tau > \tau'$ and $\tau' \neq t$ no event occurs in (τ, τ') and the next event is an observation, so we update the weights to

$$w \leftarrow w \times \frac{a_1 e^{-\sum_j a_j (\tau' - \tau)}}{e^{-\sum_j b_j (\tau' - \tau)}},$$

we update the time, $\tau \leftarrow \tau'$, and update the state $\tilde{x}_\tau$ according to an infection transition; or,

(iii) if $\tau + \Delta\tau > \tau'$ and $\tau' = t$ then no event occurs before the end of the day and the weights are updated to

$$w \leftarrow w \times \frac{e^{-\sum_j a_j(\tau'-\tau)}}{e^{-\sum_j b_j(\tau'-\tau)}}.$$

We repeatedly recalculate the rates and make updates until an update of type (iii) occurs, at which time the process ends for this time increment.

Once all particles have gone through this process the weights are averaged to form an approximation of $p(y_t|y_{1:t-1})$ and the states are resampled, using systematic resampling [89], according to the normalised weights. These samples provide initial states for calculating the next likelihood increment. Once the last observation in the time series has been generated, we generate from a modified process where $b = (0, a_{2:k+1})$ and update weights according to (i) until the epidemic dies out; this allows us to calculate the portion of the likelihood associated with the assumption that the epidemic died out after our last observation.

5.2.3 Implementation

We simulate 50 temporal and final size data sets of multiple completed outbreaks in households of size 4 with parameters $(\beta, \gamma) = (0.933, 2/3)$ under each of the three models. We use 500 particles per likelihood calculation, as this is found to be sufficient for low variance likelihood estimates. Our implementation uses the prior distribution as the importance sampling distribution over the parameter space, $q(\theta|y)$, however a more efficient sampler could be chosen if PM-MCMC is performed before model selection [23]. We begin by sampling 500 points from $q(\theta|y)$ and continue to sample in batches of 500 samples from the parameter space until 95% credible intervals of the model evidence are non-overlapping. The initial 500 samples are such that the central limit theorem can provides estimates of precision of model evidence with small bias. Sampling in batches allows for sample weights to be calculated efficiently in parallel before the precision of the evidence is calculated again. The stopping criterion is chosen so that we can accurately choose the most appropriate model; in practice other stopping criteria could be used to calculate model evidence to a given precision. Implementing a stopping criterion in this way ensures that the number of particles for point estimates of the likelihood only effects the run time of the algorithm, not the precision of model evidence estimates. We calculate posterior model probabilities using data from 50, 100 and 150 complete household outbreaks. For each of these data sets we consider two cases: where the mean infectious period and the reproduction number are known to a high level of accuracy *a priori*; and, where the prior distribution on model parameters is relatively uninformative. We refer to these as the assumption of tight priors and loose priors, respectively. We set the mean of the tight priors to their true value; we suppose that $1/\gamma$ has a gamma distributed prior with mean 3/2 and variance 0.01, and that β/γ has a Uniform$(0.933 \times 3/2 - 0.03, 0.933 \times 3/2 + 0.03)$ prior. For the loose priors we suppose that $1/\gamma$ has a gamma distributed prior with mean 2 and variance 0.75, and we assume that β/γ has a Uniform$(1, 2)$ prior; which is a typical range of β/γ for influenza. Note the likelihood estimates can have high variance in regions of the parameter space where the rate associated with observed events is large [74]. This issue is avoided as the prior distributions ensure that parameters only have support in places away from these values. In the simulation study the relatively low variance of likelihood estimates allowed the sample variance of the weights to remain low enough for convergence to occur.

5.2.4 Results

Our results are given in terms of box plots of the difference in posterior model probabilities of the true model and other candidate models in Figures 5.1 and 5.2, and in terms of the proportion of times the correct model was identified (having the highest posterior model probability) in Table 5.2. If the box plots are near one it means the posterior model probability for the correct model is nearly one; values that are negative represent times at which the other candidate models have higher posterior model probabilities.

For tight priors, the correct model was most often associated with the highest posterior model probabilities for each of the models considered (Table 5.2). Note that the lowest box in the panels on Figures 5.1 and 5.2 always corresponds to an adjacent model. This fits with the intuition that adjacent models are most often misidentified as each other, that is, models with more similar infectious period distributions are more difficult to discriminate. The SI(2)R model is most often misidentified; this agrees with intuition as this model's infectious period has a shape parameter between the other two (Table 5.2). Results were similar under loose priors, (Figure 5.2), however the box plots tended to have a larger range, showing that there was less certainty in the correct model.

There was also very little difference in the proportion of correct times the model was identified under loose and tight priors (Table 5.2). Even if the parameters are not well known, with multiple completed outbreaks (from small populations), infectious case data can be used to distinguish between these models, but the SI(2)R model is the most difficult to identify. Of the 900 runs, 744 estimates had converged in the first 500 samples, another 115 had converged in less than 10,000 samples and all of the samples converged in under 91,000 samples. Slow convergence occurs when two models are nearly equally likely, in this case it may not be relevant as to whether credible intervals are non-overlapping. This could be avoided by implementing a stopping rule which ensures posterior model probabilities are either non-overlapping or within a specified tolerance. Estimates based on tight priors converged at least as fast as those with loose priors in all but 13 cases.

We compare the results with those obtained by only considering the final size data of the same simulated data sets. Here sampling over the parameter space is the same, however, now the likelihood function is calculated exactly as in [87]. For all cases except for the SIR model with loose priors and 50 outbreaks, one-sided Wilcoxin signed-rank tests at the 95% level show that the posterior model probabilities of the true model are statistically significantly higher when the full temporal data are used. The proportion of times the correct model was identified from final size data are given in Table 5.2. Interestingly, for the SI(2)R model with loose priors and 50 outbreaks the correct model was identified less often than if we had uniformly randomly guessed between the models. This model also saw the biggest improvement from temporal data; the proportion of times the correct model was identified more than doubled. We find that in all cases, except for the SIR model with loose priors and 100 outbreaks, using the full temporal data increased the proportion of times that the correct model was identified. These results show that the full temporal data sets are useful for performing model selection, even though they are more computationally intensive to work with.

As the runtimes of the algorithm were reasonable and we chose a number of particles for likelihood estimates to ensure that estimates of precision of the model evidence had low bias, the number of state particles were not chosen to optimise efficiency. If one had a fixed computational budget and required estimates of model evidence to be as accurate as possible it would be important to make the algorithm as efficient as possible. As such, we have included a comparison of the coefficient of variation of model evidence

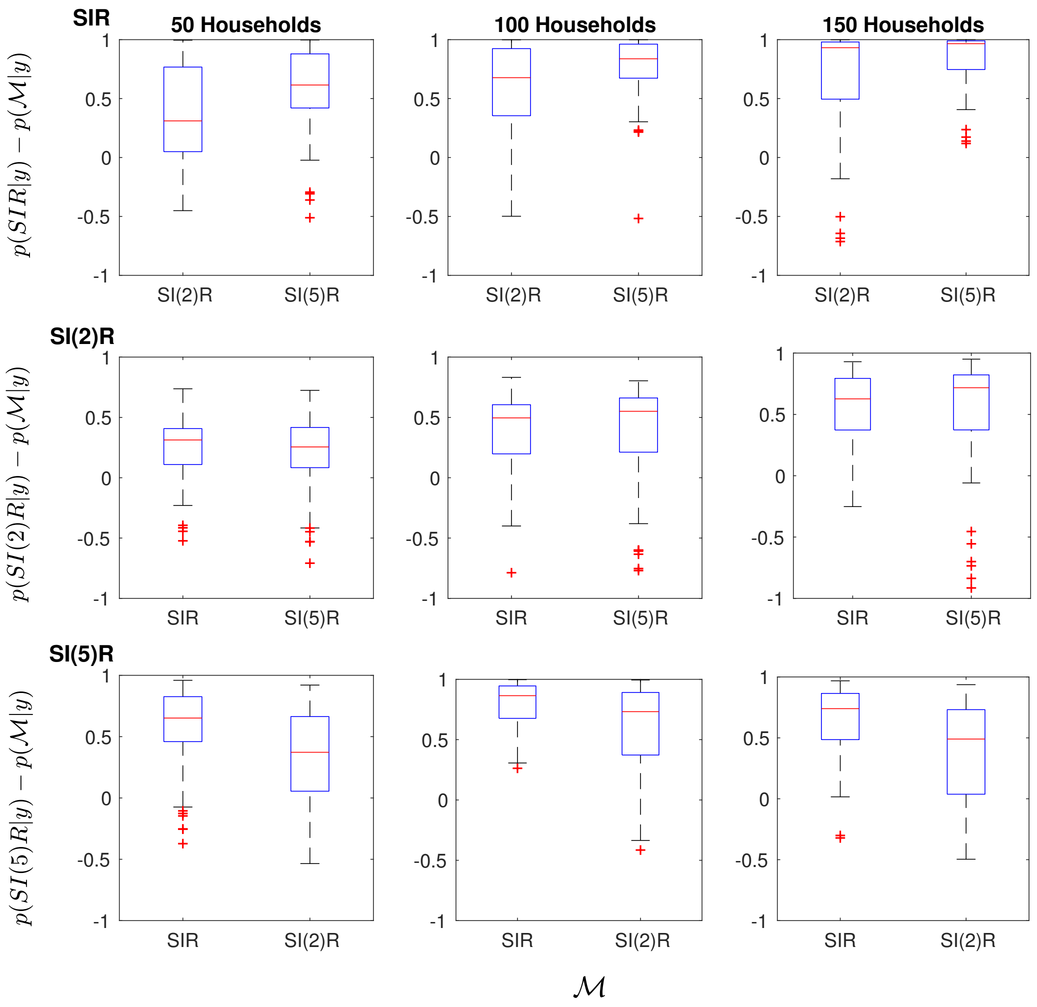

Figure 5.1: Box plots of the difference in posterior model probability of the true model and the other candidate models for the SI(k)R models with tight priors based on 50 simulated data sets. For example, in the upper left panel, boxes on the left and right of are made using 50 estimates of $p(\text{SIR}|y) - p(\text{SI(2)R}|y)$ and $p(\text{SIR}|y) - p(\text{SI(5)R}|y)$ respectively. Rows from top to bottom show results from data sets generated from the SIR, SI(2)R and SI(5)R models. Columns from left to right represent data sets containing 50, 100 and 150 independent outbreaks in households.

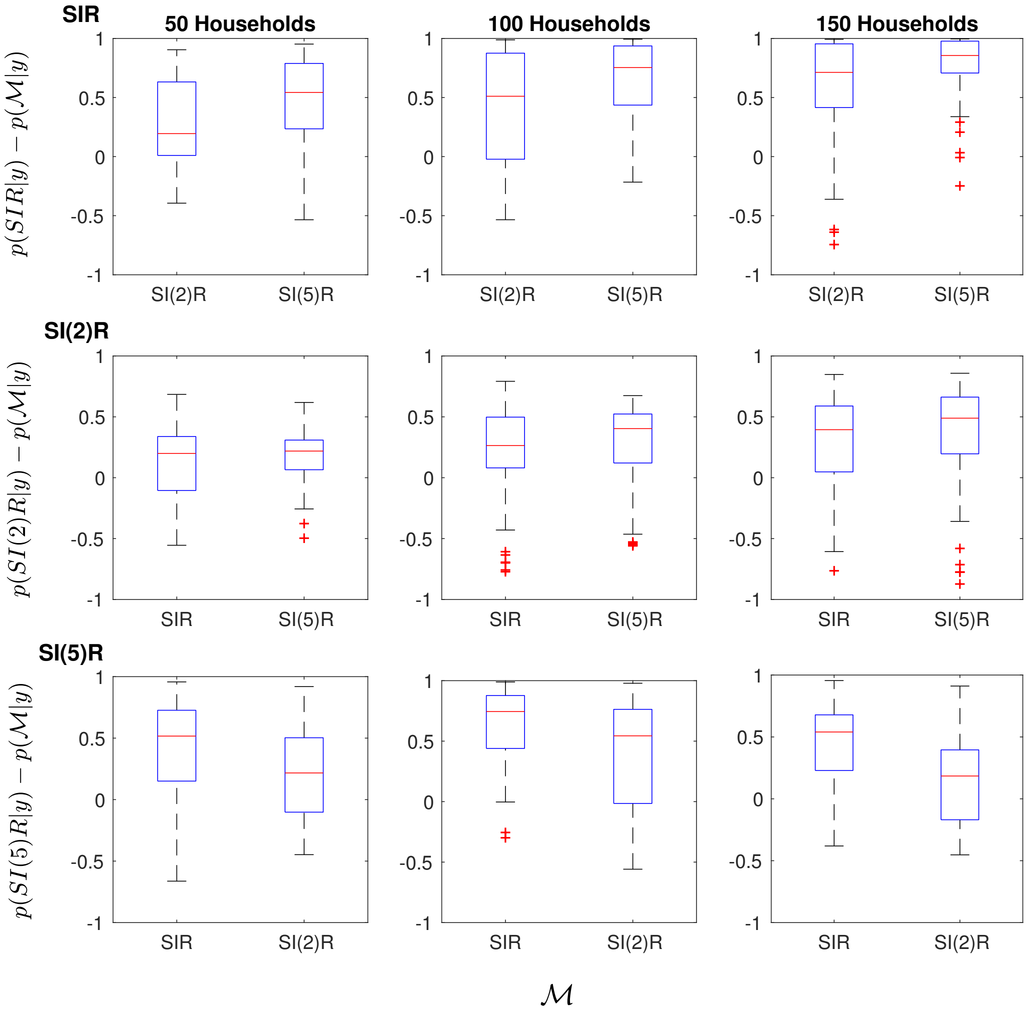

Figure 5.2: Box plots of the difference in posterior model probability of the true model and the other candidate models for the SI(k)R models with loose priors based on 50 simulated data sets. For example, in the upper left panel, boxes on the left and right of are made using 50 estimates of $p(\text{SIR}|y) - p(\text{SI(2)R}|y)$ and $p(\text{SIR}|y) - p(\text{SI(5)R}|y)$ respectively. Rows from top to bottom show results from data sets generated from the SIR, SI(2)R and SI(5)R models. Columns from left to right represent data sets containing 50, 100 and 150 independent outbreaks in households.

Table 5.2: The proportion of times the model that generated the data corresponded to the highest posterior model probability estimate from 50 data sets. Data sets were generated from each of the SIR, SI(2)R and SI(5)R models with 50, 100, and 150 outbreaks. These data sets were analysed using the full temporal data and the final size data under the assumption of loose or tight priors.

		Loose priors			Tight priors		
Number of outbreaks		50	100	150	50	100	150
Temporal data	SIR	0.76	0.72	0.82	0.78	0.86	0.86
	SI(2)R	0.44	0.6	0.62	0.58	0.72	0.8
	SI(5)R	0.68	0.74	0.86	0.8	0.84	0.96
Final size data	SIR	0.7	0.76	0.74	0.72	0.84	0.82
	SI(2)R	0.2	0.4	0.52	0.36	0.56	0.74
	SI(5)R	0.6	0.58	0.64	0.62	0.74	0.78

estimates against the number of particles used in likelihood calculations, n, in Figure 5.3. For three data sets from the SI(2)R model we ran the model selection algorithm for a fixed computation time with differing numbers of particles. The sample weights were used to estimate the coefficient of variation as $s/(\sqrt{m}\hat{p}(y))$, where m is the number of samples from the parameter space, s is the sample standard deviation of weights, and $\hat{p}(y)$ is the mean of the weights. We find that the smallest coefficient of variation is achieved at $n = 125$ for one data set and $n = 75$ for the other two. This indicates that the algorithm would run more efficiently with n around 100. As run time per sample is linear with respect to n there will be 5 times the number of samples per unit time if n were decreased to 100.

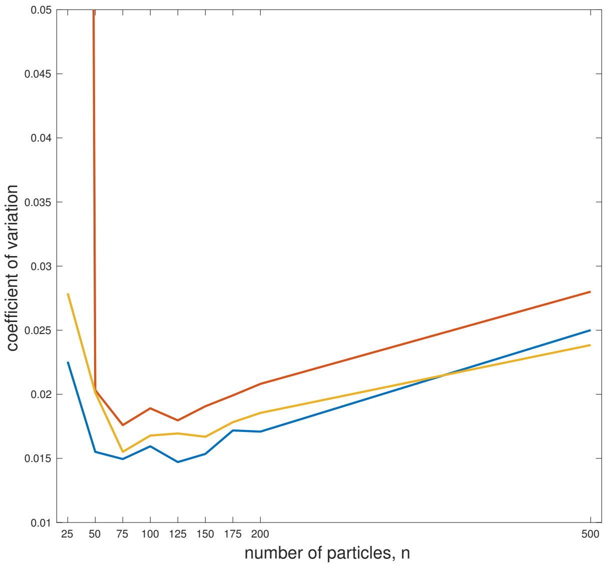

Figure 5.3: The coefficient of variation of the model evidence estimate against n, the number of particles used for likelihood calculations. For three SI(2)R data sets we estimate the coefficient of variation after running the model selection algorithm for 24 hours. The coefficient of variation is estimated as $s/(\sqrt{m}\hat{p}(y))$, where m is the number of samples from the parameter space, s is the sample standard deviation of the weights, $Z_1, \ldots, Z_m$, and $\hat{p}(y)$ is the mean of the weights.

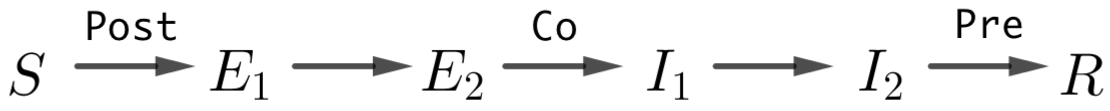

Figure 5.4: A diagram showing how we relate post-symptomatic (Post), coincidental-symptomatic (Co) and pre-symptomatic (Pre) infections to transitions in the SE(2)I(2)R model.

5.3 Case study II: Inferring symptom onset relative to infectiousness

This study considers inferring the time of infectiousness relative to symptom onset, where infectiousness could either be pre-symptomatic (Pre), coincidental-symptomatic (Co) or post-symptomatic (Post). For example, influenza may have symptoms that largely coincide with infectiousness [90], SARS is largely infectious post-symptom onset [91] and HIV is infectious pre-symptom onset [16]. We consider diseases that have a lag between exposure of individuals and infectiousness, specifically an SE(2)I(2)R model, which is detailed in Section 5.3.1. This kind of model has been used in previous work on inference using early outbreak data; it has realistic features, such as non-exponential exposed and infectious periods, while being simple enough for inference [66, 73, 92–94]. We model the observations of symptom onset as either a transition into an exposed, infectious or recovered state for the Post, Co and Pre models respectively (Figure 5.4).

5.3.1 SE(2)I(2)R model

For a population of size N, the SE(2)I(2)R model is a compartmental model that allows each individual to be in one of six compartments: they are either susceptible; exposed phase 1 or 2; infectious phase 1 or 2; or, recovered. The key differences between this model and the SI(k)R model is that the infectious period is assumed to be Erlang-2 distributed and there is an Erlang-2 distributed lag between being exposed to the disease and being able to spread it. Let S, E_i, I_j and R denote the number of susceptible, exposed phase i, infectious phase j and recovered individuals respectively.

Again, as these numbers must always be non-negative and sum to N, we have state space

$$\mathcal{S} = \left\{(S, E_1, E_2, I_1, I_2, R) \in \mathbb{N}^6 : S + E_1 + E_2 + I_1 + I_2 + R = N\right\}.$$

There are five kinds of transitions: exposure; exposed phase change; infectious; infectious phase change; and, recovery. These transitions are similar to those described for the SI(k)R model and are shown in Table 5.3.

Table 5.3: Transitions and rates for the SE(2)I(2)R model. Only compartments that change are shown, all other remain the same.

Transition Type	State Change	Rate
(1) Exposure	$(S, E_1) \to (S-1, E_1+1)$	$\frac{\beta S(I_1+I_2)}{N-1}$
(2) Exposed phase change	$(E_1, E_2) \to (E_1-1, E_2+1)$	$2\sigma E_1$
(3) Infectious	$(E_2, I_1) \to (E_2-1, I_1+1)$	$2\sigma E_2$
(4) Infectious phase change	$(I_1, I_2) \to (I_1-1, I_2+1)$	$2\gamma I_1$
(5) Recovery	$(I_2, R) \to (I_2-1, R+1)$	$2\gamma I_2$

We assume that symptom onset corresponds to a transition into either the exposed phase 1, infectious phase 1 or recovered class for the Post, Co and Pre models respectively, as shown in Figure 5.4. Hence, the process has an initial state that is either

$$(S, E_1, E_2, I_1, I_2, R) = (N-1, 1, 0, 0, 0, 0),$$

for exposure observations,

$$(S, E_1, E_2, I_1, I_2, R) = (N-1, 0, 0, 1, 0, 0),$$

for infectious observations, or a stochastic initial state for the recovery observations; this last case is discussed in the following section.

5.3.2 Importance sampling for initial state generation

For the SE(2)I(2)R model, if we observe the daily number of recovery transitions, then a population with a single exposed individual initially may be in one of several states at the time of the first observation, as multiple exposures may have occurred before the first recovery. If we see no secondary transmission in the household we can calculate the likelihood exactly as

$$p(y|\beta,\gamma,\sigma) = \left(\frac{2\gamma}{2\gamma+\beta}\right)^2 .$$

If the epidemic does not die out after the initial infection we can efficiently generate weighted initial states of the process via importance sampling.

If our data set for the household has daily observation vector $y_{1:T} = (y_1, \ldots, y_T)$ we know that each type of transition can occur at most $\psi = 1 + \sum_{t=1}^{T} y_t$ times and that recovery cannot occur if it would lead to epidemic fade out; note that the plus 1 is because the initial condition is not included in $y_{1:T}$. So if the SE(2)I(2)R process has rate vector

$$a = \left(\frac{\beta S(\tau)\left(I_1(\tau)+I_2(\tau)\right)}{N-1}, 2\sigma E_1(\tau), 2\sigma E_2(\tau), 2\gamma I_1(\tau), 2\gamma I_2(\tau)\right),$$

as per Table 5.3, we consider a modified process with rates that never lead to inconsistencies in our data. The modified rates are

$$b = a \circ \left(1_{\{z_1<\psi\}}, 1_{\{z_2<\psi\}}, 1_{\{z_3<\psi\}}, 1_{\{z_4<\psi\}}, 1_{\{z_4-z_1>1\}}\right);$$

where '$\circ$' denotes an elementwise product, z_i denotes the cumulative number of individuals that entered the ith compartment, and $1_{\{.\}}$ denotes the indicator function. The modified process is used to generate transitions until an observation occurs; the bias from simulating from this modified process is corrected for by using importance sampling weights.

The process begins with particles in state

$$(S, E_1, E_2, I_1, I_2, R) = (N-1, 0, 0, 1, 0, 0),$$

with importance weight $w = 1$. Each particle makes transitions according to probabilities $b/\sum b$, and with each transition the particles weight is updated to

$$w \leftarrow w \times \frac{a_i \sum_j b_j}{b_i \sum_j a_j},$$

where i is the type of the transition that occurred (type is numbered as per Table 5.3). When the first recovery (transition type 5) occurs the process ends. Once we have a set of particles distributed according to the initial state they are resampled according to the normalised importance weights to give initial particles for the sequential importance resampling algorithm.

5.3.3 Importance sampling for SE(2)I(2)R outbreaks

The importance sampling scheme for the SE(2)I(2)R model is similar to the SI(k)R model except that the modified rates are slightly different, and, at times, we need to force certain events to occur to ensure feasibility of samples [74]. Again, we uniformly generate observation times over $(t-1, t]$ and from time $t-1$ generate rates between observations according to a modified process with rate vector, b. We set modified rates $b_j = a_j$ for all j that correspond to rates which are not set to 0 (those set to 0 will be specified in this section). For the Post, Co and Pre models we set the modified rates $b_1 = 0$, $b_3 = 0$ and $b_5 = 0$ respectively. We also set b_5 to 0 if a recovery would lead to epidemic fadeout without the correct number of observations occurring. For these models we may also need to force events to occur to ensure that the particle generated, $\tilde{x}_{(t-1,t]}$, is a feasible realisation from an SE(2)I(2)R process. For example, for the Co model, if there are no exposed phase 2 individuals in the population and an observation is yet to occur, an exposed phase change would need to occur. If there were also no exposed phase 1 individuals then an exposure time would need to be generated before the exposed phase change; we refer to these kinds of events, and observation events, as *forced events*. More generally, if the particle is a realisation generated up to time τ, $\hat{x}_{(t-1,\tau]}$, where the next forced event occurs at time τ' and would lead to an infeasible realisation, we generate a new forced event which would allow the next forced event to be feasible. The next forced event is chosen by proposing transitions further back in the chain, or forcing an infectious event (transition type 3), until an event with a positive rate is proposed. The order in which to propose events is shown

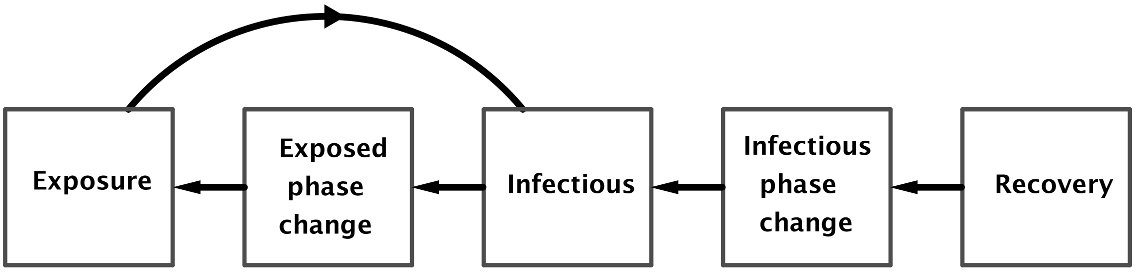

Figure 5.5: A flow chart showing the order in which transitions are proposed as forced events in order to ensure feasibility of the particle. Starting from the type of the next forced event, propose a new forced events according to arrows in the flow chart.

via a flowchart in Figure 5.5. For other CTMCs this can be done by proposing forced events according to a logic tree which is model specific. If we need to generate a forced event of type i, we generate an inter-arrival time, s, according to a $\text{TruncExp}(a_i, \tau' - \tau)$ distribution, set $b_i = 0$, set the next forced event as type i at time $\tau' = \tau + s$, where $\text{TruncExp}(a, t)$ refers to the truncated exponential distribution with rate a, truncated to $[0, t]$. The truncated exponential distribution is chosen as it has appropriate support and generates events with a similar distribution to the true process. The weights are updated according to

$$w \leftarrow w \times \frac{a_i e^{-a_i s}}{1 - e^{-a_i(\tau' - \tau)}}.$$

If no more forced events are necessary we continue to propose candidate events from time τ according to the new modified rates; as in Section 5.2.1, we let these occur, move to the next forced event, or move to the end of the time increment as per (i), (ii) and (iii) respectively. Once the cumulative number of events of type i equals the number of observations in total we set $b_i = 0$; as the epidemics are completed within households the final state must be $(N - \psi, 0, 0, 0, 0, \psi)$, so no event can occur more than ψ times. For more details and a more general description of the importance sampling scheme see [74].

5.3.4 Implementation

We simulate 20 data sets of multiple completed outbreaks in households of size 4 with parameters $(\beta, \sigma, \gamma) = (0.933, 0.5, 2/3)$ under each of the three models and calculate posterior model probabilities with data from 50, 100 and 150 completed household outbreaks. In this case we used 2000 particles per likelihood calculation; a larger number of particles were chosen because the sampling distribution is less like the true process, due to the possibility of needing to force events. Our implementation uses the prior distribution as the importance sampling distribution over the parameter space, $q(\theta|y)$. We first sample 1000 points from the parameter space and continue to sample from the parameter space until 95% credible intervals of the model evidence are non-overlapping. We consider inference based upon tight and loose priors on all model parameters. We set the mean of the tight priors to their true value; we suppose that $1/\gamma$ has a gamma distributed prior with mean 3/2 and variance 0.01, similarly we assume that $1/\sigma$ has a gamma distributed prior with mean 2 and variance 0.01 and we assume that β/γ has a $\text{Uniform}(0.933 \times 3/2 - 0.03, 0.933 \times 3/2 + 0.03)$ prior. For the loose priors we suppose that $1/\gamma$ and $1/\sigma$ have gamma distributed priors with mean 2 and variance 0.75 and we assume that $\beta/\gamma - 1$ has a gamma distributed prior with mean 1 and variance 0.75.

5.3.5 Results

Our results are given in terms of box plots of the difference in posterior model probabilities of the true model and other candidate models in Figures 5.6 and 5.7, and in terms of the proportion of times the correct model was identified in Table 5.4. The Pre model is the most difficult model to identify with data on 50 outbreaks. With data on 150 outbreaks the correct model was selected every time except for data generated from the Co model. The boxes in Figure 5.6 are all situated near 1, indicating that with tight priors, when only 50 households are infected, we are usually certain of which model is the true model. By the time 150 households are infected effectively all posterior model probabilities are close to 1, so we are almost always certain of which model is the true model. The boxes in Figure 5.7 are situated lower than those in Figure 5.6, indicating that loose priors reduces the certainty of the correct model. However, each model is easily identifiable whether or not the priors are informative and by the time 150 households have had completed outbreaks the correct model is identified in almost all cases. Figure 5.7 also shows that even with loose priors the posterior model probabilities for the correct model are almost always near 1 for Post and Pre models once 150 households are infected. It also shows that the Co model tended to be chosen with less certainty when priors were loose, with one outlier choosing an incorrect model with posterior model probability near 1. Of the 360 runs, 336 converged in the first 1000 iterations and all 360 runs converged within 5000 iterations; the fast convergence is likely due to easy identifiability of the three models.

Table 5.4: The proportion of times the model that generated the data corresponded to the highest posterior model probability estimate from 50 data sets. Data sets were generated from each of the Post, Co and Pre models with 50, 100, and 150 outbreaks. These data sets were analysed under the assumption of loose or tight priors.

	Loose Priors			Tight Priors		
Number of outbreaks	50	100	150	50	100	150
Post	0.95	1	1	0.9	1	1
Co	0.85	0.8	0.9	0.8	0.85	0.95
Pre	0.7	0.8	1	0.75	0.8	1

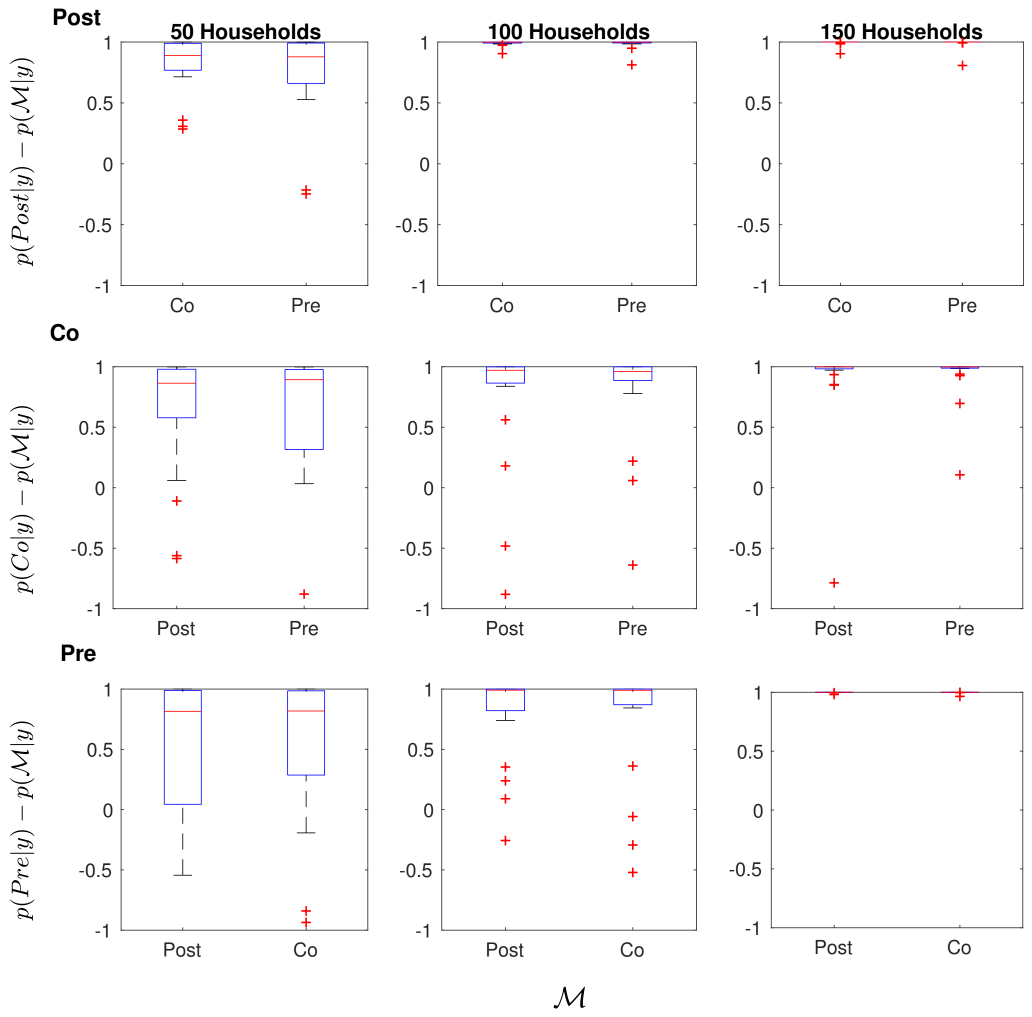

Figure 5.6: Box plots of the difference in posterior model probability of the true model and the other candidate models for the SE(2)I(2)R observation models with tight priors based on 20 simulated data sets. For example, in the upper left panel, boxes on the left and right of are made using 20 estimates of $p(\text{Post}|y) - p(\text{Co}|y)$ and $p(\text{Post}|y) - p(\text{Pre}|y)$ respectively. Rows from top to bottom show results from data sets generated from the Post, Co and Pre models. Columns from left to right represent data sets containing 50, 100 and 150 independent outbreaks in households.

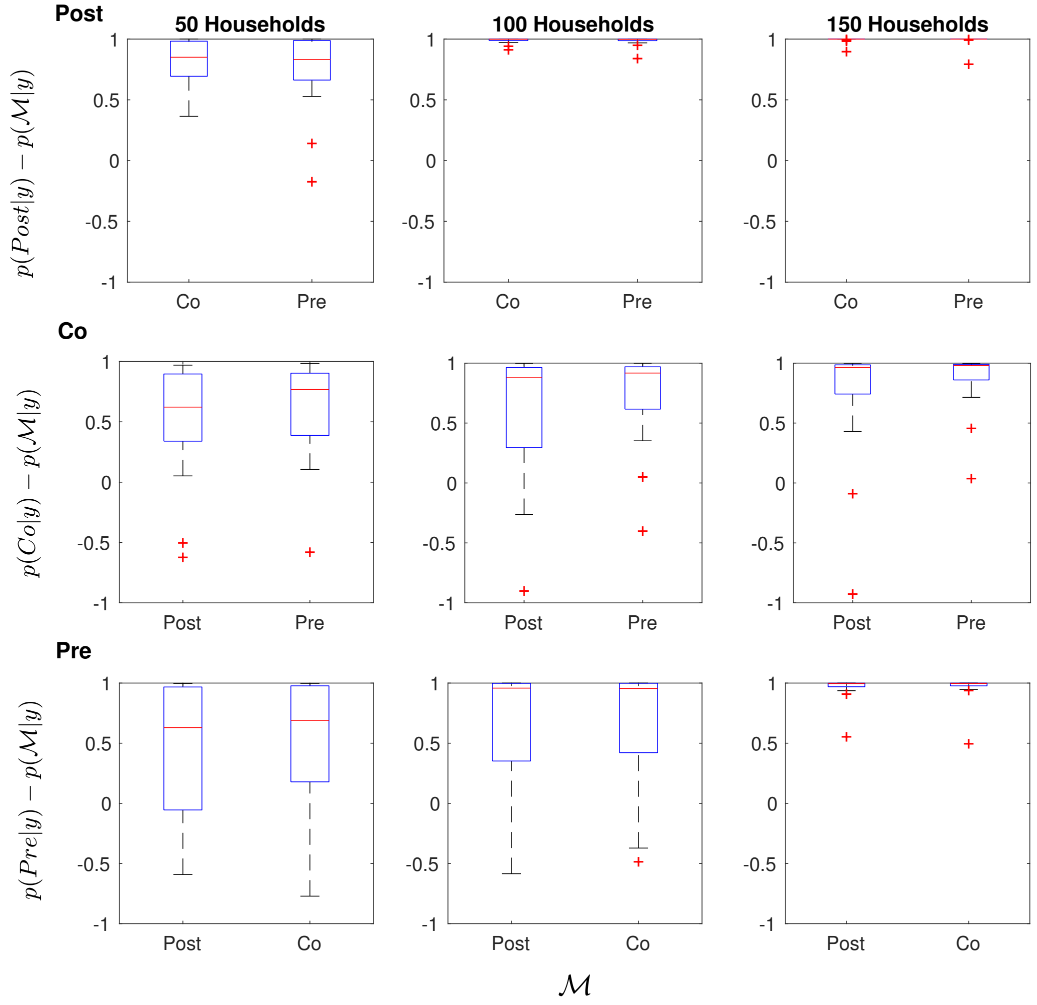

Figure 5.7: Box plots of the difference in posterior model probability of the true model and the other candidate models for the SE(2)I(2)R observation models with loose priors based on 20 simulated data sets. For example, in the upper left panel, boxes on the left and right of are made using 20 estimates of $p(\text{Post}|y) - p(\text{Co}|y)$ and $p(\text{Post}|y) - p(\text{Pre}|y)$ respectively. Rows from top to bottom show results from data sets generated from the Post, Co and Pre models. Columns from left to right represent data sets containing 50, 100 and 150 independent outbreaks in households.

5.4 Discussion

This chapter has introduced an exact method for Bayesian model selection which used importance sampling for estimating the likelihood function as well as for estimating the evidence. The novelty of this method is the use of an efficient importance sampling scheme ideal for partially-observed state space models used to estimate the likelihood function [74]. Implementation of this scheme has some overhead compared to Doob-Gillespie simulations; however, by construction all simulations agree with the observed data, hence all strictly contribute to the likelihood estimate. This is in sharp contrast to rejection-sampling approaches which do not force samples to fit with observations, such as in approximate Bayesian computation [95] or the Alive particle filter [25, 96]. Hence, it is able to greatly improve computational efficiency, particularly in estimating the tails of the likelihood. This is also an *exact* method, in that it computes unbiased estimates which converge to the target as the number of iterations increases. In contrast, other methods may not be guaranteed to converge; for example, approximate Bayesian computation, which accepts samples that do not perfectly agree with data. Further, methods may have an exact implementation, but in practice approximations are needed for computational feasibility; for example, the Alive particle filter [25, 96] can be implemented as an exact method but may require too many simulations to be practical, so the likelihood function is set to 0 in the tails instead.

If it is possible to evaluate conditional densities, a different approach is to directly calculate the evidence after use of a Gibbs sampler [97, 98]. This is problematic in that typically the Gibbs sampler for partially-observed CTMCs will require data augmentation [6, 17], making mixing slow and convergence problematic when the amount of missing data to be imputed becomes large [6, 65, 84]. Further, the conditional densities are usually only analytically tractable if priors are conjugate. In [23] data-augmentation is used to estimate the evidence, but this required a time discretisation of the model so that the forward-filtering backwards-sampling algorithm could be applied to sample from the latent process. If the model is not discretised and it is not possible to sample exactly from the latent process, data-augmentation may still be applied. However, for each point estimate of the likelihood function this requires imputing latent variables until the process reaches stationarity. Imputation is undesirable as convergence can be slow for models with high dimensional latent processes and likelihood estimates may still have a high variance. Our method circumvents this difficult imputation by instead marginalising over the latent variables in estimating the likelihood. A weakness of our approach is that we are currently restricted to Markovian models where a single event type is observed whereas data-augmented approaches are much more flexible. Current research is looking at how our method can be extended to situations where multiple types, and combinations, of events are observed.

Other methods for Bayesian model selection include reversible-jump MCMC (RJ-MCMC) [38], nested sampling [99, 100] and sequential Monte Carlo-squared (SMC-squared) [24–26]. RJ-MCMC is an algorithm in which the model is considered as one of the parameters to be sampled [38]. One issue with such an approach is that if models are non-nested it can be difficult to make sensible proposals between the different parameter spaces, particularly if data-augmentation techniques are used to address unobserved transitions [23]. Hence, at times, proposals are accepted with low probability and the mixing of the algorithm can be prohibitively slow. Nested sampling is an approach to model selection which does not require MCMC samples, however it requires a tractable likelihood [99, 100], which is uncommon for most dynamical mechanistic models. $SMC - squared$ is an alternative that jointly infers model parameters and the posterior model probabilities by using particle filters in the parameter space and in the state space [24–26]. As noted by [25], the estimates that allow for model selection may have large Monte Carlo error, but use of importance sampling after SMC-squared can be an effective way of reducing this error. During the revision process we became aware, by a personal communication, of a similar method for Bayesian model selection for partially-observed continuous-time Markov chains using importance sampling [101]. The paper uses an alive particle filter to estimate the likelihood function. The alive particle filter, in practice, introduces bias into likelihood estimates in the tails [25,101]. The particle filter used in this chapter does not introduce any bias into the likelihood function. Further, we consider different examples from epidemiology, which are important for understanding emerging infectious diseases.

Other than in [101] and this chapter, the method of importance sampling for estimating the evidence has typically only been considered for cases where the likelihood function is known [72, 80], for inference on continuous-state models [80], where data augmentation is used to estimate the likelihood function [23, 80], or in conjunction with another method which is not suitable for processes with highly variable observations [25]. This sort of implementation is inefficient for models with high dimensional parameter spaces. One way to overcome this is by using particle-marginal MCMC steps to inform a sampling distribution over the parameter space prior to model selection [23].

In this chapter we described the general method used but made no attempt to optimise the algorithm. We found that, in cases where posterior model probabilities were close, it may take many samples before estimates have non-overlapping credible intervals. Performance could be improved by implementing a stopping criterion where sampling occurs until either evidence estimates are non-overlapping or are within a specified tolerance. For the simulation study we chose a large number of state particles, n, for likelihood calculations to avoid needing a large number of samples from the parameter space. The choice of the number of particles for estimating the likelihood was chosen for reasonable performance, rather than optimal performance. There is potential to improve efficiency by informing the number of particles based on optimality criteria, such as minimising the coefficient of variation of evidence estimates after a fixed amount of computation. We estimated the coefficient of variation after running the model selection algorithm for a fixed computation time on three SI(2)R data sets with differing numbers of particles. We found that the coefficient of variation was minimised at $n = 125$ in one case and the other two were minimised at $n = 75$. If we had chosen $n = 100$, say, the runtime of the algorithm would be reduced. As run time per weight calculation is linear in n, this would represent a five-fold increase in the number of samples per unit time, though the variance of weight estimates would be larger. We also note that the models considered in this chapter are likely to give rise to household outbreak data sets that are equivalent, so computational gains could be made by re-using likelihood estimates for these data sets, as in [73]. However, these steps were not considered in this chapter as the computational benefit is model specific and is decreased when considering a slightly more complicated example, such as a model with heterogeneous household sizes. Whereas, the current implementation is versatile enough that it will readily run on models with heterogeneous household sizes.

A number of other approaches to model selection exist. Competing models are commonly discriminated using information criteria which are based on maximum likelihood estimates, such as AIC [34], AICc [102], BIC [35] or DIC [103]; note none of these account for prior information about model parameters, they depend on asymptotic results or depend on distributions being approximately Gaussian. Further, interpretation of these quantities is non-intuitive other than that they represent a function of maximum likelihood values which are penalised for each parameter in the model, and comparison of Bayesian model selection and DIC found DIC to be unreliable [104].

Our method was applied to two case studies: inferring the shape of the infectious period distribution of an SI(k)R, model and inferring the time of symptom onset relative to infectiousness for an SE(2)I(2)R model. In each of the studies we considered daily symptom onset data from completed outbreaks in multiple households. The first study showed that the data were sufficiently informative to select the correct model much more often than by randomly guessing. The study also showed that temporal data were better able to choose an appropriate model than the final size data alone, although at a higher computational cost. For example, from data generated from 150 outbreaks from the SI(5)R model with loose priors the proportion of times the correct model was identified was 0.64 from final size data compared to 0.86 from the full temporal data. Although using the full temporal data is more computationally expensive, when run on multiple CPUs the runtime is divided by the number of CPUs used, indicating that the full data sets should be used if the computational facilities are available. The SI(2)R model was the most difficult model to select correctly, intuitively this seems to be because the disease has an infectious period distribution that has a variance between the other two models. It is worth noting that the posterior model probabilities were often near 1 when the correct model was chosen and most often multiple models had reasonable support when the incorrect model was chosen. This shows that if there was insufficient information to identify the correct model, the correct model was often still given some posterior support. We find that this method does well even with datasets from 50 outbreaks in households.

In the second case study we found that the data was highly informative for being able to determine the time of symptom onset relative to infectiousness. In the case of the post-symptomatic infection model we selected the correct model every time once 100 households were infected and for the pre-symptomatic infection model we selected the correct model every time once 150 households were infected. In all cases we generally saw an increased ability to select the correct model as more data was obtained. With data on 50 outbreaks and loose priors the Pre model had the lowest proportion of correct selections at 0.7. With data on 150 outbreaks and loose priors the Co model had the lowest proportion of correct selections at 0.9. This shows that symptom onset data from multiple outbreaks is highly informative for choosing the time of symptom onset relative to infectiousness. The results of both studies show that these kinds of data sets are sufficient for discriminating between these models in the early stages of an outbreak of a novel disease, which has important implications for informing public health response [14–16].

Chapter 6

Joint Inference and Model Selection

We present a sequential Bayesian joint inference and model selection (JIMS) algorithm for partially-observed Markov chains. This is an adaptation of a standard SMC-squared algorithm (as described in Section 2.3.6) where the standard particle-marginal MCMC rejuvenation is replaced by a model selection algorithm based on importance sampling. This form of rejuvenation allows for both parameter particles and model evidence to be rejuvenated. Further, it provides a stopping criterion for the rejuvenation step, which ensures accuracy of estimates, and easily allows the sampling procedure to be adapted, which allows for fewer iterations to meet the stopping criterion. We apply JIMS algorithm and SMC-squared to simulated data sets from an SIR model and an SE(2)I(2)R model and find that JIMS algorithm out-performs a standard SMC-squared algorithm in terms of accuracy of model evidence, inferred parameters and run time. We apply JIMS algorithm to infer model parameters and infer time of symptom-onset relative to infectiousness from multiple outbreak data from an SE(2)I(2)R model.

6.1 JIMS and SMC-squared

To jointly infer the evidence of a model, $p(y_{1:t})$, and the posterior distribution, $p(\theta|y_{1:t})$, is a difficult inference problem. For partially-observed models with intractable likelihoods the only existing algorithm that does both is SMC-squared [24, 25]. Unfortunately, the evidence estimates from the SMC-squared algorithm have no guaranteed error bounds, and error is accumulated as the algorithm progresses. We present a new algorithm that switches between a particle filter for parameter inference and an importance sampling scheme for estimating model evidence. Switching between these two steps allows both model selection and parameter inference to be more accurate. It allows parameter inference to inform an appropriate importance sampling distribution for model selection; this is because the model evidence estimates converge fastest when the importance sampling density is a good approximation to $p(\theta|y_{1:t})$. The model selection component of the algorithm recalculates the model evidence to a given accuracy and provides weighted samples from the posterior distribution of interest, these provide a set of *rejuvenated* particles for the next iteration of the algorithm.

Both JIMS and SMC-squared use a particle filter to sequentially update particles in an inference step. Then, if there are too many samples from the parameter space with low posterior support (according to effective sample size criteria described in Section 2.3.4), they swap to a resampling and rejuvenation step. In standard SMC-squared the rejuvenation step occurs according to PM-MCMC [60, 79]; each particle is shifted according to a Metropolis-Hastings step where an unbiased estimate is used in place of the likelihood function. For JIMS algorithm the rejuvenation step is made according to a model selection algorithm, as presented in Chapter 5, which uses importance sampling to estimate the model evidence and resamples to rejuvenate parameters [39]. Here, the importance sampling distribution is an approximation to the posterior distribution at the current time point, which allows for efficient sampling of the parameter space. This efficient, independent, sampling allows for implementation of a stopping criterion for the rejuvenation, whereas there is no clear, sensible, stopping criterion for PM-MCMC rejuvenation (this is discussed further in Section 6.1.3).

6.1.1 Inference step

The inference step for JIMS algorithm is the same as for SMC-squared, except that likelihood estimates are not needed as they are only used in SMC-squared for PM-MCMC rejuvenation. Here, for simplicity, we assume that resampling of state particles is performed during the particle filter step, so state particles associated with a parameter particle have equal weight. In practice, this could be generalised to allow state particles to have unequal weights, which can reduce variance in parameter particle weights.

Algorithm 7 shows the inference step at iteration t for JIMS algorithm, this assumes inputs of the observation, y_t, the number of parameter particles, N_θ, the number of state particles, N_x, the ESS cutoff for rejuvenation, ζN_θ, the evidence estimate from the previous step, $\hat{p}(y_{1:t-1})$, and particles from the previous step, $\left\{\theta^i, \left[\{x_t\}_{j=1}^{N_x}\right]^i, W^i\right\}$, for $i = 1, \ldots, N_\theta$. In addition, for SMC-squared, likelihood estimates from the previous step, $\hat{p}(y_{1:t-1}|\theta^i)$, are required. The function 'particlefilter' in Algorithm 7 is assumed to be a particle filter that takes a parameter value, θ^i, equally weighted state particles, $\{x_s\}_{j=1}^{N_x}$, and observations, $y_{s:t}$, and returns an unbiased estimate of the likelihood increment, $\hat{p}(y_{s:t}|y_{1:s-1}, \theta^i)$, and updated, equally weighted, state particles, $\{x_t\}_{j=1}^{N_x}$. Our implementation uses the sequential importance resampling particle filter, where likelihood increments are estimated by the same method as described in Chapter 5.

```
Update parameter weights
for $i = 1 : N_\theta$ do
    $\left[\left[\{x_t\}_{j=1}^{N_x}\right]^i, \hat{p}\left(y_t|y_{1:t-1}, \theta^i\right)\right] = \text{particlefilter}\left(\theta^i, \left[\{x_{t-1}\}_{j=1}^{N_x}\right]^i, y_t\right)$;
    Likelihood estimate (SMC-squared only):
    $\hat{p}(y_{1:t}|\theta^i) = \hat{p}(y_{1:t-1}|\theta^i)\hat{p}(y_t|y_{1:t-1}, \theta^i)$;
    Weights: $w^i = \hat{p}(y_t|y_{1:t-1}, \theta^i)W^i$;
end
Normalised weights: $W^{1:N_\theta} = w^{1:N_\theta} / \sum_{i=1}^{N_\theta} w^i$;
Effective sample size: $\text{ESS} = 1/\sum_{i=1}^{N_\theta} \left(W^i\right)^2$;
Model evidence: $\hat{p}(y_{1:t}) = \left(\sum_{i=1}^{N_\theta} w^i\right) \hat{p}(y_{1:t-1})$;
```

Algorithm 7: Inference Step of JIMS and SMC-squared algorithms

6.1.2 Importance sampling densities

If ESS $< \zeta N_\theta$ at the 'Update model evidence' section of the inference step (Algorithm 7), rejuvenation steps are required to avoid particle degeneracy. Before the rejuvenation step is performed an appropriate importance sampling density must be chosen. JIMS algorithm uses independent samples over the parameter space to rejuvenate particles and estimate the model evidence. So, the sampling density is not of the form

$q_t(\cdot|\theta)$, as is the case for standard SMC-squared. Hence, it may be the case that JIMS scales worse for high dimensional inference problems. However, for many epidemiological models of interest the dimension of the parameter space is relatively low, so this may not be an issue.

Note the variance in estimates of the model evidence are minimised when the sampling density, $q_t(\cdot)$, is the posterior distribution [23]. Hence, we choose a sampling density based on an approximation of the posterior distribution fitted to the current set of particles. As the posterior distribution of interest is typically non-Gaussian, we approximate it by a Gaussian mixture. This approximation to the posterior is evaluated by resampling particles according to their weights and using these in an inbuilt MATLAB routine, *fitgmdist*, for fitting Gaussian mixtures via the EM algorithm [105]. The number of Gaussian distributions in the mixture, k, is decided by fitting mixtures for $k = 1, \ldots, 4$ and choosing the mixture that corresponds to the minimum AIC value. Defence mixtures are found to improve efficiency by guarding against sampling distributions with too little variance [23, 106]; these are distributions that are a mixture of an estimate of the target distribution and a distribution with heavier tails. We choose a sampling distribution which is a defence mixture that uses Gaussian mixture approximations to the posterior distribution. Suppose at iteration t the Gaussian mixture fitted to the posterior distribution is denoted $\hat{p}(\theta|y_{1:t})$, we choose a sampling density of the form

$$q_t(\theta) = p_0 p(\theta) + \sum_{s=1}^{t} p_s \hat{p}(\theta|y_{1:s}), \tag{6.1}$$

where $\{p_0, \ldots, p_t\}$ is a probability vector which is increasing, that is, $p_s < p_u$ for $s < u$. Rather than only sampling from the current approximate posterior distribution this mixture allows one to sample from a mixture of all of the approximate posterior distributions and the prior. These are weighted such that the most recent approximation to the posterior distribution is sampled from most often and the prior is sampled from least often. One choice of defence mixture is to choose a mixture of an approximation to the posterior distribution and prior [23]. The rationale behind the distribution in Equation (6.1) is that the tails should be heavier than the posterior distribution but lighter than the prior, so the sampling density will be suitable even if the posterior has much smaller support than the prior. Samples from this kind of defence mixture are displayed in Figure 6.1. The current set of equally weighted particles is represented by blue dots, samples from the sampling distribution of the form given in Equation (6.1) are represented by red dots. It appears that most of the samples from the importance sampling distribution cover the same region as the particles, however there are some samples further in the tails, as desired.

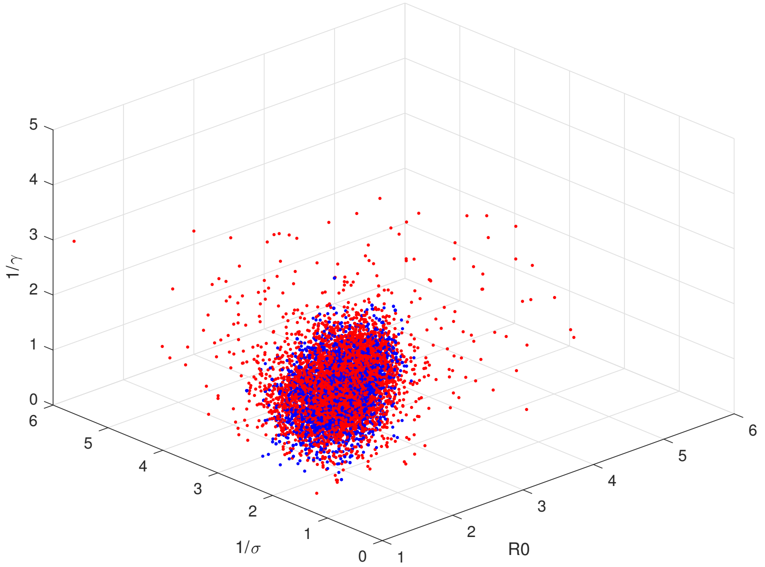

Figure 6.1: Blue dots are a sample from the posterior distribution inferred via a particle filter for a simulated data set from an SEIR model. Red dots represent sampling from a Gaussian mixture approximation based on the posterior distribution samples, mixed with the prior distribution. Here $\{p_0, \ldots, p_t\}$ is based on a truncated geometric distribution with parameter 0.2 which is reversed so that the distribution is increasing.

6.1.3 Model selection step

Once a sampling density has been evaluated the rejuvenation step may be performed. This is essentially the model selection algorithm from Chapter 5, however, the importance sampling distribution is chosen as described in the previous section and may be adapted. That is, if the stopping criterion is not met after many iterations the sampling distribution can be changed (or adapted) to reduce the number of samples required. The rejuvenation step is given in Algorithm 8, the rejuvenation for SMC-squared is given in Algorithm 9 for comparison. Algorithms 7 and 8 are combined to give JIMS algorithm in full in Algorithm 10. JIMS algorithm allows for ESS to be evaluated and for error bounds on the model evidence to be obtained, each of which allow for sensible stopping criteria to be chosen. By comparison, the ESS for standard SMC-squared is unknown and the model evidence has unknown error; so a sensible choice for the number of steps of rejuvenation, R, is unclear. It has been proposed that R could be chosen via a method which has been applied in an MCMC scheme [25], where

$$R = \frac{\log(c)}{\log(1 - p_{acc})},$$

where p_{acc} is an estimate of the acceptance probability from the previous rejuvenation step and c is a fixed chosen probability that a resampled particle does not get moved [107]. However, this rule is only appropriate if the initial value of R is sufficiently large that the acceptance probability estimate is accurate and the acceptance probability does not change sufficiently between consecutive rejuvenation steps. Unfortunately, as a particle filter is used to estimate the likelihood in SMC-squared, given a fixed number of state particles the acceptance probability will decrease between rejuvenation steps (as variance in likelihood estimates increase) [25]. If the number of state particles is increased between rejuvenation steps the acceptance probability will still be affected, though it is unclear as to whether this will increase or decrease the runtime.

```
$R = 0$;
while some condition holds do
    $R = R + 1$;
    sample $\theta^R$ from density $q_t(\cdot)$ ;
    $\left[\left[\{x_t\}_{j=1}^{N_x}\right]^R, \hat{p}\left(y_{1:t}|\theta^R\right)\right] = particlefilter\left(\theta^R, \left[\{x_0\}_{j=1}^{N_x}\right]^R, y_{1:t}\right)$;
    $W^R = \frac{\hat{p}\left(y_{1:t}|\theta^R\right)p(\theta^R)}{q_t(\theta^R)}$;
    $\hat{p}(y_{1:t}) = \frac{1}{R}\sum_{i=1}^{R} W^i$;
    ESS $= 1/\sum_{i=1}^{R}\left(W^i\right)^2$;
end
```

Algorithm 8: Rejuvenation step of JIMS Algorithm

The stopping criterion for rejuvenation in JIMS algorithm has an impact on the

Resample

Use systematic resampling to sample N_θ indices, $a_1, \ldots, a_{N_\theta}$, from $1, \ldots, N_\theta$ with weights $W^1, \ldots, W^{N_\theta}$;

Set new particles according to indicies $a_1 : a_{N_\theta}$, that is, set

$$\left\{\theta^{1:N_\theta}, \left[\{x_t\}_{j=1}^{N_x}\right]^{1:N_\theta}, W^{1:N_\theta}, \{\hat{p}(y_{1:t}|\theta^i)\}_{i=1}^{N_\theta}\right\} =$$

$$\left\{\theta^{a_1:a_{N_\theta}}, \left[\{x_t\}_{j=1}^{N_x}\right]^{a_1:a_{N_\theta}}, 1/N_\theta, \{\hat{p}(y_{1:t}|\theta^{a_i})\}_{i=1}^{N_\theta}\right\};$$

Run PM-MCMC rejuvenation

```
for $i = 1, \ldots, N_\theta$ do
    for $r = 1, \ldots, R$ do
        sample $\tilde{\theta}$ from density $q_t(\cdot|\theta^j)$ ;
        $\left[\left[\{\tilde{x}_t\}_{j=1}^{N_x}\right]^i, \hat{p}\left(y_{1:t}|\tilde{\theta}\right)\right] = particlefilter\left(\tilde{\theta}, \left[\{x_0\}_{j=1}^{N_x}\right]^i, y_{1:t}\right)$;
        Sample $u \sim \text{Uniform}(0, 1)$;
        if $\frac{\hat{p}\left(y_{1:t}|\tilde{\theta}\right)p(\tilde{\theta})q_t(\theta^i|\tilde{\theta})}{\hat{p}(y_{1:t}|\theta^i)p(\theta^i)q_t(\tilde{\theta}|\theta^i)} < u$ then
            $\theta^i = \tilde{\theta}$;
            $\hat{p}\left(y_{1:t}|\theta^i\right) = \hat{p}\left(y_{1:t}|\tilde{\theta}\right)$;
            $\left[\{x_t\}_{j=1}^{N_x}\right]^i = \left[\{\tilde{x}_t\}_{j=1}^{N_x}\right]^i$;
        end
    end
end
```

Algorithm 9: Rejuvenation step of SMC-squared algorithm

runtime of the algorithm and accuracy of estimates. In Chapter 5 we sampled until credible intervals of evidence estimates were non-overlapping. In practice, this stopping criterion is difficult to implement for JIMS algorithm; it requires that the algorithm run on all models at once and that rejuvenation is performed at the same iterations for every model, which could lead to unnecessary rejuvenation steps for some models. One sensible stopping criterion is to sample until the ESS reaches a given threshold value, say, N_θ. Another sensible choice is to sample until the credible interval, $[L, U]$, for the model evidence estimate, $\hat{p}(y_{1:t})$, is a given width. As specifying the width for the intervals can be difficult, we instead specify that the width of the credible interval be less than a proportion, ρ, of the model evidence estimate. That is

$$U - L \leq \rho \hat{p}(y_{1:t}),$$

for $\rho > 0$. As $\hat{p}(y_{1:t})$ is a probability, this rule is only sensible for $\rho < 1/\hat{p}(y_{1:t})$.

Note that some of the stopping criteria for Algorithm 8 lead to a stochastic number of parameter particles, R, whereas standard SMC-squared rejuvenation always returns a set of N_θ particles. If a constant number of particles is desired for JIMS algorithm (to constrain memory requirements, for example), a post-rejuvenation resampling step may be performed. That is, after rejuvenation is performed one can use systematic resampling to sample N_θ indices, $a_1, \ldots, a_{N_\theta}$, from $1, \ldots, R$ with weights $W^1, \ldots, W^n$. For $i = 1, \ldots, N_\theta$ the new particles are given by $\left\{\theta^i, \left[\{x_t\}_{j=1}^{N_x}\right]^i, W^i\right\} = \left\{\theta^{a_i}, \left[\{x_t\}_{j=1}^{N_x}\right]^{a_i}, 1/N_\theta\right\}$.

Rejuvenation can also be performed in a way such that the algorithm adapts. That is, the sampling procedure may be changed to reduce the number of iterations in the rejuvenation step, or possibly to reduce the number of times rejuvenation is needed. One example that has been applied to SMC-squared is to double the number of state particles, N_x, if the acceptance-rate for PM-MCMC steps drops below a threshold [26]. For JIMS algorithm there is no acceptance rate, so this kind of criterion can not be implemented. An alternative measure of efficiency is ESS divided by the number of iterations of rejuvenation. Hence, the algorithm can be adapted by calculating ESS/n and doubling N_x if this is below a threshold. Another option is to make a rule based on the width of the credible intervals of the model evidence. We also implement a criterion in JIMS algorithm which chooses a new sampling distribution $q_t(\cdot)$, if the algorithm does not converge after a given number of samples. The new $q_t(\cdot)$ is a defence mixture as given in Equation (6.1), where $\hat{p}(\theta|y_{1:t})$ is a Gaussian mixture fit to the samples from the rejuvenation step. That is, if rejuvenation is run for S iterations, resample particles and fit a Gaussian mixture to obtain a new $\hat{p}(\theta|y_{1:t})$, then begin rejuvenation again.

```
Inputs: The number of particles, $N_\theta$, the number of state particles, $N_x$, and observed data, $y_{1:T}$.
Outputs: The evidence, $\hat{p}(y_{1:T})$ and weighted particles, $\{\theta^i, W^i\}_{i=1}^{N_\theta}$.
for  $i = 1, \dots, N_\theta$ do
    Generate an initial parameter set $\theta^i \sim p(\theta)$;
    Set initial weight $W^i = 1/N_\theta$;
    For $j = 1, \dots, N_x$ generate initial state $x_0^j \sim q_0(\cdot)$;
    Resample initial states, where index $j$ is sampled in proportion to $\frac{p(x_0^j)}{q_0(x_0^j)}$;
end
for t=1:T do
    for $i = 1 : N_\theta$ do
        $\left[\left[\{x_t\}_{j=1}^{N_x}\right]^i, \hat{p}\left(y_t|y_{1:t-1}, \theta^i\right)\right] = particlefilter\left(\theta^i, \left[\{x_{t-1}\}_{j=1}^{N_x}\right]^i, y_t\right)$;
        Weights: $w^i = \hat{p}(y_t|y_{1:t-1}, \theta^i) W^i$;
    end
    Normalised weights: $W^{1:N_\theta} = w^{1:N_\theta} / \sum_{i=1}^{N_\theta} w^i$;
    Effective sample size: ESS $= 1/\sum_{i=1}^{N_\theta} (W^i)^2$;
    Model evidence: $\hat{p}(y_{1:t}) = \left(\sum_{i=1}^{N_\theta} w^i\right) \hat{p}(y_{1:t-1})$;
    if  $ESS < \zeta N_\theta$ then
        Evaluate $q_t$;
        $R = 0$;
        while some condition holds do
            $R = R + 1$;
            sample $\theta^R$ from density $q_t(\cdot)$ ;
            $\left[\left[\{x_t\}_{j=1}^{N_x}\right]^R, \hat{p}\left(y_{1:t}|\theta^R\right)\right] = particlefilter\left(\theta^R, \left[\{x_0\}_{j=1}^{N_x}\right]^R, y_{1:t}\right)$;
            $W^R = \frac{\hat{p}\left(y_{1:t}|\theta^R\right) p(\theta^R)}{q_t(\theta^R)}$;
            $\hat{p}(y_{1:t}) = \frac{1}{R} \sum_{i=1}^{R} W^i$;
            ESS $= 1/\sum_{i=1}^{R} (W^i)^2$;
        end
    end
end
```

Algorithm 10: JIMS Algorithm

6.2 Comparison based on SIR model

We compare twenty runs the standard SMC-squared algorithm with JIMS algorithm on a single, randomly selected, simulated data set from the SIR model. Note that in this section there are no competing models, we obtain model evidence estimates, which could be used for model selection, and compare the algorithms in terms of the precision of these estimates. The simulated data were from completed outbreaks in a populations of $N = 160$ individuals, with transmission rate $\beta = 1.5$ and recovery rate $\gamma = 0.6$ and consists of 143 infectious cases over 16 days. The model is initialised with a single infectious individual at time 0. We choose a prior distribution defined in terms of the reproduction number and mean infectious period, where $\beta/\gamma \sim \text{Uniform}(1, 4)$ and $1/\gamma \sim \text{Uniform}(1, 7)$.

Each of the algorithms is implemented using the same particle filter, as described in Chapter 5. The algorithms also use the same independent sampler in the rejuvenation step, a Gaussian mixture fit to particle weights as described in Section 6.1.2. The Gaussian mixtures were weighted according to a truncated geometric distribution with parameter 0.4 which is reversed so that the distribution is increasing. That is, at time t, the mixture weights, $\{p_0, \dots, p_t\}$, are given by

$$p_s = c(0.6)^{t-s},$$

where c is a normalisation constant. We set the number of parameter particles to $N_\theta = 1000$ and the number of state particles to $N_x = 50$. For the SMC-squared we set the number of PM-MCMC steps per state particle to $R = 5$. The rejuvenation step for JIMS algorithm ends when the width of the credible interval for the evidence estimate, $\hat{p}(y_{1:t})$, is less than $0.1\hat{p}(y_{1:t})$ (after an initial N_θ samples). The rejuvenation step for JIMS algorithm ends with resampling to keep a consistent number of parameter particles.

6.2.1 Results

Our results are given in terms of box plots of the model evidence estimates from the twenty runs in Figures 6.2 and 6.3, scatter plots of particle filter estimates of the mean of the reproduction number, R_0, and the mean infectious period, $1/\gamma$, from the twenty runs in Figure 6.4; and box plots of the total runtime in Figure 6.5.

The medians of boxes in Figures 6.2 and 6.3 agree well between the two methods, however the estimates of model evidence from JIMS algorithm have lower variance, as desired. Further, estimates of the mean parameter values in Figure 6.4 have lower variance, which is likely due to JIMS algorithm rejuvenating particles more effectively. Figure 6.5 shows that JIMS algorithm takes less time to run in every iteration, and by the final iteration the total runtime of SMC-squared was twice that of JIMS algorithm. The jumps in the cumulative runtime in Figure 6.5 correspond to rejuvenation steps of the algorithms, these jumps in runtime are larger for SMC-squared than for JIMS algorithm; indicating that the stopping criterion in the rejuvenation step allows the algorithm to run faster. By all measures JIMS algorithm outperformed standard SMC-squared in this case.

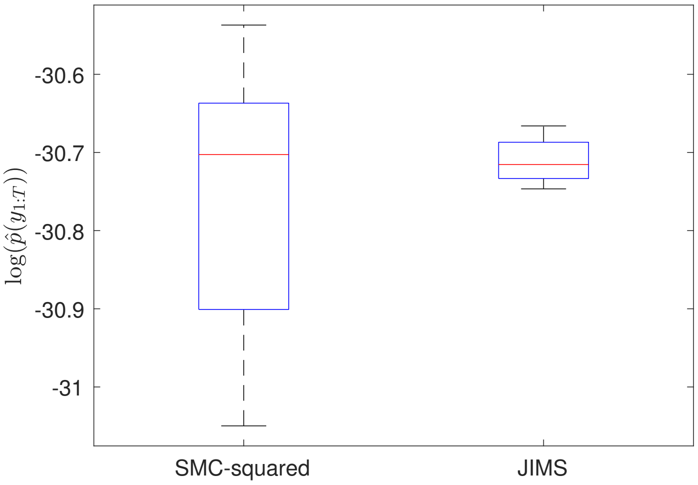

Figure 6.2: Box plots of the posterior model probability estimates from 20 runs of SMC-squared and JIMS algorithm on a single simulated SIR data set.

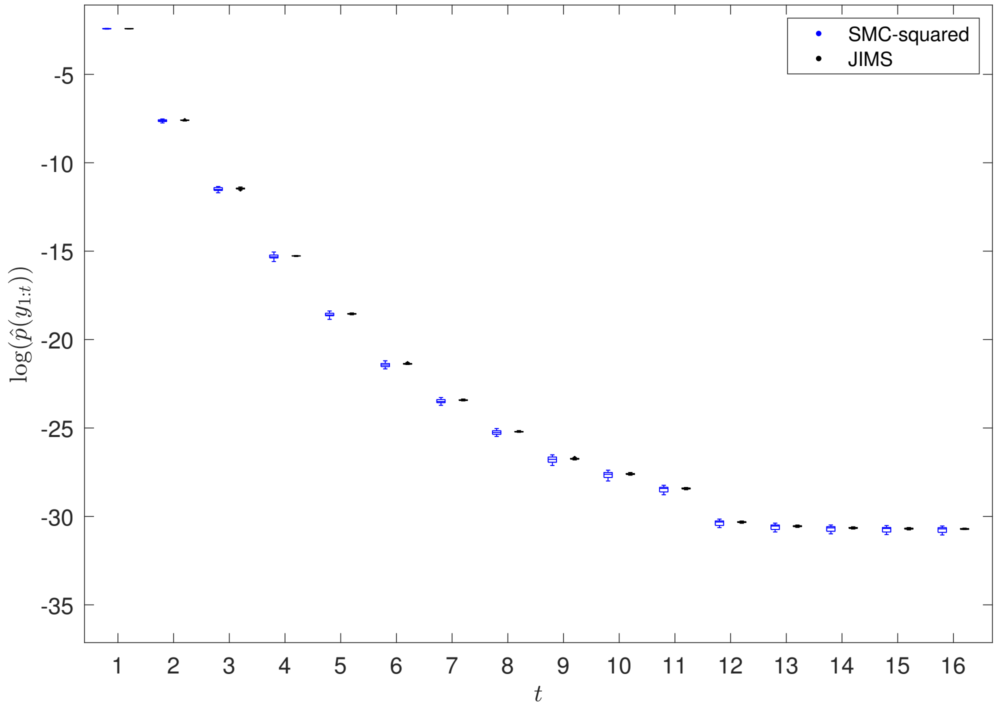

Figure 6.3: Box plots of the posterior model probability estimates at each iteration from 20 runs of SMC-squared and JIMS algorithm on a single simulated SIR data set. The black boxes are estimates from JIMS algorithm and the blue boxes are estimates from SMC-squared.

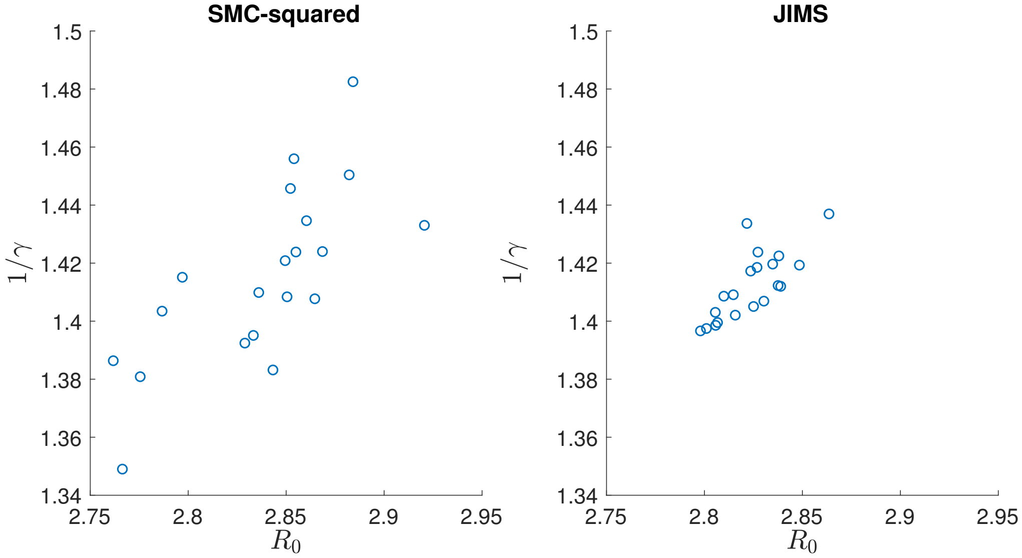

Figure 6.4: Scatter plots of the mean estimates of R_0 and $1/\gamma$ from 20 runs of SMC-squared and JIMS algorithm on a single simulated SIR data set. The true parameter values were $R_0 = 2.5$ and $1/\gamma \approx 1.667$.

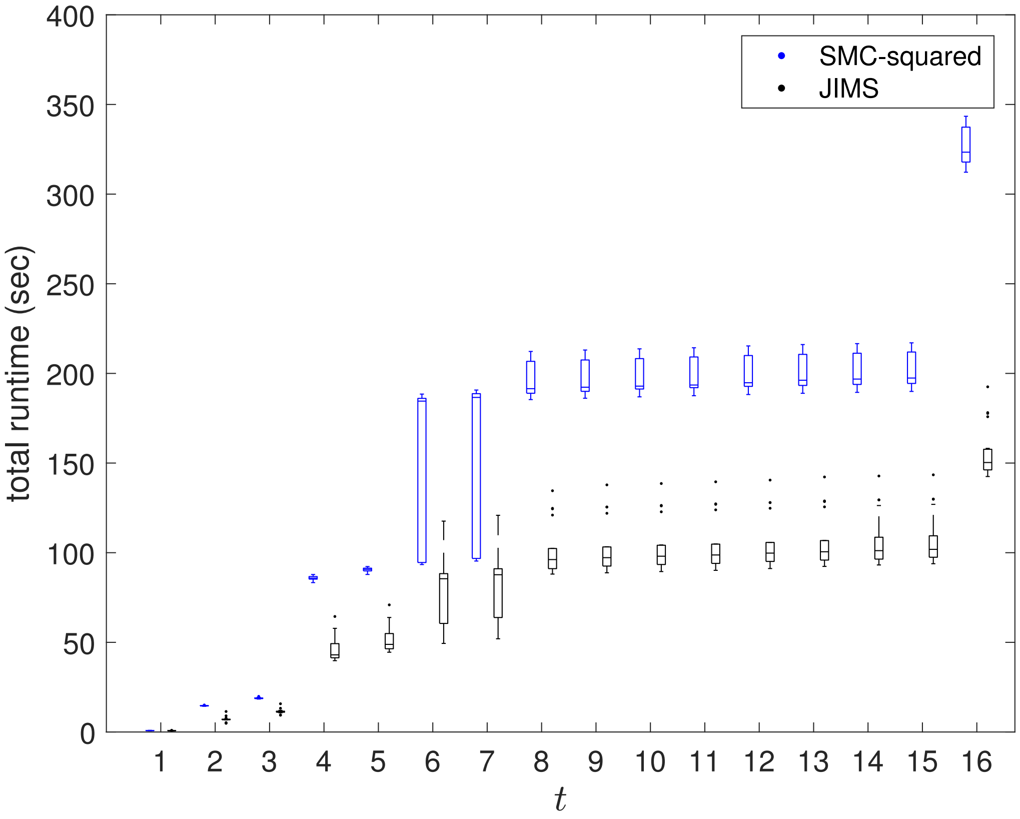

Figure 6.5: Boxplots of the total runtime, in seconds, from 20 runs of SMC-squared and JIMS algorithm on a single simulated SIR data set. The total runtime is given for each iteration of the algorithm.

6.3 Comparison based on SE(2)I(2)R model

We compare twenty runs the standard SMC-squared algorithm with JIMS algorithm on a single, randomly selected, simulated data set from the SE(2)I(2)R model. Again, in this section there are no competing models, we obtain model evidence estimates under one model and compare the algorithms in terms of the precision of these estimates. The simulated data were from a completed outbreak in a population of $N = 300$ individuals, with transmission rate $\beta = 0.933$, infectiousness rate $\sigma = 0.5$, and recovery rate $\gamma = 0.6$, and consists of 184 infectious cases over 36 days. The model is initialised with a single infectious individual at time 0. We choose a prior distribution defined in terms of the reproduction number, mean latent period and mean infectious period, where $R_0 - 1$, $1/\sigma$ and $1/\gamma$ are each gamma distributed with means 2 and variances 0.75.

Again, the algorithms are implemented using the same particle filter, as described in Chapter 5 and an independent sampler is chosen for the rejuvenation step in the same way as in the previous section. We set the number of parameter particles to $N_\theta = 2000$ and the number of state particles to $N_x = 50$. For the SMC-squared algorithm we set the number of PM-MCMC steps per state particle to $R = 5$. The rejuvenation step for JIMS algorithm ends when the width of the credible interval for the evidence estimate, $\hat{p}(y_{1:t})$, is less than $0.1\hat{p}(y_{1:t})$ (after an initial N_θ samples). The rejuvenation step for JIMS algorithm ends with resampling to keep a consistent number of parameter particles.

6.3.1 Results

Our results are given in terms of box plots of the model evidence estimates from the twenty runs in Figure 6.6; box plots of the mean estimates of the reproduction number, R_0, the mean exposed period, $1/\sigma$, and the mean infectious period, $1/\gamma$, in Figure 6.7; variances of the mean estimates in Figure 6.8; and box plots of the total runtime in Figure 6.9.

Again, the medians of boxes in Figure 6.6 agree well between the two methods, however, estimates of model evidence from JIMS algorithm have lower variance, as desired. Figure 6.4 shows that the estimates of the mean parameter values in agreed well at all time points between the two methods, however, Figure 6.8 shows that the estimates from JIMS algorithm have lower variance. Lastly, Figure 6.5 shows that JIMS algorithm takes less time to run in every iteration, and by the final iteration the total runtime of SMC-squared was generally around three times that of JIMS algorithm. By all measures JIMS algorithm outperformed standard SMC-squared in this case.

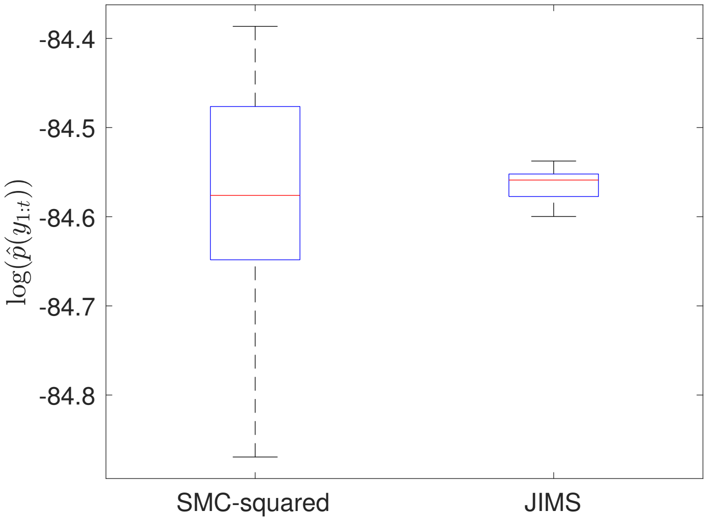

Figure 6.6: Box plots of the posterior model probability estimates from 20 runs of SMC-squared and JIMS algorithm on a single simulated SE(2)I(2)R data set.

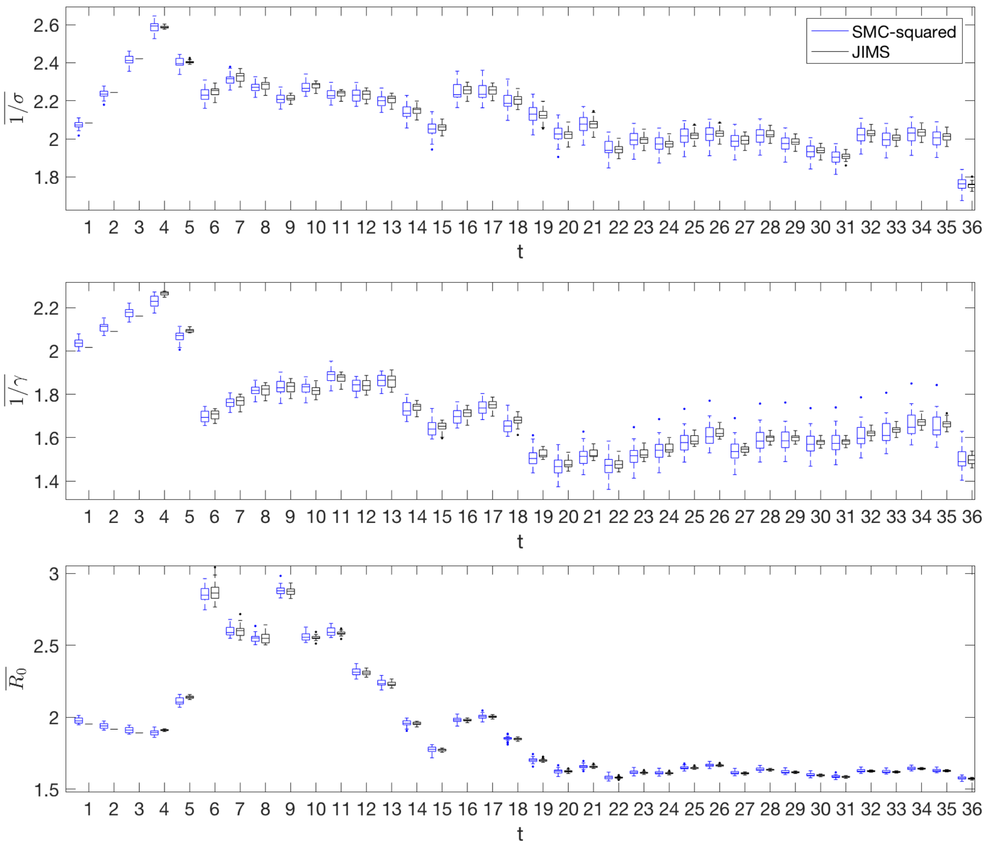

Figure 6.7: Box plots of the mean estimates of $1/\sigma$, $1/\gamma$, β/γ from 20 runs of SMC-squared and JIMS algorithm on a single simulated SE(2)I(2)R data set. The true parameter values were $R_0 = 1.555$, $1/\gamma \approx 1.667$ and $1/\sigma = 2$.

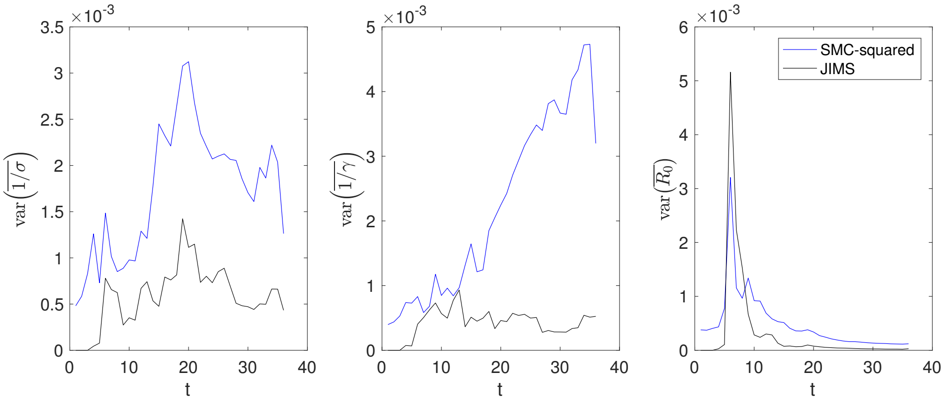

Figure 6.8: Variance estimates of the mean estimates of $1/\sigma$, $1/\gamma$ and β/γ from 20 runs of SMC-squared and JIMS algorithm on a single simulated SE(2)I(2)R data set.

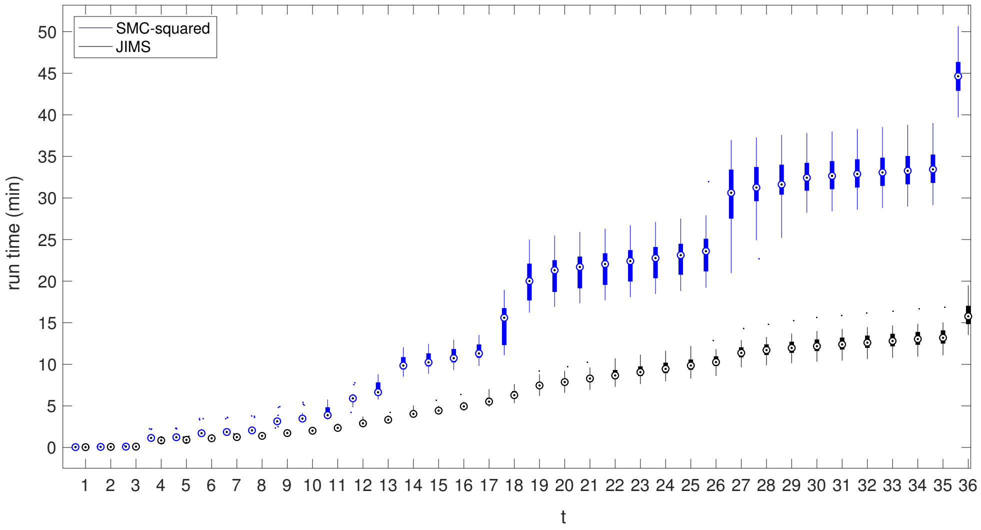

Figure 6.9: Box plots of run times, in minutes, from 20 runs of SMC-squared and JIMS algorithm on a single simulated SE(2)I(2)R data set. The total runtime is given for each iteration of the algorithm. Runtimes from JIMS algorithm is given by black boxes and runtimes of SMC-squared are given by blue boxes.

6.4 Application to multiple outbreaks data

We employ JIMS algorithm to jointly infer model parameters and select between models for multiple outbreaks data. Here we apply JIMS algorithm to perform model selection on few large outbreaks from an SE(2)I(2)R model and infer the time of symptom onset from daily case data. We apply model selection to identify whether observed symptoms correspond to a transition from the E_1 to E_2 class, the E_2 to I_1 class, or, the I_1 to I_2 class. This extends from Chapter 5 in that it considers larger outbreaks (in populations of size 100) with some longer time series, which is a more computationally challenging problem. Further, it attempts to distinguish between models which are more similar than in Chapter 5, so model evidences are likely to be closer. Hence, estimates of model evidence will need to be inferred more precisely to effectively choose between models.

6.4.1 Implementation

The model considered here is a modified version of the SE(2)I(2)R multiple outbreaks models considered in Chapter 4, with transmission rate $\beta = 0.933$, infectiousness rate $\sigma = 0.5$, and recovery rate $\gamma = 2/3$. Rather than considering a fixed number of completed outbreaks, we let the number of outbreaks be stochastic and simulate new outbreaks in populations of 100 individuals until at least 300 individuals are infected. Each subpopulation is initialised with a single infection at time 0. As outbreaks in subpopulations are independent, their seeding time only affects inference in terms of the time between the seeding time and the first observation (which is assumed to be 0 here). This ensures that data sets contain at least three hundred cases, so more realistic complete FF100 study data sets are generated. This is also more realistic with respect to how resources may be used; if there is a single infection in a population, data collection will be less resource intensive than an outbreak with many individuals. That is, the size of the outbreak has an impact on the resources left for surveillance.

For each of the three models we simulate 30 data sets and run JIMS algorithm to estimate the model evidence and infer parameters for each of the three models. We choose a prior distribution defined in terms of the reproduction number, mean exposed period and mean infectious period, where $R_0 - 1$, $1/\sigma$ and $1/\gamma$ are each gamma distributed with means 2 and variances 0.75. The number of outbreaks in simulated data sets varied from 7 up to 29 outbreaks with a maximum duration of 63 days. The chosen parameters were such that epidemic fadeout was likely, so many of these 'outbreaks' only contain one or two cases.

Again, we use the same particle filter, as described in Chapter 5. We choose an independent sampler in the rejuvenation step in the same way as in the previous section. We set the number of parameter particles to $N_\theta = 2000$ and the number of state particles to $N_x = 200$. The rejuvenation step for JIMS algorithm ends when the width of the credible interval for the evidence estimate, $\hat{p}(y_{1:t})$, is less than $0.2\hat{p}(y_{1:t})$ (after an initial N_θ samples). After $5N_\theta$ samples of rejuvenation the sampling distribution is adapted if $ESS < 0.2N_\theta$, that is, the samples from the rejuvenation step are resampled according to their weights and used to fit a multivariate Gaussian mixture to approximate the posterior distribution. The rejuvenation step for JIMS algorithm ends with resampling to keep a consistent number of parameter particles.

6.4.2 Results

Bar graphs of posterior model probability estimates and kernel density estimates of marginal posterior distributions of R_0, $1/\sigma$ and $1/\gamma$ are shown for the three simulations over time in Figures 6.10 and 6.11. In each case the highest posterior model probability estimate corresponded to the model that generated the data by the final time point. For each of these simulated data sets the $E_1 \to E_2$ is always preferred after $t = 1$. The $E_2 \to I_1$ model, at times, preferred the $E_1 \to E_2$ model, but by day 40 the preferred model was the correct model. The $I_1 \to I_2$ model prefers the $E_1 \to E_2$ model for some early time points, however, after time $t = 10$ the correct model is clearly preferred. In all cases the kernel density estimates peak near the true model parameters, in particular, the posterior distribution of R_0 peaks near the true value and with relatively low variance.

Boxplots of expected values of r, R_0, $1/\sigma$ and $1/\gamma$ evaluated as the means from each candidate model weighted by the posterior model probabilities are given in Figure 6.12. The boxes of estimates of transmissibility, r and R_0, are generally centred around the true parameter values. Boxplots of estimates of $1/\sigma$ and $1/\gamma$ appear negatively and positively biased respectively. For the $E_1 \rightarrow E_2$, $E_2 \rightarrow I_1$ and $I_1 \rightarrow I_2$ models the true model corresponded with the highest posterior model probability in 19, 20 and 19 of the 30 simulated data sets respectively.

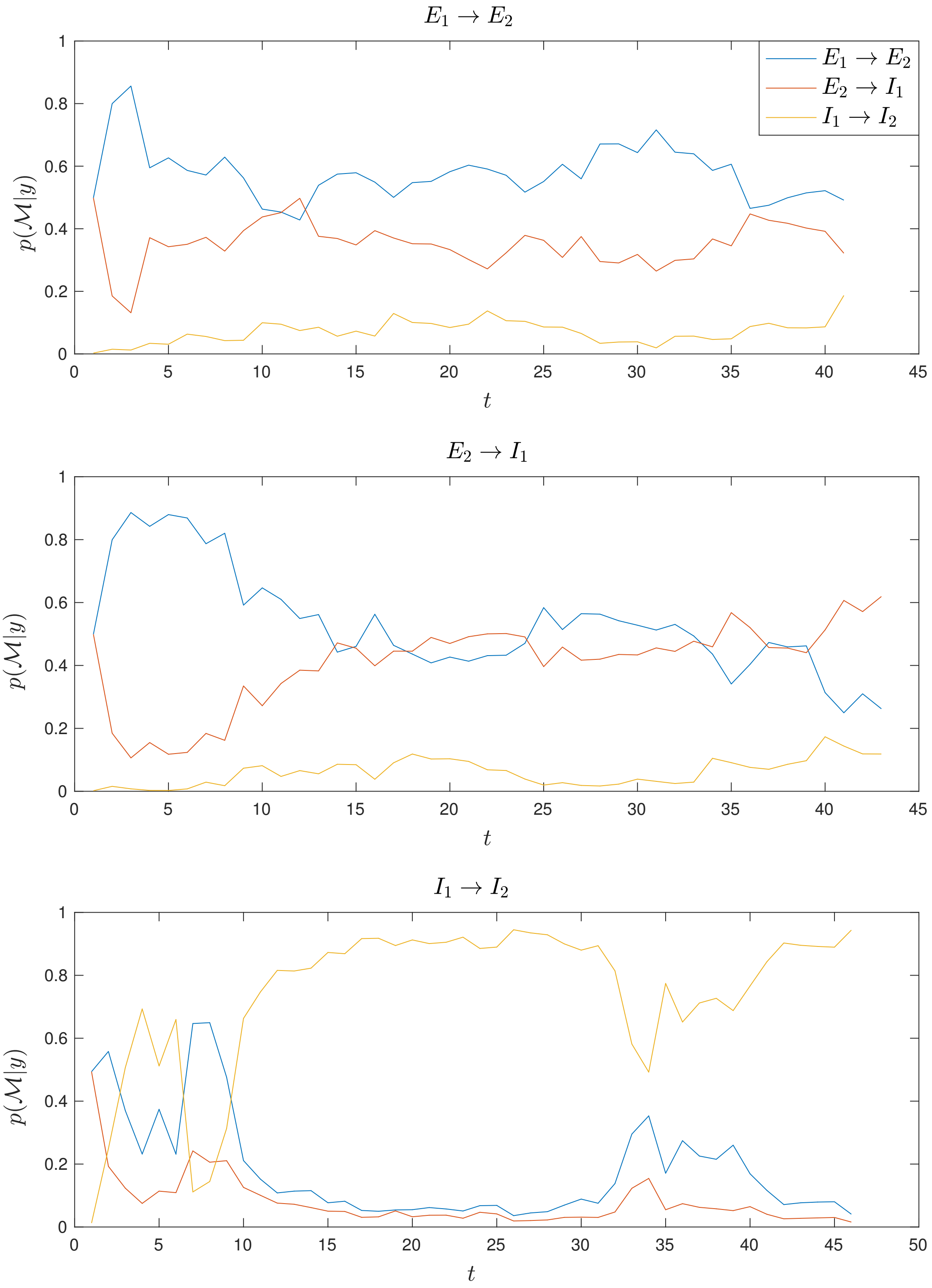

Figure 6.10: Plots of posterior model probabilities over time from simulated SE(2)I(2)R multiple outbreak data for three models.

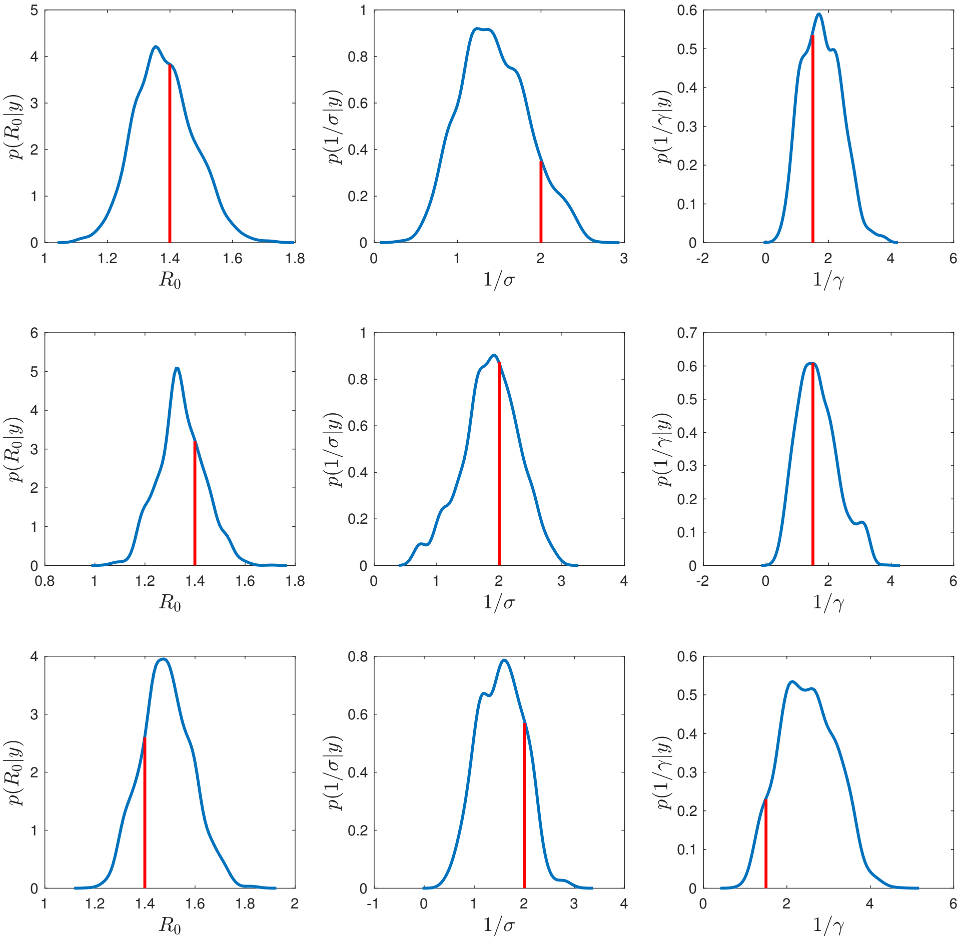

Figure 6.11: Kernel density estimates of the marginal posterior of R_0, $1/\sigma$ and $1/\gamma$. The input parameter values for the simulated data sets, $(R_0, 1/\sigma, 1/\gamma) = (1.3995, 2, 1.5)$, are marked by a red line in each panel.

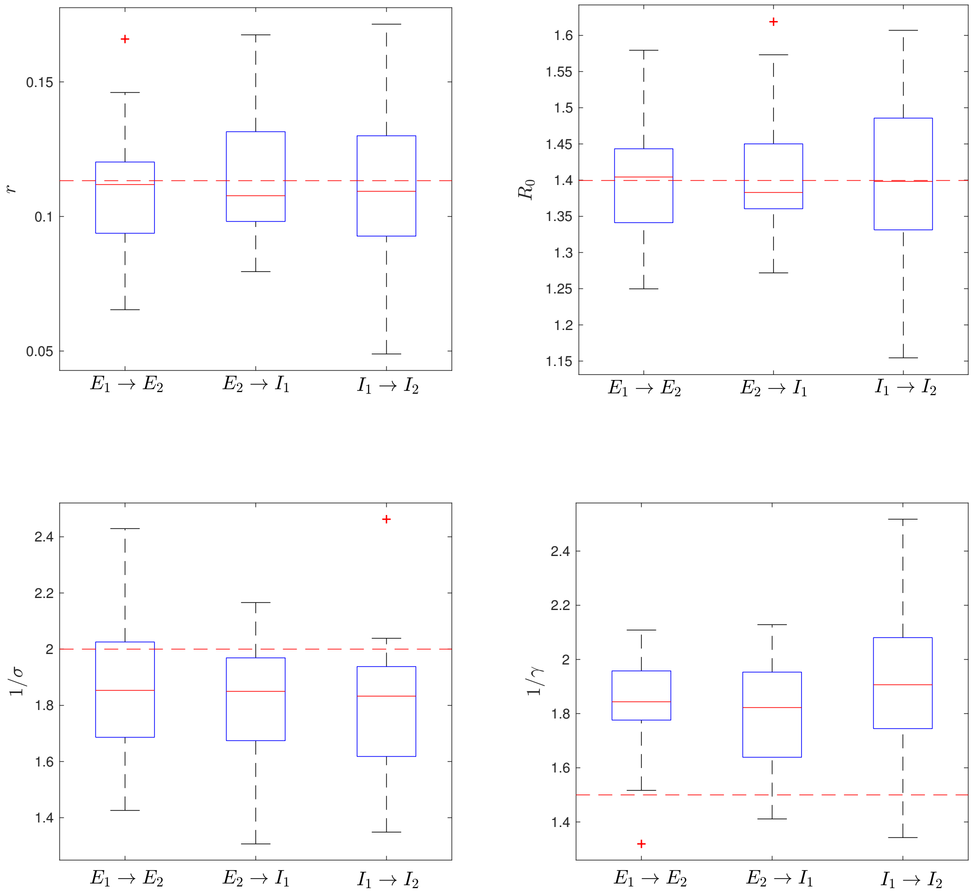

Figure 6.12: Box plots of the means of r, R_0, $1/\sigma$ and $1/\gamma$ from 30 simulated data sets. Each estimate is the sum of means from each model weighted by posterior model probabilities. The input parameter values for the simulated data sets, $(r, R_0, 1/\sigma, 1/\gamma) = (0.1133, 1.3995, 2, 1.5)$, are marked by a red dashed line in each panel.

6.5 Discussion

Sections 6.2.1 and 6.3.1 compared JIMS algorithm with SMC-squared in a way which used the same samplers, the same particle filters and the same number of particles to allow for a fair and clear comparison. Our results show that JIMS algorithm provides an effective way of rejuvenating particles, via resampling and using a stopping criterion as opposed to performing a specified number of PM-MCMC steps. This rejuvenation allowed model evidence to be estimated accurately (by design), but also had flow on effects in terms of low variance parameter estimates and faster run times. The ability to rejuvenate and ensure estimates of the model evidence allows for estimates to be inferred precisely enough to distinguish between models. The lack of precision in model evidence can be an issue with SMC-squared, as highlighted in [25], where the variance in model evidence estimates was too high to precisely identify the preferred model. Further, identifying this issue with model evidence estimates required multiple runs of SMC-squared, whereas the error in model evidence can be estimated in a single run of JIMS algorithm.

Section 6.4 employed JIMS algorithm to perform model selection and infer parameters based on multiple outbreaks from an SE(2)I(2)R model in populations of size 100 where the time of symptom onset corresponded to a $E_1 \to E_2$, $E_2 \to I_2$, or $E_1 \to E_2$ transition. This is a difficult model selection problem due to the size of the data set in terms of the number of outbreaks, the number of cases and the length of the time series (the longest outbreak went for 69 days). The length of the time series, in particular, poses a challenge as it leads to rejuvenation steps requiring many samples. However, JIMS algorithm was able to estimate the model evidence precisely enough to choose between models effectively, and, for these simulations the data were informative enough that the correct model was chosen in at least 19 out of 30 simulations from each model. Further, JIMS algorithm returned unbiased estimates of measures of transmissability, the early growth rate and reproduction number. Estimates of $1/\sigma$ and $1/\gamma$ appear negatively and positively biased respectively; this is due to a lack of identifiability in model parameters and agrees with the results from [73]. We note that the kernel density estimates of the posterior distributions from JIMS algorithm were not particularly smooth; increasing the number state and parameter particles could smooth these and would lead to more precise estimates. A sensible method is to double the number of state particles, N_x, in the rejuvenation step, as suggested in [24]. For JIMS algorithm N_x could be doubled when for the jth iteration of rejuvenation ESS$/j$ drops below a threshold. This was not implemented here so that overly large arrays did not need to be stored. For these data sets the computational requirements from JIMS algorithm were intensive and highly stochastic; with 25 CPUs, 2000 parameter particles and 200 state particles, the longest job ran in just under a day and a half and the shortest job ran in just under an hour.

This chapter presented only one way of implementing SMC-squared and JIMS algorithm, and highlighted a number of alternative implementations. The stopping criterion in the rejuvenation step in JIMS algorithm were based on estimates of the error in model evidence, but alternatively could be based on estimates of the effective sample size during rejuvenation. The rejuvenation step in JIMS algorithm was also followed by a resampling step, which in practice may not always be necessary. The algorithms were run with a fixed number of state particles, however, it is proposed that these be increased as the algorithm progresses to reduce variance in likelihood estimates. The comparison of SMC-squared and JIMS algorithm in Sections 6.2.1 and 6.3.1 were based on an SMC-squared algorithm with a fixed number of rejuvenation steps per particle, R. In practice the number of rejuvenation steps could be adapted based on estimates of the acceptance probability from a previous step as proposed in [25]. It is unclear as to how effective this rule for adapting R will be in practice for SMC-squared as the likelihood values are estimated via a particle filter, which makes the acceptance probability decrease between rejuvenation steps. The proposal distribution implemented in this chapter was chosen to be sensible and to allow for reasonable performance. However, weights of the Gaussian mixture approximations of the posterior distributions at each time point were chosen in an ad-hoc way (based on a truncated geometric distribution). These weights were chosen to ensure that the approximation to the posterior distribution from the previous iteration was sampled more often than those from earlier iterations. An alternative is to weigh these distributions based upon ESS from that iteration, which could account for iterations where the posterior distribution approximation has high error. Although the sampling distribution was implemented in a way that stabilises the particle filter, large changes in the posterior distribution could make the algorithm inefficient; this would likely occur when applied to inference problems with high dimensional parameter spaces.

Chapter 7

Concluding Remarks

This thesis has developed methods that use First Few Hundred (FF100) study-type data to accurately characterise an emerging infectious disease. The data consisted of daily cases of symptom onset stratified by the household of infectious individuals. When an individual in the population is identified with symptoms they, and members within their household, may be surveilled. A natural description of an epidemic model that uses this kind of household stratified data is stochastic and either incorporates household structure or considers independent outbreaks in households. As there is a lack of real FF100 study data, this thesis used simulation studies to assess methods for characterising diseases. As the epidemic process which generated the data is known this allowed results to be validated.

From simulated FF100 study data this thesis characterised transmissibility by performing Bayesian model selection (to appropriately describe disease dynamics) and Bayesian inference (to quantify transmissibility). These two tasks are made difficult by the fact that underlying disease dynamics in a population are largely unobserved, which makes direct point-wise evaluation of the likelihood function (the probability of observations given a parameter set and model) difficult. This thesis circumvented this issue by considering data-augmentation and particle filter methods which allow asymptotically exact Bayesian inference and model selection to be performed via sampling.

This thesis contributes to the mathematical epidemiology literature and provides tools and insight into pandemic preparedness. It describes and implements efficient, Bayesian, methods for inferring parameters of household epidemic models, and models of independent outbreaks in households, from FF100 study data. We provide insight into how a population should be surveilled to most accurately and robustly characterise a disease; this highlights the importance of model selection for pandemic preparedness. This thesis develops new efficient Bayesian model selection algorithms that use importance sampling, particle filters and constrained simulation to infer model evidence (and thus the probability of observations under a model) to a given accuracy. These methods were applied to two model selection problems: inferring the shape of the infectious period distribution; and, inferring the time of symptom onset relative to infectiousness. We have shown that FF100 study data are informative enough to effectively identify these important model characteristics, which has implications for the controllability of emerging infectious diseases.

This thesis developed and implemented efficient methods for inferring parameters of partially-observed stochastic epidemic models on households or independent outbreaks in Chapter 3 using FF100 type data. First, we implemented efficient data-augmented MCMC (DA-MCMC) for an SIR household model with households of size three, and, a SE(2)I(2)R household model with the same household distribution as Adelaide, Australia. The DA-MCMC method for the SIR household model was compared with a branching process approximation (BPA) that was developed earlier. We found that DA-MCMC, as an exact method, had lower bias than the BPA method, but was less efficient. For the SE(2)I(2)R model we found that it was feasible to get accurate estimates of the household reproduction number, early growth rate from FF100 study data. We also implemented a sequential version of DA-MCMC which used ideas from sequential Monte Carlo to sequentially update estimates of the posterior distribution as new data are acquired. This was found to be more efficient than DA-MCMC when applied to a single outbreak from a homogeneous SIR model. It also was able to perform inference on the infectious period shape from multiple outbreaks from an SI(n)R model, where mixing of DA-MCMC was prohibitively slow.

Applying DA-MCMC to household models was computationally taxing; for the SE(2)I(2)R household model this involved multiple independent runs on a cluster to obtain a sufficient number of samples from the posterior distribution after burn-in. These computational issues are inflated when considering models with unobserved cases, as they increases the dimension of the parameter space, increase the dimension of the hidden process, and increases the correlation between consecutive samples. More complex models may be fit by improving computational efficiency via non-centering [55] or optimisation of the number of Hastings steps per iteration. The sequential DA-MCMC was only briefly considered due to limitations that prevent it from being applied to more complex models. It depends on calculation of a kernel density estimate from the samples of the posterior distribution at each time point and point estimates from these kernel densities need to be made at each iteration. This is a computational challenge which prevents this method from being more widely applicable when the dimension increases. For example, we were unable to successfully apply this method to the homogeneous SEIR model.

Chapter 4 conducted an analysis on the optimal way to surveil a population given a fixed total number of individuals to surveil. This was done by simulating multiple outbreak data on a six hundred individuals divided into subpopulations of different sizes (representing, for example, households, schools and workplaces). This leads to a trade off between observing many small outbreaks, and hence many samples from final size distributions, and few large outbreaks, and hence richer temporal information. These simulated data sets were analysed using DA-MCMC under the models that they were generated from. We found that many repetitions of small outbreaks lead to more precise estimates than fewer outbreaks from larger populations, that is, the most precise estimates were made from surveilling many small subpopulations. We also tested how robust estimates were to a misspecified model. For the models considered, the least biased results came from outbreaks in larger populations (where at least one subpopulation experienced a large outbreak). To balance the need for parameters to have both low variance and low bias under model misspecification we found that surveilling a mix of many small subpopulations and a few large populations were preferable; this allowed for a balance of many final size samples and some longer sets of temporal data. However, we found that there is no clear way of surveilling subpopulations in a way that the reproduction number and growth rate is guaranteed to have low bias under a misspecified model.

The analysis in Chapter 4 has some limitations due to simplifying assumptions. We considered that subpopulations with a combined total of six hundred individuals would be surveilled without accounting for the number of infections in each subpopulation. This implies that the same resources were spent on surveilling a subpopulation with one infectious case as opposed to a subpopulation with many infectious cases. In reality, surveilling a subpopulation which has no secondary infection is unlikely to be as resource intensive as surveilling a subpopulation with a large outbreak. Resource allocation likely acts as a tradeoff between the number of individuals which are surveilled and the number of infections observed, however, this relationship is unknown. Although we did not assume that all cases in the population were observed (as assumed in Chapter 3), we assumed that if any case in a subpopulation is observed then all cases within the subpopulation are observed. This limits the results of the analysis to diseases with easily identifiable symptoms. We also only considered frequency-dependent transmission, which could be reasonable depending on the disease considered [69], but in reality infectious diseases may have a transmission rate that is somewhere between frequency-dependent and density-dependent [67,68]. The analysis considered transmissibility within subpopulations, without considering the transmissibility between subpopulations; so this should be considered in the context of characterising outbreaks within schools, households and workplaces rather than at a population level.

As choosing an inappropriate model lead to biased estimates in Chapter 4, we considered Bayesian model selection methods in Chapter 5. We developed a Bayesian model selection method that used an efficient particle filter for calculating point estimates of the likelihood function within an importance sampling scheme. The particle filter simulates epidemic trajectories in a way that sample paths are always feasible and always match observations. The method takes advantage of a stopping criteria to ensure the accuracy of model evidence estimates and hence ensures that the most appropriate model is chosen for a given data set. We employed this method to identify two important features for evaluating how to control an outbreak: the shape of the infectious period distribution; and, the timing of infectiousness relative to symptom onset. In both cases we found that FF100 study data were informative enough to identify these features (even with loose priors) and that our method was effective at computing the model evidence estimates. In particular, in almost all cases, by the time one-hundred and fifty households were infected we could identify the time of infectiousness relative to symptom onset.

The method described in Chapter 5 was implemented without being optimised. We note that estimates of the coefficient of variation of model evidence estimates could be used as a heuristic to decide the number of state particles. As the method was implemented by independently sampling from the prior distribution it may perform poorly in high dimensional parameter spaces where very few samples have posterior support and many samples may be needed to infer the model evidence accurately. Performance could be improved by choosing a more sensible sampling distribution, for example, samples from PM-MCMC can be used to inform an efficient sampling distribution [23]. Chapter 5 utilised a stopping criteria which required model evidence estimates to have non-overlapping credible intervals. When model evidences were close this made convergence slow. A solution is to sample until model evidence estimates are within a specified tolerance. This implementation is practically useful as considering all models with posterior support and quantifying transmissibility in each case allows for a more complete characterisation of the epidemic process as opposed to only considering one "best" model.

Samples from the posterior distribution can be used to improve the model selection algorithm from Chapter 5 and samples the model selection algorithm can be used to improve parameter estimates. Chapter 6 combines these ideas to develop a novel version of SMC-squared, dubbed JIMS algorithm, which uses the model selection algorithm from Chapter 5 to rejuvenate particles when particle degeneracy occurs. Samples in the model selection step are made by fitting Gaussian mixture distributions to the set of particles. JIMS algorithm and a standard SMC-squared were applied to data on daily infectious cases from homogeneous SIR and SE(2)I(2)R models. For each of these models, JIMS algorithm outperformed standard SMC-squared; it resulted in lower variance estimates of model evidence and parameter means, and, had a shorter run time. JIMS algorithm was also applied to the difficult problem of choosing the time of symptom onset relative to infectiousness, given multiple outbreaks in large subpopulations (of size one hundred). In this case we supposed symptoms corresponded to a transition into the exposed phase 2, infectious phase 1 or infectious phase 2 class;

these transitions are more similar to those considered in Chapter 5, which increases the difficulty of the model selection problem further. For each model we were able to effectively perform model selection and inference simultaneously. The highest posterior model probability corresponded to the true model in at least 19 out of 30 simulated data sets for each model. Further, mean estimates of transmissibility were unbiased for data sets generated under each model.

Chapter 6 gave an implementation of JIMS algorithm but highlighted many alternative implementations. There were multiple ways that stopping criteria could be implemented, importance sampling distributions could be chosen and ways in which algorithmic parameters (such as the number of state particles) could be adapted throughout the algorithm. This highlights that there are many aspects of JIMS algorithms which are yet to be optimised. The comparison of JIMS algorithm and SMC-squared considered an implementation of SMC-squared with a fixed number of samples of PM-MCMC per particle in the rejuvenation step. It has been suggested that the number of samples in the rejuvenation step of SMC-squared could be adapted based on an estimate of the acceptance probability of PM-MCMC [25]. An implementation of SMC-squared with an adapted number of samples in the rejuvenation step is still unlikely to outperform JIMS algorithm, as JIMS algorithm outperformed SMC-squared in terms of both accuracy and run time. The model selection problem in Chapter 4 considered the problem of surveilling a fixed number of individuals in subpopulations. In Chapter 6 we considered multiple outbreaks data again, however, we supposed that there were resources to surveil subpopulations until three hundred infectious individuals are observed. Realistically, the utilisation of resources will be somewhere between these two models; as individuals without symptoms will be surveilled, but surveilling a population with few secondary infections will be less resource intensive than surveilling a large outbreak. Resources allocation will depend on both the total number of individuals in surveilled populations and the number of cases in each of these populations. While JIMS algorithm outperforms SMC-squared in the cases considered, we note that independently sampling from the parameter space may perform worse than proposing small moves to a set of particles in high dimensional parameter spaces. This indicates that, for models with many parameters, SMC-squared with a PM-MCMC rejuvenation step that does not use an independent sampler may outperform JIMS algorithm.

Future Considerations

The models in this thesis did not consider asymptomatic infection, or otherwise unobserved cases of infection. Methods based on household models in Chapter 3 assumed that all infectious cases in the population were observed. The methods applied to multiple outbreaks data assumed that all cases in each of the subpopulations were observed. Dealing with unobserved infections poses a difficult inference problem; it increases the dimension of the parameter space (as there may be an observation probability parameter), the initial time of infection may be unknown, the initial state of the process is unknown and there are more latent variables. We briefly looked into implementing DA-MCMC for a homogeneous SIR model where there was a fixed probability of observing infectious cases, the dimension of the latent variables and the lack of parameter identifiability made the mixing slow. However, parameter identifiability for these models can improved by considering multiple outbreak data, as shown in [73]. Currently, a collaborator is aiming to apply the model selection method from Chapter 5 to models of influenza and SARS with unobserved cases.

In Chapter 3 we only gave a brief outline of the sequential DA-MCMC method due to its limitations. The ability to sequentially update posterior distributions in a data-augmented framework is appealing, and this method provides a solution. The largest computational limitation of the method is its reliance on kernel density estimates. Methods that avoid the need for kernel density estimation could be a useful alternative, for example, if the posterior distributions can be well approximated by a Gaussian mixture this could be used instead (though this may introduce some error).

The analysis in Chapter 4 motivated the need to perform model selection. With methods developed in Chapters 5 and 6 we intend to return to the optimal surveillance problem where the model is chosen and parameters are inferred. Further, we aim to perform a sensitivity analysis to determine the sensitivity of estimates with respect to the shape of the exposed and infectious period distributions. As mentioned earlier, we assumed that there were resources to surveil a fixed number of individuals, however, in reality surveilling many people in one subpopulation may be easier than surveilling many people in many subpopulations. An analysis of the resources available for surveillance and how efficiently subpopulations of different sizes can be surveilled would allow for a more realistic model of FF100 style data under different surveillance protocols.

In Chapter 5 we successfully identified the infectious period shape and the time of symptom onset relative to infectiousness separately. Ideally we would seek to apply to model selection algorithm to choose between these two features simultaneously, and, perhaps identify the shape of the exposed period too. Of course, if there are many models to select between, it becomes difficult to implement a stopping rule which requires non-overlapping credible intervals of model evidence.

In Chapter 6 we discuss the ability to increase the number of state particles and to use effective sample size in a stopping rule for the rejuvenation step for JIMS algorithm. Neither of these approaches were applied, so implementing these and comparing run times, the number of iterations of rejuvenation and the precision of estimates is of interest. JIMS algorithm was implemented with a sampling distribution that was a defence mixture weighted according to a chosen distribution (we chose a truncated geometric distribution), comparing different choices of distributions for the most efficient and reliable implementation would be useful. This method could also be used to prepare for pandemic influenza by performing model selection on historical influenza data for a variety of models to obtain an informed prior over the model space, rather than using a uniform prior on the candidate models.

To conclude, in this thesis I have developed Bayesian methods for inferring parameters and selecting between competing models for epidemics in populations of households based on FF100 study data, and, identified a number of areas of further research. This thesis shows that FF100 study data are informative enough to infer measures of transmissibility for complex models. It analyses how to surveil emerging infectious diseases and highlight the importance of model selection. Further, it provides methods for solving difficult model selection problems and shows that FF100 study data are informative enough to identify important features of epidemics, which has real implications for the controllability of emerging infectious diseases. All methods in this thesis were assessed via simulating FF100 study data, due to a lack of real FF100 study data. In Australia tests of FF100 protocol are likely to take place during a seasonal influenza outbreak in the next few years; this thesis provides methods that can be readily applied during these studies.

Appendix A

DA-MCMC for multiple outbreaks

A subpopulation of size N is modelled as a SI(n)R or a SE(n_1)I(n_2)R CTMC. These models are initialised by seeding an infection in the subpopulation at a Uniform$(0, 1)$ distributed time, t_1, such that $S_{t_1} = N - 1$ and $I^1_{t_1} = 1$. Our data only reveals time of symptom onset at a daily resolution, and we make the modelling assumption that the onset of symptoms corresponds to an infection transition. We assume that our data reveals all infections that will occur in each subpopulation over the entire course of the epidemic, that is, we assume no more infections will occur in surveilled subpopulations. To use data from these subpopulations to infer the transmission rate in a subpopulation, β, the mean latent period $1/\sigma$ and the mean infectious period $1/\gamma$, we assume that each subpopulation acts independently after their initial infection.

Due to independence of outbreaks in each subpopulation the likelihood function is the product of the likelihood functions of each individual outbreak. Let $x_{[t_1,T]}$ be the complete epidemic process, from the time of initial infection, t_1, until the last recovery at time T. Let $y_{1:T}$ be the data vector, with each entry corresponding to the number of infections observed each day for a single subpopulation. Here we present the augmented likelihood for an SE(n_1)I(n_2)R epidemic, but note that it is simple to modify this to an SI(n)R model by removing terms related to the exposed period. Assume that $x_{[t_1,T]}$ agrees with $y_{1:T}$, meaning that $x_{[t_1,T]}$ describes a feasible SE(n_1)I(n_2)R process in which the number of infections each day equals the entries of $y_{1:T}$. The augmented likelihood function for an outbreak in a subpopulation is given by

$$f(x_{[t_1,T]}, y_{1:T}|\beta,\sigma,\gamma,t_1) = \prod_{i\in A} \frac{\beta I_{t_{i-1}}}{N-1} \exp\left\{-\int_{t_1}^{T} \frac{\beta S_t I_t}{N-1} dt\right\} \times \prod_{k=2}^{\bar{Y}} g(\Delta_k, n_1+1, n_1\sigma) \prod_{l=1}^{\bar{Y}} g(\delta_l, n_2+1, n_2\gamma),$$

where $g(a,b,c)$ represents a gamma distribution pdf evaluated at a with shape parameter b and scale parameter c; I_t is the total number of infectious individuals in the population at time t; A represents indices of transitions corresponding to infections (excluding the first infection); $\bar{Y}$ is the total number of cases; Δ_k is the latent period of the kth infected individual; and, δ_l is the infectious period of the lth infected individual.

Let θ be the set of model parameters; in the case of the SE(n_1)I(n_2)R model this is $\theta = \{\beta,\sigma,\gamma\}$. Due to independence of each subpopulation we can write the full augmented likelihood function as

$$f\left(x^{1:M}, y^{1:M}|\theta, t_1^{1:M}\right) = \prod_{j=1}^{M} f\left(x^j, y^j|\theta, t_1^j\right),$$

where y^j, x^j and t_1^j are shorthand for $y_{1:T}$, $x_{[t_1,T]}$ and t_1 in subpopulation j, respectively. The objective of the inference is to calculate the posterior distribution, $f\left(\theta|y^{1:M}\right)$, which we calculate by sampling from the augmented posterior,

$$f\left(\theta, t_1^{1:M}, x^{1:M}|y^{1:M}\right) \propto f\left(x^{1:M}, y^{1:M}|\theta, t_1^{1:M}\right) p\left(\theta, t_1^{1:M}\right),$$

and integrating over $t_1^{1:M}, x^{1:M}$. Here $p\left(\theta, t_1^{1:M}\right)$ is the joint prior distribution of θ and $t_1^{1:M}$. We choose priors such that $\beta/\gamma \sim \text{Uniform}(0.25, 4)$, $1/\gamma \sim \text{Uniform}(0.25, 7)$, $1/\sigma \sim \text{Uniform}(0.25, 7)$ and by our modelling assumptions $t_1^{1:M}$ have an i.i.d Uniform$(0,1)$ prior distribution. Let $\bar{Y}^j$ denote the total number of cases in subpopulation j. By taking the augmented posterior distribution up to proportionality with respect to each variable we obtain marginal distributions for the Gibbs sampler,

$$\beta|\sigma,\gamma,t_1^{1:M},x^{1:M},y^{1:M} \sim \text{Gamma}\left(1+\sum_{m=1}^{M}(\bar{Y}^m-1), \int_{t_1}^{T}\frac{S_t I_t}{N-1}dt\right),$$

$$\sigma|\beta,\gamma,t_1^{1:M},x^{1:M},y^{1:M} \sim \text{Gamma}\left(-1+n_1\sum_{m=1}^{M}(\bar{Y}^m-1), n_1\sum_{m=1}^{M}\sum_{k=2}^{\bar{Y}^m}\Delta_k^m\right),$$

and

$$\gamma|\beta,\sigma,t_1^{1:M},x^{1:M},y^{1:M} \sim \text{Gamma}\left(-2+n_2\sum_{m=1}^{M}\bar{Y}^m, n_2\sum_{m=1}^{M}\sum_{l=1}^{\bar{Y}^m}\delta_l^m\right).$$

Lastly, the distribution of $x^{1:M}, t_1^{1:M}|\theta, y^{(1:M)}$ can be sampled through Hastings steps. Note that these Hastings steps can be made independently for each outbreak. In this case Hastings steps can be made by uniformly randomly selecting a subpopulation and choosing one of the following moves with probability p and $1-p$ respectively:

(i) uniformly randomly select an individual (other than the first infected individual) who became infectious at time t. Sample a Uniform$(\lfloor t \rfloor, \lceil t \rceil)$ candidate infection time, $\hat{t}$, a $\hat{t} -$ Gamma$(n_1, n_1\sigma)_{[0,\hat{t}-t_1]}$ exposed period and a $\hat{t} +$ Gamma$(n_2, n_2\gamma)_{[0,T-\hat{t}\ ]}$ infectious period, to obtain candidate sample path $\hat{x}^j$. Accept this path with probability

$$\min\left\{1, \frac{f(\hat{x}^j, y^j|\theta)}{f(x^j, y^j|\theta)}\right\};$$

(ii) sample a candidate infection time for the first infected individual $\hat{t}_1 \sim$ Uniform$(0, t_2)$, and a $\hat{t}+$Gamma$(n_2, n_2\gamma)_{[0,T-\hat{t}\ ]}$ distributed recovery time. Accept this path with probability

$$\min\left\{1, \frac{f(\hat{x}^j, y^j|\theta)}{f(x^j, y^j|\theta)}\right\}.$$

To infer the shape of the exposed and infectious period via DA-MCMC too, consider the marginal distributions for the shape parameters. These are

$$f(n_1|\beta,\gamma,\sigma,n_2,x^{1:M},y^{1:M},t^{1:M}) \underset{n_1}{\propto} \frac{(\prod_{m=1}^{M}\prod_{p=2}^{\bar{Y}^m}\sigma n_1 \Delta_p^m)^n p(n_1)}{(n_1!)^{\sum_{m=1}^{M}\bar{Y}^m - M}},$$

and

$$f(n_2|\beta,\gamma,\sigma,n_1,x^{1:M},y^{1:M},t^{1:M}) \underset{n_2}{\propto} \frac{(\prod_{m=1}^{M}\prod_{p=1}^{\bar{Y}^m}\gamma n_2 \delta_p^m)^{n_2} p(n_2)}{(n_2!)^{\sum_{m=1}^{M}\bar{Y}^m}};$$

where $p(n_1)$ and $p(n_2)$ are the prior distributions of the shape parameters. These can be used to infer the shape parameters by normalising and taking Gibbs samples in the n_1 and n_2 dimensions. For data on only final epidemic size we can perform DA-MCMC inference on an SI(n)R model by holding $\gamma = 1$ fixed and estimating $\beta = R_0$.

Bibliography

[1] World Health Organization. WHO guidance for surveillance during an influenza pandemic, 2017.

[2] Health Protection Agency, Health Protection Scotland, Communicable Disease Surveillance Centre Northern Ireland, and National Public Health Service for Wales. The First Few Hundred (FF100) project: epidemiological protocols for comprehensive assessment of early swine influenza cases in the United Kingdom, 2009.

[3] Australian Department of Health and Ageing. Australian health management plan for pandemic influenza, 2014.

[4] E. McLean et al. Pandemic (H1N1) 2009 influenza in the UK: clinical and epidemiological findings from the first few hundred (FF100) cases. *Epidemiol. Infect.*, 138:1531–41, 2010.

[5] A. B. van Gageldonk-Lafeber, van der Sande M. A., A. Meijer, I. H. Friesema, G. A. Donker, J. Reimerink, et al. Utility of the first few100 approach during the 2009 influenza A(H1N1) pandemic in the Netherlands. *Antimicrob. Resist. Infect. Control*, 1:30, 2012.

[6] James N. Walker, Joshua V. Ross, and Andrew J. Black. Inference of epidemiological parameters from household stratified data. *PLOS ONE*, 12:1–21, 2017.

[7] T. House, N. Inglis, J. V. Ross, F. Wilson, S. Suleman, O. Edeghere, et al. Estimation of outbreak severity and transmissibility: influenza A(H1N1)pdm09 in households. *BMC Medicine*, 117:1–7, 2012.

[8] J. M. McCaw, K. Glass, G. Mercer, and J. McVernon. Pandemic controllability: a concept to guide a proportionate and flexible operational response to future influenza pandemics. *Journal of Public Health*, 36:5–12, 2013.

[9] Andrew J. Black, Nicholas Geard, James M. McCaw, Jodie McVernon, and Joshua V. Ross. Characterising pandemic severity and transmissibility from data collected during First Few Hundred studies. *Epidemics*, 2017.

[10] N. T. J. Bailey. *The mathematical theory of infectious diseases and its applications. 2nd edition.* Charles Griffin and company limited, 1975.

[11] RM Anderson and RM May. *Infectious diseases of humans.* Oxford University Press, 1991.

[12] Lorenzo Pellis, Neil M. Ferguson, and Christophe Fraser. Epidemic growth rate and household reproduction number in communities of households, schools and workplaces. *Journal of Mathematical Biology*, 63:691–734, 2011.

[13] J Wallinga and M Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society of London B: Biological Sciences*, 274:599–604, 2007.

[14] Alun L. Lloyd. Realistic distributions of infectious periods in epidemic models: Changing patterns of persistence and dynamics. *Theoretical Population Biology*, 60:59 – 71, 2001.

[15] Helen J Wearing, Pejman Rohani, and Matt J Keeling. Appropriate models for the management of infectious diseases. *PLOS Medicine*, 2, 2005.

[16] Caroline Fraser, Steven Riley, Roy M. Anderson, and Neil M. Ferguson. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences of the United States of America*, 101 16:6146–51, 2004.

[17] P. D. O'Neill and G. O. Roberts. Bayesian inference for partially observed stochastic epidemics. *J. R. Stat. Soc. A*, 162:121–130, 1999.

[18] S. Cauchemez and N. M Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *J. R. Soc. Interface*, 5:885–897, 2008.

[19] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 20:801–836, 2003.

[20] M J Keeling and J V Ross. On methods for studying stochastic disease dynamics. *J R Soc Interface*, 5:171–181, 2008.

[21] Trevelyan McKinley, Alex R Cook, and Robert Deardon. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5, 2009.

[22] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12:656–704, 2009.

[23] Naif Alzahrani, Peter Neal, Simon E.F. Spencer, Trevelyan J. McKinley, and Panayiota Touloupou. Model selection for time series of count data. *Computational Statistics & Data Analysis*, 122:33 – 44, 2018.

[24] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:397–426, 2012.

[25] Christopher C. Drovandi and Roy A. McCutchan. Alive SMC2: Bayesian model selection for lowcount time series models with intractable likelihoods. *Biometrics*, 72:344–353, 2016.

[26] Andrew Golightly and Theodore Kypraios. Efficient SMC^2 schemes for stochastic kinetic models. *Statistics and Computing*, 28:1215–1230, 2018.

[27] Philip D. O'Neill, David J. Balding, Niels G. Becker, Mervi Eerola, and Denis Mollison. Analyses of infectious disease data from household outbreaks by markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 49:517–542, 2000.

[28] N. Demiris and P. D. O'Neill. Bayesian inference for epidemics with two levels of mixing. *Scand. J. Statist.*, 32:265–280, 2005.

[29] Peter Neal. Efficient likelihood-free bayesian computation for household epidemics. *Statistics and Computing*, 22:1239–1256, 2012.

[30] Chris P Jewell, Theodore Kypraios, Peter Neal, Gareth O Roberts, et al. Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4:465–496, 2009.

[31] Nikolaos Demiris, Theodore Kypraios, and L. Vanessa Smith. On the epidemic of financial crises. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177:697–723, 2013.

[32] Daniel Merl, Leah R. Johnson, Robert B. Gramacy, and Marc Mangel. A statistical framework for the adaptive management of epidemiological interventions. *PLOS ONE*, 4:1–9, 2009.

[33] Pieter Trapman, Frank Ball, Jean-Stéphane Dhersin, Viet Chi Tran, Jacco Wallinga, and Tom Britton. Inferring R_0 in emerging epidemics–the effect of common population structure is small. *Journal of the Royal Society Interface*, 13:20160288, 2016.

[34] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[35] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[36] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

[37] Panayiota Touloupou, Naif Alzahrani, Peter Neal, Simon E. F. Spencer, and Trevelyan J. McKinley. Efficient model comparison techniques for models requiring large scale data augmentation. *Bayesian Anal.*, 13:437–459, 2018.

[38] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[39] James N. Walker, Andrew J. Black, and Joshua V. Ross. Bayesian model discrimination for partially-observed epidemic models. *Mathematical Biosciences*, 317:108266, 2019.

[40] Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

[41] Roger B. Sidje. Expokit: A software package for computing matrix exponentials. *ACM Trans. Math. Softw.*, 24:130–156, 1998.

[42] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[43] Frank Ball, Denis Mollison, and Gianpaolo Scalia-Tomba. Epidemics with two levels of mixing. *Ann. App. Prob.*, 7:46–89, 1997.

[44] L. Pellis, F. Ball, and P. Trapman. Reproduction numbers for epidemic models with households and other social structures I: Definition and calculation of R0. *Math. Biosci.*, 235:85–97, 2009.

[45] F. Ball, L. Pellis, and P. Trapman. Reproduction numbers for epidemic models with households and other social structures II: comparisons and implications for vaccination. *Math. Biosci.*, 2016.

[46] E. Goldstein, K. Paur, C. Fraser, E. Kenah, J. Wallinga, and M. Lipsitch. Reproductive numbers, epidemic spread and control in a community of households. *Math. Biosci.*, 221:11–25, 2009.

[47] F. Ball. Stochastic and deterministic models for SIS epidemics among a population partitioned into households. *Math. Biosci.*, 156:41–67, 1999.

[48] P. K. Pollett and V. T. Stefanov. Path integrals for continuous-time Markov chains. *Journal of Applied Probability*, 39:901–904, 2002.

[49] Joshua V. Ross, Thomas House, and Matt J. Keeling. Calculation of disease dynamics in a population of households. *PLOS ONE*, 5:1–9, 2010.

[50] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[51] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman and Hall/CRC, 1995.

[52] S. Brooks, A. Gelman, G. L. Jones, and X. Meng. Handbook of Markov Chain Monte Carlo. *Biometrics*, 69:801–801, 2013.

[53] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *Ann. Stat.*, 38:2916–2957, 2010.

[54] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Non-centered parameterisations for hierarchical models and data augmentation. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, volume 307. Oxford University Press, USA, 2003.

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

[56] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

[57] Michael K. Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171:134 – 151, 2012.

[58] Genshiro Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25, 1996.

[59] Mark Briers, Arnaud Doucet, and Simon Maskell. Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, 62:61, 2010.

[60] Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37:697–725, 2009.

[61] J. N. Walker. Inference methods for first few hundred studies. Master's thesis, School of Mathematical Sciences, University of Adelaide, 2015.

[62] L. Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, 1986.

[63] Australian Bureau of Statistics. Household and family projections, Australia, 2016 to 2041, 2016.

[64] D. Vats, J. M. Flegal, and G. L. Jones. Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106:321–337, 2019.

[65] C. M. Pooley, S. C. Bishop, and G. Marion. Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes. *J. R. Soc. Interface*, 12:20150225, 2015.

[66] Andrew J. Black, Thomas House, M. J. Keeling, and J. V. Ross. Epidemiological consequences of household-based antiviral prophylaxis for pandemic influenza. *Journal of The Royal Society Interface*, 10:20121019, 2013.

[67] Alessia Melegaro, NJ Gay, and GF Medley. Estimating the transmission parameters of pneumococcal carriage in households. *Epidemiology & Infection*, 132:433–441, 2004.

[68] T House and M J Keeling. Household structure and infectious disease transmission. *Epidemiol Infect*, 137:654–661, 2009.

[69] Michiel van Boven, Tjibbe Donker, Mariken van der Lubben, Rianne B. van Gageldonk-Lafeber, Dennis E. te Beest, Marion Koopmans, Adam Meijer, Aura Timen, Corien Swaan, Anton Dalhuijsen, Susan Hahné, Anneke van den Hoek, Peter Teunis, Marianne A. B. van der Sande, and Jacco Wallinga. Transmission of novel influenza a(H1N1) in households with post-exposure antiviral prophylaxis. *PLOS ONE*, 5:1–10, 2010.

[70] Michiel van Boven, Marion Koopmans, Mirna Du Ry van Beest Holle, Adam Meijer, Don Klinkenberg, Christl A Donnelly, and Hans (J. A. P.) Heesterbeek. Detecting emerging transmissibility of avian influenza virus in human households. *PLOS Computational Biology*, 3:1–9, 2007.

[71] Matt J. Keeling and J.V. Ross. Optimal prophylactic vaccination in segregated populations: When can we improve on the equalising strategy? *Epidemics*, 11:7 – 13, 2015.

[72] A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:501–514, 1994.

[73] Andrew J. Black, Nicholas Geard, James M. McCaw, Jodie McVernon, and Joshua V. Ross. Characterising pandemic severity and transmissibility from data collected during First Few Hundred studies. *Epidemics*, 19:61 – 73, 2017.

[74] Andrew J. Black. Importance sampling for partially observed temporal epidemic models. *Statistics and Computing*, 29:617–630, 2019.

[75] Janet S. Sinsheimer, Marc A. Suchard, and Robert E. Weiss. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*, 18:1001–1013, 2001.

[76] Ephraim M Hanks, Mevin B Hooten, Mat W Alldredge, et al. Continuous-time discrete-space models for animal movement. *The Annals of Applied Statistics*, 9:145–165, 2015.

[77] Youyi Fong, Peter Guttorp, and Janis Abkowitz. Bayesian inference and model choice in a hidden stochastic two-compartment model of hematopoietic stem cell fate decisions. *Ann. Appl. Stat.*, 3:1695–1709, 2009.

[78] DP Kroese, T Taimre, and ZI Botev. *Handbook of Monte Carlo Methods*. Wiley, 2011.

[79] Mark A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160, 2003.

[80] Naif Alzahrani, Peter Neal, Simon E.F. Spencer, Trevelyan J. McKinley, and Panayiota Touloupou. Model selection for time series of count data. *Computational Statistics & Data Analysis*, 122:33 – 44, 2018.

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

[82] Jun S Liu, Rong Chen, and Tanya Logvinenko. A theoretical framework for sequential importance sampling with resampling. In *Sequential Monte Carlo methods in practice*, pages 225–246. Springer, 2001.

[83] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[84] T. J. McKinley, J. V. Ross, R. Deardon, and A. R. Cook. Simulation-based Bayesian inference for epidemic models. *Comput. Stat. Data Anal.*, 71:434–447, 2014.

[85] M J Keeling and P Rohani. *Modeling Infectious Diseases in Humans and Animals.* Princeton University Press, New Jersey, 2007.

[86] Andrew J. Black, Alan J. McKane, Ana Nunes, and Andrea Parisi. Stochastic fluctuations in the susceptible-infective-recovered model with distributed infectious periods. *Phys. Rev. E*, 80:021922, 2009.

[87] Andrew J Black and Joshua V Ross. Computation of epidemic final size distributions. *Journal of Theoretical Biology*, 367:159–165, 2015.

[88] Frank Ball. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. App. Prob.*, 18:289–310, 1986.

[89] Randal Douc and Olivier Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69. IEEE, 2005.

[90] Lincoln LH Lau, Benjamin J Cowling, Vicky J Fang, Kwok-Hung Chan, Eric HY Lau, Marc Lipsitch, Calvin KY Cheng, Peter M Houck, Timothy M Uyeki, JS Malik Peiris, et al. Viral shedding and clinical illness in naturally acquired influenza virus infections. *The Journal of Infectious Diseases*, 201:1509–1516, 2010.

[91] Roy M Anderson, Christophe Fraser, Azra C Ghani, Christl A Donnelly, Steven Riley, Neil M Ferguson, Gabriel M Leung, Tai H Lam, and Anthony J Hedley. Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359:1091–1105, 2004.

[92] Thomas House, Marc Baguelin, Albert Jan Van Hoek, Peter J White, Zia Sadique, Ken Eames, Jonathan M Read, Niel Hens, Alessia Melegaro, W John Edmunds, et al. Modelling the impact of local reactive school closures on critical care provision during an influenza pandemic. *Proceedings of the Royal Society B: Biological Sciences*, 278:2753–2760, 2011.

[93] Andrew J. Black and Joshua V. Ross. Estimating a Markovian epidemic model using household serial interval data from the early phase of an epidemic. *PLOS ONE*, 8:1–8, 2013.

[94] Andrew J. Black and Joshua V. Ross. Contact tracing and antiviral prophylaxis in the early stages of a pandemic: the probability of a major outbreak. *Mathematical Medicine and Biology: A Journal of the IMA*, 32:331–343, 2014.

[95] Donald B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12:1151–1172, 1984.

[96] Pierre Del Moral, Ajay Jasra, Anthony Lee, Christopher Yau, and Xiaole Zhang. The alive particle filter and its use in particle Markov chain Monte Carlo. *Stochastic Analysis and Applications*, 33:943–974, 2015.

[97] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[98] Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

[99] John Skilling. Nested sampling. *AIP Conference Proceedings*, 735:395–405, 2004.

[100] R. Salomone, L.F. South, C.C. Drovandi, and D.P. Kroese. Unbiased and consistent nested sampling via sequential Monte Carlo. *arXiv preprint arXiv:1805.03924*, 2018.

[101] T. J. McKinley, P. Neal, S. E. Spencer, A. Conlan, and L. Tiley. Bayesian model choice for partially observed processes: with application to an experimental transmission study of an infectious disease. *Bayesian Analysis*, (to appear) 2019.

[102] Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

[103] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:583–639, 2002.

[104] C. M. Pooley and G. Marion. Bayesian model evidence as a practical alternative to deviance information criterion. *Royal Society Open Science*, 5:171519, 2018.

[105] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[106] Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37:185–194, 1995.

[107] C. C. Drovandi and A. N. Pettitt. Estimation of parameters for macroparasite population evolution using approximate bayesian computation. *Biometrics*, 67:225–233, 2011.