Bayesian Computational Methods for Stochastic Epidemics

Jessica E. Stockdale

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy

April 6, 2019

Abstract

Mathematical modelling has become a useful and commonly-used tool in the analysis of infectious disease dynamics. Understanding disease spread is of considerable importance for public health planning and the prevention of future outbreaks, and mathematical analysis of disease outbreaks offers insight which may not be so easily obtained through direct biological study.

One key aspect, in mathematical analysis of infectious diseases specifically, is that generally the epidemic process is only partially observed. We might be able to identify the time at which infective individuals become symptomatic or recover, but rarely are we able to observe when infection began, or from whom it was transmitted. This leads to a number of complications with analysis, which will be a focus of this work.

The first part of this thesis describes a full Bayesian analysis for such an outbreak with only partial observation of the disease process. We will perform the first Bayesian analysis of the Abakaliki smallpox data, which have been widely cited within the infectious disease modelling literature, to include the full data. In order to do this, we use data augmented Markov Chain Monte Carlo (DA-MCMC) techniques to perform parameter estimation. Analysis involves interpretation of these parameter estimates as well as model assessment with simulation-based methods. We also compare our results to a previous analysis which used an approximate likelihood expression.

The second part of this thesis describes novel approximate likelihood methods, motivated in part by the results of the Abakaliki study. Although DA-MCMC

is generally considered the standard tool for analysis of partial epidemic data, it often struggles for large population sizes and large amounts of missing data, both through issues of highly correlated missing data and of potentially limiting computation times. We suggest that likelihood approximation methods are a useful tool for dealing with these issues. We develop a series of such methods, which essentially assume some independence in the outbreak population in order to obtain likelihood expressions which do not depend on any missing data. These methods will be motivated and developed, and then illustrated both by simulation study and by application to real data.

List of Relevant Publications

- The work contained in Chapter 2 is published in Stockdale et al. (2017).
- The work contained in Chapters 3 and 4 is currently being prepared for submission to a peer-reviewed journal.

Acknowledgements

This PhD project was carried out under the supervision of Phil O'Neill and Theo Kypraios, whom I would like to thank for all of their help. This work was supported by the UK Engineering and Physical Sciences Research Council, grant number EP/M506588/1.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In this thesis, we will explore different aspects of Bayesian analysis for infectious disease data. Bayesian methods are of particular use for epidemic modelling since data are typically only partially observed. Data Augmented Markov Chain Monte Carlo (DA-MCMC) is currently the standard computational Bayesian method employed, and this will be the focus of the first part of this thesis in an application to a much-cited data set concerning a smallpox outbreak. The second part of this thesis will then introduce novel approximate likelihood methods for Bayesian inference. These are motivated by a number of computational problems of DA-MCMC. A series of such likelihood methods will be described, as well as applied to various simulated and real data sets.

First, this introduction will provide relevant background material. We begin with a discussion of the history of infectious disease data analysis, and the use of stochastic models for this. We define important aspects of the data and models which are used, and describe current computational methods. We particularly consider methods which involve approximation, since this will be the focus of much of this thesis.

1.1 Mathematical Infectious Disease Modelling

1.1.1 Background

Mathematical models have seen increasing popularity as tools for furthering the understanding of infectious disease epidemiology. Often, experimental study on the spread of disease in humans and animals is ethically difficult and resource intensive, but mathematical models offer an alternative in seeking to replicate the underlying factors driving disease dynamics, to allow estimation of parameters which, for example, govern disease transmission. Models aim to describe transmission of disease between hosts by incorporating contact patterns and a realistic representation of the disease itself (e.g. lengths of infectious periods, latent periods and so on). Predictions may then be made about key parameters such as infection rates or vaccine efficacies. In fact, the aims of analysis may largely be categorised into three main areas: furthering understanding of the mechanisms by which diseases spread, predicting future spread, and discovery of the methods which best control this (Daley and Gani, 2001). Development of vaccination strategies, health care interventions, and public health initiatives may all be informed by this mathematical research.

The development of mathematical theories on the spread of infectious diseases can be traced back to at least the ancient Greeks, although real progress was arguably only made from the 19th century with the discovery of the connection of microorganisms to disease. This laid the foundations for the development of more rigorous mathematical descriptions of disease outbreaks. Although there had been previous studies such as the famed Broad Street Pump study of 1854/5, in which a contaminated water pump was identified as the source of a cholera outbreak in London (see e.g. Newsom, 2006), these had been largely empirical rather than theoretical. In the last century, however, the field has seen great advancement, driven both by development of mathematical theory and increasing computational resources. From the development of individual-based, deterministic models where outbreak progression depends

specifically on the numbers of susceptible and infectious individuals, to the introduction of stochastic models which incorporate probabilities of events occurring, development has been rapid. In Kermack and McKendrick (1927), the first complete, deterministic mathematical model to receive attention was introduced, named the *general epidemic model*, with its stochastic counterpart in McKendrick (1926), though this received less attention. Previously, these kinds of models were generally only used within mathematical theory, but since the 1990s or so there has been an increased interest in the field by more applied scientists as well as public health officials and policy makers. Central texts which further describe the history of the field as well as key methods and applications include Bailey (1975), Becker (1989), Anderson and May (1991) and Andersson and Britton (2000). Today, with increasing volumes of data driven by increasing computational power, there is growing demand on mathematical models which capture more complex populations, as well as methods which can translate these into real-world conclusions.

1.1.2 Stochastic Models

The set of models for analysis of infectious disease is often split into deterministic and stochastic. Deterministic models are usually defined through a set of ordinary (or partial) differential equations, which seek to describe the flow of individuals between different disease stages over time. Stochastic models, although often less straightforward to analyse, are generally considered more realistic than their deterministic counterparts. In capturing the variability of real-life events, they represent the natural stochasticity of disease outbreaks.

The use of stochastic models has allowed for considerable development in infectious disease analysis. From the 1930s, when the idea of using binomial distributions to represent successive crops of new cases was introduced (described in Bailey, 1975), there has been an increase in their use. They are especially useful when the number of individuals in the population is small, since in these cases the innate randomness of the processes involved is more

pronounced and hence deterministic models less well describe the epidemic spread. Models are generally constructed on an individual-based level, where we consider individual units (humans, animals, or even groups of individuals) within some population. The total population might refer to a household, a city, or a country, for example.

One model which has received significant attention (although it was never published) was proposed by Reed and Frost. This is a discrete time chain-binomial model, in which outbreaks are described as evolving in generations where each infected individual infects each susceptible individual independently, with a fixed probability p. The individuals infected by those in some generation g then form generation $g + 1$, and those in g are assumed to recover. The number of infectives in each generation is therefore binomially distributed, with probability dependent upon the number of infectives present in the previous generation, and the entire outbreak is described by a chain of such binomial random variables. A more detailed description of the Reed-Frost model can be found in Andersson and Britton (2000), Section 1.2 or Abbey (1952). The so-called *general stochastic model* for an epidemic process was then formulated by Bartlett (1949). This was the first work to define a stochastic model using Markovian processes, and most stochastic models since have been defined as such, in either discrete or continuous time. From this point on, we will restrict our attention to stochastic rather than deterministic models, since these are the focus of this work.

1.1.2.1 SIR and SEIR Stochastic Epidemic Models

Models used for analysis of infectious diseases generally include some kind of state of health of the individuals concerned. These are commonly known as compartmental models, since they categorise individuals into a discrete set of disease states.

The standard terminology is to define all individuals who are currently able to become infected by the disease as '*susceptible*', all currently infectious (i.e. able

to pass the disease on to others) individuals as '*infective*' and all individuals who are no longer infectious but are also not able to contract the disease again as '*removed*'. This removed state might correspond to a number of causes: death, removal from the population, recovery and immunity from reinfection, or perhaps recovery with the timescale of the outbreak too short for reinfection to reasonably occur. These states may be combined to form what is called an SIR (susceptible-infective-removed) model, though there are others which may be considered and are often applied. For example, if we also introduce a '*latent*' or '*exposed*' period before the infectious period, wherein individuals are infected but not yet able to infect others, we may use an SEIR (susceptible-exposed-infective-removed) model. If we do not wish to include a removal period, but instead to allow all recovered individuals to become immediately infected again, we may use an SIS (susceptible-infective-susceptible) model. In this thesis, however, we will largely focus on SIR and SEIR models, and extensions thereof.

Figure 1.1 depicts a typical timeline for the stages of a disease that an individual might pass through. The model used here is an extension of the SEIR model, which we will see more of in Chapter 2, where we have split the infectious period into two parts according to whether the individual is symptomatic or asymptomatic. These periods might be categorised by differing infection rates. Of course, compartmental models may be defined with any compartments desired, but we consider here those which are frequently seen in practice.

We will discuss different types of infectious disease data further in Section 1.2, but typically outbreak data contain the removal times only (or, say, case detection times treated as removal times). The length of each individual's infectious period, and the latent period if we include this, are then assumed random and independent of other individuals. In general, these lengths of time are assumed to be random samples from distributions with known parameters.

Transmission of the disease is modelled as being the result of *infectious con-*

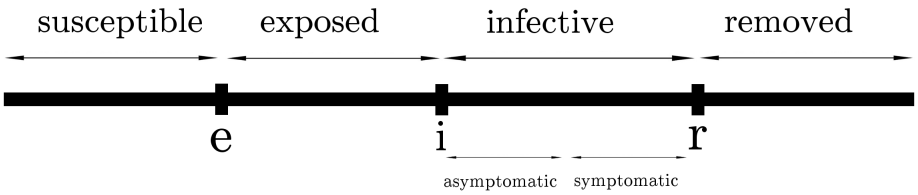


Figure 1.1: Typical disease timeline for an SEIR model with 2 stage infectious period. Times "*e*", "*i*" and "*r*" refer to the start of the individual's exposed (infected, but not yet infectious), infective (able to infect others) and removed (plays no further part in the outbreak) periods, respectively.

tacts. Infectious individuals in the population are assumed to have contacts with others at some defined rate, where a contact is defined as an interaction close enough for an infection to occur. In reality, this contact may refer to a physical meeting of these individuals, but also a proxy for the infective dispersing pathogens in their environment and the contacted individual making contact with this pathogenic material. Examples include using objects they have encountered, breathing in the same air, or, in hospital settings, medical staff being the intermediary link. An infectious contact with a susceptible is normally assumed to result in the susceptible's immediate infection.

We now more rigorously define the SIR model, which may be extended to any compartmental model of choice. We first define a closed population of size N, which does not include any demographic changes (i.e. births or deaths). We assume, for now, that this population is homogeneously mixing (i.e. all individuals mix uniformly), so that the chance that any two individuals meet is independent of the choice of individuals. All individuals will at all times $t \geq 0$ be in one of the three states: susceptible, infective and removed. The total numbers of individuals in these categories at any time t are given by $S(t)$, $I(t)$ and $R(t)$, respectively, where for all t, $S(t) + I(t) + R(t) = N$. We assume external infection of the initial infective, so that initially $S(0) = N - 1$,

$I(0) = 1$ and $R(0) = 0$.

Any given infective individual will contact any other given individual at times given by the points in a homogeneous Poisson process of rate β, where all Poisson processes are mutually independent. Any contact between an infective and a susceptible is assumed to result in immediate infection of the susceptible. The length of any infective's infectious period x is assumed independent of all others, and identically distributed with arbitrarily defined probability density (or mass) function $f_I(x \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ includes the parameters controlling the length of the infectious periods. The outbreak ends when there are no more infectives in the population.

The *general stochastic epidemic model* for infectious periods following an exponential distribution with rate γ (so that $f_I(x \mid \boldsymbol{\theta}) = \gamma e^{-\gamma x}$, $x > 0$) may then be defined as a continuous-time Markov chain $\{(S(t), I(t)) : t \geq 0\}$ with transition probabilities:

$$\mathbb{P}\big[(S(t+h), I(t+h)) = (s-1, i+1) \mid (S(t), I(t)) = (s, i)\big] = \beta hsi + o(h)$$
$$\mathbb{P}\big[(S(t+h), I(t+h)) = (s, i-1) \mid (S(t), I(t)) = (s, i)\big] = \gamma hi + o(h), \text{ as } h \to 0$$

where the first equation corresponds to an infection and the second to a removal. At time t, infections then occur at rate $\beta S(t) I(t)$ and removals at rate $\gamma I(t)$.

Under this model, the infectious periods are independent exponentially distributed lengths of time with mean $\frac{1}{\gamma}$. In this standard form of the model there are hence two parameters; the infection rate β and the removal rate γ. We will also often employ gamma distributed infectious periods within this thesis, with mean $\frac{m}{\gamma}$ for shape parameter m and rate parameter γ. Gamma (or Erlang, for integer valued m) distributions are frequently used since exponentially distributed infectious periods, although leading to convenient mathematical results, may be unrealistic in practice (see e.g. Lloyd, 2001, Streftaris and Gibson, 2004).

1.1.2.2 Population Structures

Although in Section 1.1.2.1 we assumed that all individuals mixed homogeneously, in reality we often assume some heterogeneity in the mixing behaviour of individuals. The structure for the populations within which outbreaks occur can have considerable impact on the behaviour of the model, and more complex choices of population structure are becoming more commonly used in practice to meet the demands of real data analyses. Models must be complex enough to accurately represent the dynamics of the real population, but simple enough to allow consideration of the impact of modelling assumptions on the outcomes, as well as not being over-parameterised.

There are numerous options for population structures which have been explored (see e.g. Britton et al., 2015 or Mollison, 1995). Global contact structures (or homogeneously mixing structures) as defined in Section 1.1.2.1 essentially assume no structure at all, and individuals within the population do not differ in their interactions, infectivity or in the average length of time spent infected. Although simple and comparatively easy to implement, this assumption is often unrealistic (especially in larger populations rather than smaller communities e.g. households), and so a more detailed description of the population may be necessary.

For this we require a heterogeneously mixing population model, within which we may define a number of structure subcategories. A multi-type model would include a set of structured subgroups in the population, categorising for example by age, sex, or social grouping. Ball et al. (1997) defined a two level mixing household model, where individuals have local contacts within their household and global contacts between households. This was extended to three levels by Britton et al. (2011), who considered global, household and school/workplace contacts in the context of a measles outbreak.

Network models, on the other hand, deal with a more complex structure of interaction between individuals. Considering each individual in the population

as a node, the contact rate between any two individuals is given by the weight of the edge between them. These have become widely applied, see e.g. Keeling and Eames (2005), Barthélemy et al. (2005) and Newman (2002), as well as Britton and O'Neill (2002) who developed methods for inference. Definition of network models is difficult since any individual-based model may be represented as a network of the individuals (for example, a homogeneously mixing population structure is just a completely connected network where the weight of all edges is equal). For simplicity, we will consider them here as any population structure where members vary individually in their infectivity or mixing behaviour, rather than by some structured grouping as in a multi-type model. Danon et al. (2011) provide a detailed discussion of many forms of network models for infectious diseases. One such example is a spatial model, whereby individuals' infectivity/mixing behaviour in some way depends on their 'distance' from others. This could be geographically (for example the distance between individuals' area of residence, as in Chowell et al., 2007), or otherwise (for example the 'genetic distance' between individuals' DNA samples, as in Worby et al., 2016).

1.1.2.3 Reproduction Numbers

Reproduction numbers play an important role in infectious disease analysis. Usually referred to as R, the exact definition of these varies but generally we define, as in Becker (2015):

$R =$ the average number of infections that a single infective will cause.

The name reproduction number comes from the fact that models for disease transmission may essentially be considered as birth-death processes, where an infection describes a 'birth' and a recovery describes a 'death'. Then R is the mean number of 'offspring' produced by an infective. Although we may obtain an estimate of R for an entire outbreak, of course in reality R is constantly changing with the number of susceptibles left in the population, and also po-

tentially with the reaction of the population to the outbreak. We hence define the more interpretable, and more commonly used, *basic reproduction number* R_0:

$$R_0 = \text{the average number of infections that a single infective will cause, in a large and entirely susceptible population.}$$

In the calculation of R_0 we therefore require that none of the population is immune, by vaccination or otherwise, and that all individuals are able to interact. Although this requirement of complete susceptibility makes R_0 less interpretable in its meaning for any given outbreak where this is not the case, it makes it much more comparable between different data sets and diseases. It offers a general measure of the overall infectivity of a disease.

Other reproduction numbers may also be defined, and we will explore a number of these in Chapter 2. We may, for example, define a reproduction number during a particular section of the infectious period, or within a particular subgroup of the population.

The exact formula for any reproduction number will of course depend on the model used for the infection and removal rates. In general, for stochastic models we may define

$$R_0 = \text{infection rate} \times \text{number of susceptibles} \times \text{mean length of the infectious period},$$

and for the Markovian SIR model defined in Section 1.1.2.1 above,

$$R_0 = \frac{\beta N}{\gamma}.$$

This is since each individual will be infectious for, on average, time $\frac{1}{\gamma}$, and the average number of susceptibles infected per unit time is βN.

A key interpretation of R_0 is in its relation to the epidemic spread as a threshold quantity. In an infinitely large population, it is possible to show with probability one that if $R_0 \leq 1$ then the epidemic will die out (only a finite number

of individuals become infected). Correspondingly, a major outbreak in a large population is possible if and only if $R_0 > 1$, by the threshold limit theorem (Andersson and Britton, 2000, Chapter 4). Estimating the value of R_0 from the model parameter estimates allows us to gain an understanding of how large a future outbreak could be. This is particularly useful in informing vaccination strategies, for example, since we can calculate the proportion of a population that would need to be vaccinated to prevent a widespread outbreak.

For a more detailed discussion of this, and reproduction numbers in general, we direct the reader to Andersson and Britton (2000) or Heesterbeek and Dietz (1996), for example.

1.2 Data

So far in this chapter we have focused on defining models for the spread of infectious disease, but of course in a statistical context we are also concerned with inference about the model parameters. This requires data from disease outbreaks, which we now discuss.

Disease outbreak data are most commonly mathematically collected in a temporal form. These typically contain a time series of outbreak events, usually case detection times or removal times. This is often aggregated e.g. into daily or weekly data. Final size data are also common, consisting just of the initial number of susceptibles and which of these were eventually infected. However, this thesis will not focus on final size data, and from this point all data will be assumed temporal.

There is frequently some extra information available as well: for example age, sex or location of residence of the individuals within the population. These may inform the population structure aspects of the model. Information on the vaccination of individuals may also be present, as will be particularly relevant in Chapter 2. This may be incorporated into the model to, for example, consider a proportion of the population effectively initially in the 'removed'

stage. Estimation of the efficacy of the vaccine given a particular outbreak may then be performed. Daley and Gani (2001), Chapter 7 and Anderson and May (1982) include a more detailed discussion of immunization, and particularly its relation to R_0.

Disease outbreak data are commonly only partially observed. This is since the transmission process itself is generally unobserved, for example most infectives are only identified at the moment they become symptomatic. Further than this, if the system is only observed at discrete intervals (e.g. weekly hospital tests or similar), then there will be a large amount of uncertainty as to exactly when outbreak events occurred.

Missing data may also come in the form of unobserved cases, whether due to asymptomatic individuals, misdiagnosis, or under-reporting of cases. Population sizes may also include some element of uncertainty, either in the total number of people living in some particular area or in the proportion of this which is initially susceptible, due, for example, to prior immunity. We generally assume that the population is closed for the duration of the outbreak, but this is course will not always be accurate (particularly for outbreaks which occur over longer lengths of time).

Performing statistical analysis on only partially observed data often proves complicated, even with simple models. Missing data may lead to likelihood expressions which are analytically intractable. We will explore this further in Section 1.3.2.

1.3 Methods for Analysis

The majority of methods for inference from stochastic models use likelihood expressions. As we have discussed in Section 1.2, many of these likelihoods will be intractable, due to a combination of partially observed data and the model used to describe them.

Tractable likelihoods do arise, usually as a result of simplifying assumptions such as fixed length infectious periods. Inference is then possible using standard techniques such as *maximum likelihood estimation* through numerical optimization of the likelihood function. Intractable likelihoods, on the other hand, may occur when infection times (or indeed any event times, though we state infection times for simplicity) are unobserved. The likelihood can then be considered as the integral over all possible infection times of the likelihood expression when augmented with the infection times. The high dimensional and often complex region of integration, however, is what leads to intractability, often analytically and numerically. There are a number of methods which have been explored for working with these likelihoods.

Initial approaches for intractable likelihoods generally took a frequentist approach. *Martingale methods*, as described in Becker (1989), may be applied to final size and temporal data for both parametric and nonparametric inference. Martingales are random processes evolving over time, defined in part by the martingale property which requires the expected change of a martingale over time to be zero. A Martingale process hence must be unbiased, and will usually arise from a counting process in an epidemic context. These count the occurrence of events happening randomly in time, where at each event the process increases by size one. In the context of infectious disease data, this could count the number of observed infections. From this representation of the disease process, maximum likelihood estimators may be obtained, as well as asymptotic results using properties of the Martingales. However, Martingale methods for incomplete data involve reconstruction of the infection process, which requires various approximations or simplifying assumptions (for example, homogeneous mixing structures). Data augmentation and MCMC offer an alternative in the same spirit, but avoiding the need for simplification (Kypraios, 2007 Section 2.1.8, Becker and Britton, 1999).

Data augmentation techniques, as we will more rigorously define in Section 1.3.4, have become more common when working with partially observed data.

These treat missing data as model parameters. The *Expectation-Maximisation* (EM) algorithm is one such technique, in which we create an artificial 'complete' data set in order to perform maximum likelihood estimation with the EM algorithm (Becker, 1997, Dempster et al., 1977). Essentially, we define some quantity $\mathbf{Z}$ which represents the complete data, i.e. a combination of observed data $\mathbf{X}$ and unobserved event times $\mathbf{Y}$. In an epidemic context, this might involve combining the observed removal times with the unknown infection times. The algorithm then uses iterative steps to move from an initial estimate of the model parameters $\boldsymbol{\theta}^{(0)}$ to new estimate $\boldsymbol{\theta}^{(1)}$ and so on, where each iteration increases the likelihood. An iteration is formed of two stages: E and M. The E-step (expectation) involves calculating the expectation of the log-likelihood function of the complete data, with respect to observed data $\mathbf{X}$ and under current parameter estimates $\boldsymbol{\theta}^{(t)}$. The M-step then determines $\boldsymbol{\theta}^{(t+1)}$ by maximising this expectation, and the two-step process is repeated until convergence is achieved.

As demonstrated in Becker (1997), the EM algorithm has been successfully used in applications including HIV/AIDS data. Here, transmissibility aspects of the disease are often ignored since the disease has a considerable incubation period as well as multiple methods of transmission in practice. Simpler models capturing part of the case generation process then fit well with the EM algorithm, due to their natural partial observation. However, in other epidemiological contexts where transmission models are required, the conditional expectation in the E-step will often be difficult to calculate. This is due to the interactions between infective and susceptible individuals, the number of which of course change with time.

It has become more common to fit models within a Bayesian framework, using for example MCMC methods, particularly combined with data augmentation. We will explore this fully in Section 1.3.3, but first provide some necessary background on Bayesian inference.

1.3.1 Bayesian Inference

Bayesian inference (Bernardo and Smith, 1994, Lee, 2012, Gelman et al., 2013) relies upon Bayes' Theorem to derive parameter estimates from a model given data. It is a widely-used statistical framework, with the benefit that it provides a natural way of combining data with prior beliefs. The approach involves deriving a posterior probability from the combination of a likelihood function, which is the conditional distribution of the data given the parameters, and a prior probability, which represents our beliefs about these parameters before the data is taken into account. This posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{X})$ is therefore the conditional distribution of the unknown parameters $\boldsymbol{\theta}$ given data $\mathbf{X}$. Bayes' Theorem states that:

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \mathbf{X}) &= \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{X} \mid \boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta})\pi(\mathbf{X} \mid \boldsymbol{\theta}) \ d\boldsymbol{\theta}} \\ &\propto \pi(\boldsymbol{\theta})\pi(\mathbf{X} \mid \boldsymbol{\theta}), \end{aligned} \tag{1.3.1}$$

where the denominator in the first expression is a normalising constant, and the integral is a sum if $\boldsymbol{\theta}$ is discrete. This constant is typically analytically intractable, especially in high dimensional problems as are usual in infectious disease analysis.

In order to make inference about $\boldsymbol{\theta}$, we require this posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{X})$ which is formed of the likelihood $\pi(\mathbf{X} \mid \boldsymbol{\theta})$ and prior density $\pi(\boldsymbol{\theta})$. The likelihood expression will depend on the choice of model, and the prior distribution must be chosen in advance depending on our beliefs about the parameters.

Any choice of prior distribution may be used for Bayesian inference, but certain choices have proved most popular. *Conjugate priors*, for example, are selected to be 'conjugate' to the likelihood, meaning the resulting posterior will be of the same family of distributions as the prior. This often leads to easier computation. If we have some existing knowledge about the parameters, we might use an *informative prior* which captures that. However, this prior information about the parameters may not always be available. In these cases we use a *non-informative prior*, which aims to contain as little information about

the parameters as possible. This results in the posterior being almost entirely informed by the data. Choice of prior can greatly impact the analysis, and so must be selected carefully.

The posterior distribution obtained in Bayesian inference then contains all the information from the data, as well as our prior beliefs about the parameters. Inference may be performed from the posterior, in order to obtain estimates of the model parameters.

1.3.2 Inference with Missing Data

As we have discussed, infectious disease data are frequently only partially observed. Event times are often unknown, and additionally there may be unreported cases. There is hence considerable importance in methods for inference which can handle missing data.

If we define the missing data as $\mathbf{Y}$, then the pair $(\mathbf{X}, \mathbf{Y})$ form what is known as the *augmented data*, which under any model will have a specified distribution dependent upon parameters $\boldsymbol{\theta}$. In our Bayesian framework using Equation (1.3.1), the conditional distribution of the parameters given the observed data is given, up to proportionality, by:

$$\pi(\boldsymbol{\theta} \mid \mathbf{X}) \propto \pi(\boldsymbol{\theta}) \int_{\mathbf{Y}} \pi(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta}) \; d\mathbf{Y}. \tag{1.3.2}$$

This essentially applies what was discussed in Section 1.3, in that we integrate over all possible values of missing data to obtain the posterior distribution. This integral, however, is usually not analytically or even numerically feasible, particularly if the missing data is of high dimension. Sampling from this therefore usually requires other techniques, such as data augmented MCMC (see Section 1.3.4). First, however, we describe standard MCMC methods.

1.3.3 Markov Chain Monte Carlo Methods

The last three decades have seen an increasing use of Monte Carlo methods within infectious disease modelling, particularly Markov Chain Monte Carlo (MCMC) algorithms. The use of more realistic stochastic models, as we have discussed in Section 1.1.2, has led to challenges in inference from highly dimensional and analytically intractable expressions. MCMC methods have proved a useful tool for dealing with these problems.

First introduced for use in particle physics by Metropolis et al. (1953) but not utilised for Bayesian inference until Gelfand and Smith (1990), MCMC for use with epidemic models was first introduced by O'Neill and Roberts (1999) and Gibson and Renshaw (1998). Since, it has become arguably the standard method for analysis. The literature on MCMC applied to epidemic models is too numerous to list in full, but includes Demiris and O'Neill (2005), who applied MCMC to a model with two levels of mixing, O'Neill and Becker (2001), who first applied MCMC for non-Markovian infectious period models, and Neal and Roberts (2004), who performed MCMC for a model incorporating a spatial component of the distance between households. MCMC, combined with data augmentation as we will discuss in Section 1.3.4, allows for inference of data where the epidemic has only been partially observed and as such would be too complicated for standard statistical techniques. Specifically when infection times are unknown and so likelihood functions cannot be easily computed, as is common, inference can be made about both the parameters of interest and the missing data themselves.

More specifically, MCMC methods allow us to draw samples from a given distribution π (which we call the target distribution), even if π cannot be written down analytically. Using the notation we defined in Section 1.3.1, the target distribution would be the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{X})$, and referring back to Bayes' Theorem in Equation (1.3.1) we recall that the denominator is the part which is typically intractable. However, this is simply a normalising constant, which an MCMC approach does not require to be calcu-

lated explicitly. Instead, samples are drawn from the posterior distribution by constructing an ergodic Markov chain $\{Z_n\}$ which has stationary distribution $\pi(\boldsymbol{\theta} \mid \mathbf{X}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{X} \mid \boldsymbol{\theta})$. So long as we can calculate the likelihood and prior distribution, we can sample from the posterior distribution. After an initial 'burn-in' period (the early iterations of the MCMC algorithm which we discard), provided we have specified suitable mixing parameters, we can be confident that the chain has reached equilibrium (regardless of the starting location Z_0). The steps of the Markov chain then approximate samples from the target (posterior) distribution. We will now describe some well-known algorithms for obtaining these samples from the posterior distribution.

1.3.3.1 The Gibbs Sampler

The *Gibbs sampler* (Geman and Geman, 1984) samples from high dimensional distributions by breaking them down into lower dimensional sections. If $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ so that the posterior distribution is of dimension d, then for all $i \in \{1, \ldots, d\}$ a Gibbs sampler will simulate component θ_i from the conditional distribution $\pi(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d, \mathbf{X})$. These are referred to as the *full conditional distributions*. A summary of the Gibbs algorithm is given in Algorithm 1.

Although we often update parameters individually as in Algorithm 1, it is also possible to group related parameters together and perform block updates, using the full conditional distribution given all remaining parameters and the data. Blocking correlated parameters can improve convergence of the chain since correlation can lead to high rejection rates for individual updates. The sampler defined in Algorithm 1 is also known as a 'deterministic scan' Gibbs sampler, since we update all parameters deterministically in order. We can alternately use a 'random scan' Gibbs sampler which, at each iteration, picks at random one (or more) parameter(s) to update.

Algorithm 1 The Gibbs sampler, for obtaining I samples from a d-dimensional posterior distribution

1. Choose initial $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)}\right)$;

2.

for $i = 1$ **to** I **do**

i. Draw $\theta_1^{(i)} \sim \pi\left(\theta_1 \mid \theta_2^{(i-1)}, \ldots, \theta_d^{(i-1)}, \mathbf{X}\right)$;

ii. Draw $\theta_2^{(i)} \sim \pi\left(\theta_2 \mid \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_d^{(i-1)}, \mathbf{X}\right)$;

iii. Draw $\theta_3^{(i)} \sim \pi\left(\theta_3 \mid \theta_1^{(i)}, \theta_2^{(i)}, \theta_4^{(i-1)}, \ldots, \theta_d^{(i-1)}, \mathbf{X}\right)$;

...

d. Draw $\theta_d^{(i)} \sim \pi\left(\theta_d \mid \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{d-1}^{(i)}, \mathbf{X}\right)$;

end for.

1.3.3.2 The Metropolis-Hastings Algorithm

The Metropolis algorithm was introduced in Metropolis et al. (1953), and generalized to obtain the *Metropolis-Hastings* (MH) algorithm in Hastings (1970). Whereas the Gibbs sampler requires full conditional distributions to compute, the MH algorithm provides an alternative when this is not possible (as is common). Most MCMC algorithms can be considered as a special case of Metropolis-Hastings. Details of the procedure are given in Algorithm 2.

The MH algorithm requires a choice of proposal density q. For each proposed value $\boldsymbol{\theta}^*$, we calculate the acceptance probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, which is the ratio of the likelihood multiplied by this proposal density, evaluated at the proposed parameter value and the current parameter value. This describes how likely the proposed value is compared to the current value, and we accept the proposal with probability α. The acceptance probability is defined in this way to ensure that the stationary distribution of the Markov chain is the target posterior distribution as desired, and also that the chain strikes a balance between tending to visit high probability areas but also satisfactorily exploring the parameter space.

The simplest choice of q is known as the *independence sampler*. In this, q is inde-

Algorithm 2 The Metropolis-Hastings algorithm, for obtaining I samples from a d-dimensional posterior distribution

1. Choose initial $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)}\right)$;

2.

for $i = 1$ **to** I **do**

i. Draw candidate value $\boldsymbol{\theta}^*$ from proposal density $q\left(\boldsymbol{\theta}^{(i-1)}, \cdot\right)$;

ii. Calculate $\alpha\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*\right) = \min\left(1, \frac{\pi\left(\boldsymbol{\theta}^* \mid \mathbf{x}\right) q\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(i-1)}\right)}{\pi\left(\boldsymbol{\theta}^{(i-1)} \mid \mathbf{x}\right) q\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*\right)}\right)$;

iii. Draw $u \sim \text{Unif}(0, 1)$;

if $u \leq \alpha\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*\right)$ **then**

Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$

else

Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$

end if

end for.

pendent of the current value of the parameters, so that $q\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^*\right)$. Alternatively, *symmetric random walk Metropolis* (as introduced in Metropolis et al., 1953) sets $q\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(i-1)}\right)$, which causes the proposal densities in the acceptance ratio $\alpha\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*\right)$ to cancel. This method has become particularly popular since, in avoiding calculation of the proposal density in the accept/reject ratio, many calculations are avoided.

The choice of proposal density is key for the MH algorithm. The unrestricted choice of $q(\cdot, \cdot)$ is what allows the algorithm its wide generality, but these different choices may have great impact on performance. A low acceptance probability may lead to poor mixing of the Markov chain, whereas a high acceptance probability may lead to slow convergence. In reality, a balance must be achieved. Notably, Roberts et al. (1997) identified an asymptotically optimal acceptance rate of 0.234 for Gaussian random walk algorithms, so long as the target density consists of a product of i.i.d. components for each parameter. Tuning of the parameters which control the proposal density can be

performed manually, but the development of adaptive algorithms which automatically tune has also received much attention, for example Haario et al. (2014) and Andrieu and Thoms (2008).

1.3.3.3 Convergence and Dependency of the Markov Chain

One important factor when using an MCMC algorithm is the speed at which convergence to the equilibrium occurs, in practice determining the length of the burn-in period. We would like to ensure that equilibrium has been reached by the end of the burn-in period, to minimise the effect of the initial values chosen on the samples obtained. However, with each calculation of the likelihood expression being potentially costly, there is often great benefit in a shorter burn-in, and hence in ensuring fast convergence to the equilibrium.

In addition to this, there will be some degree of dependence between successive simulated values, with high dependence causing slow convergence also. The search for methods which allow for fast convergence and low sample dependence has received considerable attention.

Non-centered parameterisations (NCPs) are one such technique which aims to improve the efficiency of convergence. Full details and discussion can be found in Kypraios (2007) and Papaspiliopoulos et al. (2003). In essence, rather than the standard *centered parameterisation (CP)* of the unknown quantities $\boldsymbol{\theta}$, $\mathbf{Y}$, a non-centered parameterisation finds some alternative $(\boldsymbol{\theta}, \mathbf{Y}) \rightarrow (\boldsymbol{\theta}, \tilde{\mathbf{Y}})$ where new missing data $\tilde{\mathbf{Y}}$ is *a priori* independent of $\boldsymbol{\theta}$.

There has been considerable discussion around the use of NCPs, and if their use really provides benefit over standard CPs. Gelfand et al. (1996) argued strongly for the use of CPs, which can often be applied with fast Gibbs samplers when NCPs cannot, potentially undoing any computational advantage of a non-centered approach. However, as Kypraios (2007) argues, NCPs may offer considerable improvement in convergence in cases where the dependence between missing data and model parameters is high. One major issue

with the use of NCPs, however, is a requirement of orthogonality between $\boldsymbol{\theta}$ and $\tilde{\mathbf{Y}}$, which is often hard to achieve (Papaspiliopoulos et al., 2003). This has limited their application in practice.

In reality, the standard methods for achieving convergence and low dependence are somewhat crude. The length of the burn-in period is normally determined using trace plots of the parameter values from sample runs of the MCMC algorithm, either visually or using a diagnostic tool (e.g. Geweke, 1992). To minimise autocorrelation of MCMC samples (which is usually identified with ACF plots), the general solution is to only keep every n^{th} draw from the posterior. Known as thinning, this lowers sample dependence, but at the cost of more (potentially expensive) computations of the likelihood to obtain the same number of samples (see Gilks et al., 1995).

1.3.4 Data Augmented Markov Chain Monte Carlo

Data augmentation, as first seen in Tanner and Wong (1987) but also in Gelfand and Smith (1990), involves sampling from the predictive distribution of the missing data to obtain samples from the posterior. It has become probably the most widely used technique for Bayesian inference in missing data problems. In short, we assume knowledge of the missing data $\mathbf{Y}$ to obtain samples of both it and the parameters of interest, rather than having to compute the (usually intractable) integral in Equation (1.3.2). As we have discussed in Section 1.3, methods such as the expectation-maximisation algorithm do exist for dealing with missing data, and may be preferable in simpler cases, but data augmented MCMC (DA-MCMC) helps ensure identification of a global rather than a local maximum, as well as providing improved computation times when working in a high dimension.

When modelling with missing data, target density $\pi(\boldsymbol{\theta} \mid \mathbf{X})$ is now the joint posterior distribution of the parameters and the missing data, $\pi(\boldsymbol{\theta}, \mathbf{Y} \mid \mathbf{X})$. By augmenting knowledge of the unknown data (usually event times for infec-

tious disease data), simulation from the conditional distributions $\pi(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{X})$ and $\pi(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X})$ is tractable, and we do not require integration over the missing data (as discussed previously in Section 1.3.2).

To obtain samples from the new posterior density $\pi(\boldsymbol{\theta}, \mathbf{Y} \mid \mathbf{X})$, we may therefore use a two-component Gibbs sampler, as in Algorithm 1 but where we alternate simulations from $\pi(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{X})$ for parameters $\boldsymbol{\theta}$ and from $\pi(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X})$ for missing data $\mathbf{Y}$.

In reality, the full conditional distributions required for a Gibbs sampler are often not all available. In these cases, of particular use is the *Metropolis within Gibbs* algorithm (as suggested in Tierney, 1994), a hybrid of the MH algorithm and Gibbs sampler. In this, calculation of the full conditional distributions is replaced with direct simulation through a Metropolis-Hastings step. Although clearly useful when the conditional distributions are unavailable, the introduction of Metropolis steps can greatly decrease the speed of convergence. However, with augmented data we often have the case that $\pi(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{X})$ is available in closed form (and hence a simple Gibbs sampler may be used) but that $\pi(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X})$ is not, and an MH algorithm is required. In this case Metropolis within Gibbs is very useful, since we may perform a combination of Gibbs and Metropolis within Gibbs.

1.3.5 DA-MCMC for SIR models

Although we have discussed MCMC methods in general, in this section we more specifically focus on MCMC for SIR epidemic models. We will define the SIR likelihood expression which will be used throughout this thesis, as well as the general DA-MCMC algorithm used for inference.

We define the likelihood for an SIR model with homogeneous mixing as in Britton and O'Neill (2002), similar to e.g. O'Neill and Roberts (1999) and Gibson and Renshaw (1998). We assume a closed population of fixed size N within which occurs an outbreak of final size $n \leq N$, that is to say n in-

dividuals have been infected by the end of the outbreak. This version of the likelihood requires infected individuals to be labelled $1, 2, \ldots, n$, and associated with their corresponding infection and removal times i.e. each infective j has removal time r_j and infection time i_j. Individuals are ordered by removal time, so that $r_1 \leq r_2 \leq \cdots \leq r_n$. We define the set of removal times as $\mathbf{r} = \{r_j : j = 1, 2, \ldots, n$, where $r_1 \leq r_2 \leq \cdots \leq r_n\}$, and the set of infection times as $\mathbf{i} = \{i_j : j = 1, 2, \ldots, \kappa - 1, \kappa + 1, \ldots, n\}$. Individual κ is the initial infective, who is assumed to have been infected by an external source. The identity of individual κ is not usually known from the data. For ease of exposition of the likelihood defined below, we also define $i_j = \infty$ for all non-infectives $j = n + 1, \ldots, N$. The likelihood is then written as a product of the contributions from each individual. A model with unlabelled cases (where both removal times and infection times are ordered, so that r_j and i_j do not necessarily correspond to the same individual) was proposed by Bailey and Thomas (1971), but in this thesis we will only be concerned with labelled individuals.

The likelihood is built from three components which, as in Britton and O'Neill (2002), we refer to as L_1, L_2 and L_3. Firstly, the product term L_1 contains the contribution to the likelihood of the $n - 1$ infections which occur during the outbreak (we ignore the contribution from the initial infective κ, since they are assumed to have been externally infected).

To define this, we introduce the concept of *infectious pressure*, which susceptible individuals receive from current infectives. For an individual j who is susceptible at time t, we define β as the infectious pressure acting upon them from infective k, so that

$$\mathbb{P}\big(k \text{ infects } j \text{ in } (t, t + \delta t]\big) = \beta \delta t + o(\delta t).$$

Then, any susceptible j receives infectious pressure β at their time of infection from infected individual k if and only if $i_k < i_j < r_k$. That is, if and only if k was infectious at j's infection. L_1 is therefore given by the total infectious

pressure on each individual at their time of infection, so that:

$$L_1 = \prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta \, \mathbb{1}_{\{i_k < i_j < r_k\}}. \tag{1.3.3}$$

The remainder of the infection part of the likelihood is given by L_2, which contains the total infectious pressure exerted over the course of the epidemic. For each infective, this corresponds to them 'failing' to infect all other individuals, both those who ultimately become infected and those who remain susceptible. This is given by

$$L_2 = \exp\Big(-\beta \sum_{j=1}^{n} \sum_{k=1}^{N} (r_j \wedge i_k - i_j \wedge i_k) \Big), \tag{1.3.4}$$

where $a \wedge b$ is the minimum of a and b. Here, $r_j \wedge i_k - i_j \wedge i_k$ represents the time period for which infective j places infectious pressure on any individual k.

The contribution of the removal process is contained in L_3. This is given by the probability density function (PDF) (or probability mass function, though for simplicity we will usually refer to just PDFs) of the infectious period distribution, which in Section 1.1.2.1 we defined as $f_I(r_j - i_j \mid \boldsymbol{\theta})$, for infectious periods $r_j - i_j$ of all infectives j. This construction of the likelihood allows for any choice of infectious period distribution $f_I(\cdot)$. The infectious periods are assumed independent, and therefore,

$$L_3 = \prod_{j=1}^{n} f_I(r_j - i_j \mid \boldsymbol{\theta}). \tag{1.3.5}$$

Combining Equations (1.3.3), (1.3.4) and (1.3.5), the likelihood is given by

$$\begin{aligned} \pi(\mathbf{i}, \mathbf{r} \mid \beta, \boldsymbol{\theta}, \kappa, i_\kappa) \;=\; & \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta \, \mathbb{1}_{\{i_k < i_j < r_k\}} \right) \exp\Big(-\beta \sum_{j=1}^{n} \sum_{k=1}^{N} (r_j \wedge i_k - i_j \wedge i_k) \Big) \\ & \times \prod_{j=1}^{n} f_I(r_j - i_j \mid \boldsymbol{\theta}). \end{aligned} \tag{1.3.6}$$

Without knowledge of the infection times $\mathbf{i}$ however, this likelihood is intractable. Although theoretically possible, integrating out all of the unknown

infection times is unfeasible in practice for any more than a handful of infectives, since the region of integration becomes very complex. DA-MCMC is therefore usually implemented for inference of parameters β and $\boldsymbol{\theta}$, where the augmented data correspond to the unknown infection times.

Before we define the MCMC algorithm, prior distributions need to be defined for the initial infective and their infection time, the infection rate β and the infectious period parameters contained in $\boldsymbol{\theta}$. For simplicity, we assume here that the infectious periods are exponentially distributed with rate γ, so that $\boldsymbol{\theta}$ contains this single parameter. As in O'Neill and Roberts (1999), we assume conjugate gamma distributed prior distributions for β and γ. If $\Gamma(\sigma, \nu)$ represents a gamma distribution with shape σ and rate ν, we assume that β and γ have conjugate gamma distributed prior distributions with parameters $(\sigma_\beta, \nu_\beta)$ and $(\sigma_\gamma, \nu_\gamma)$, respectively.

For the initial infective κ, we may choose from a variety of potential prior distributions. We may use some prior knowledge to determine their identity, or we may place a uniform prior distribution over all infectives so that each is equally likely to be the initial one. In practice, we often assume $\kappa = 1$ is known for simplicity. For i_κ, we define a uniform prior distribution on $(-\infty, r_1)$.

We multiply the likelihood expression in Equation (1.3.6) by these prior distributions to obtain the posterior distribution:

$$\pi(\beta, \gamma, i_\kappa, \mathbf{i} \mid \mathbf{r}) \quad \propto \quad \pi(\mathbf{r}, \mathbf{i} \mid \beta, \gamma, i_\kappa)\pi(\beta)\pi(\gamma)\pi(\kappa)\pi(i_\kappa) \tag{1.3.7}$$

This is then the target density of the MCMC algorithm. In order to perform Gibbs updates for β and γ, however, we also require their full conditional distributions.

To obtain the full conditional distribution for β, we consider Equations (1.3.7) and (1.3.6), and see that

$$\pi(\beta \mid \gamma, \kappa, i_\kappa, \mathbf{i}, \mathbf{r}) \propto \beta^{n-1} \mathrm{e}^{-\beta A} \pi(\beta), \tag{1.3.8}$$

where $A = \sum_{j=1}^{n} \sum_{k=1}^{N} (r_j \wedge i_k - i_j \wedge i_k)$.

Similarly for γ:

$$\pi(\gamma \mid \beta, \kappa, i_\kappa, \mathbf{i}, \mathbf{r}) \propto \gamma^n \mathrm{e}^{-\gamma B} \pi(\gamma), \tag{1.3.9}$$

where $B = \sum_{j=1}^{n} (r_j - i_j)$, since the only terms in the likelihood with γ dependence are the n exponential PDFs of the infectious periods in the removal part.

Combining Equations (1.3.8) and (1.3.9) with the gamma conjugate prior distributions defined above, the full conditional distributions for β and γ are given as follows:

$$\begin{aligned} \beta \mid \gamma, \kappa, i_\kappa, \mathbf{i}, \mathbf{r} &\sim \Gamma(\sigma_\beta + n - 1, \nu_\beta + A) \\ \gamma \mid \beta, \kappa, i_\kappa, \mathbf{i}, \mathbf{r} &\sim \Gamma(\sigma_\gamma + n, \nu_\gamma + B). \end{aligned} \tag{1.3.10}$$

Overall, the basic DA-MCMC algorithm for an SIR model given removal data is given in Algorithm 3. Samples of the model parameters β and γ may be obtained using Gibbs steps as the full conditional distributions are available, but the infection times must be updated with Metropolis Hastings steps since sampling directly from their posterior is not possible. For an individual j selected uniformly at random from $1, \ldots, n$, we propose candidate value i_j^* from $q(\cdot \mid i_j)$. This is accepted with probability:

$$\min\left(1, \frac{\pi(\mathbf{i}^*, \mathbf{r} \mid \beta, \gamma, \kappa^*, i_\kappa^*) q(i_j \mid i_j^*)}{\pi(\mathbf{i}, \mathbf{r} \mid \beta, \gamma, \kappa, i_\kappa) q(i_j^* \mid i_j)}\right).$$

A common choice for the proposal density is $f_I(\cdot \mid \gamma)$, which causes the proposal densities in the acceptance probability to cancel. After the infection times have been updated, we set κ accordingly, as the individual with the earliest current infection time.

The likelihood and MCMC algorithm may be similarly extended to SEIR and other compartmental models, as well as heterogeneous mixing models.

Although the technique has been widely adopted, there are a number of limitations to data augmented MCMC techniques for epidemic data analyses. Especially with growing demand for fast (potentially real-time) analysis, computationally efficient analysis is key. However, DA-MCMC methods for epidemic data often struggle with the issues of high dependence and slow chain

Algorithm 3 DA-MCMC algorithm for obtaining J samples of the model parameters, for an SIR model with known removal times, unknown infection times, infection rate β and infectious period parameter γ. Here, x^j represents the jth sample of parameter x.

1. Start the Markov chain from initial values β^0, γ^0, $\mathbf{i}^0, \kappa^0, i_\kappa^0$;
2. **for** $j = 1$ **to** J **do**
   i. Update β using a Gibbs step to draw from $\pi(\beta \mid \gamma^{j-1}, \kappa^{j-1}, i_\kappa^{j-1}, \mathbf{i}^{j-1}, \mathbf{r})$ and obtain sample β^j;
   i. Update γ using a Gibbs step to draw from $\pi(\gamma \mid \beta^j, \kappa^{j-1}, i_\kappa^{j-1}, \mathbf{i}^{j-1}, \mathbf{r})$ and obtain sample γ^j;
   iii. Choose uniformly at random one or more infection times i_k, for $k = 1, \ldots, n$ (including the initial infective). Update each using a Metropolis Hastings step to obtain $\mathbf{i}^j / i_\kappa^j$. Update κ correspondingly as required;

end for.

convergence discussed in Section 1.3.3.3, particularly for large outbreaks or complex model structures. This will be explored in more detail in Chapter 3, but essentially the unknown infection times and the parameters governing the infectious period lengths often have a high posterior correlation, leading to slow mixing of the Markov chain. This motivates the development of alternate methods which either solve, or avoid, these problems.

1.4 Approximation Methods for Infectious Disease Modelling

Without any restrictions, of course an ideal mathematical model would be a complete virtual representation of the real world. However, due to limitations in computational power this is of course not possible, and efforts must be made to obtain the best possible partial description of reality. All mathematical

models contain some degree of simplification and approximation, and when introducing greater complexity we must consider the computational cost.

Current methods for infectious disease analysis such as DA-MCMC, as we have discussed, are known to become computationally demanding for large population sizes or complex mixing structures. When working with large amounts of data it often becomes necessary to use very simplistic models due to computational restraints, especially when performing time-sensitive analyses. We suggest that likelihood approximation methods may be a useful tool for dealing with this, and these will be the focus of chapters 3 and 4. If we can obtain likelihood expressions which, although including extra levels of approximation to the 'true' likelihoods used in data augmentation, result in faster computation, this allows for more realistic models to be used in turn.

This is not an area in which there has been much previous work. As we will explore more fully in Chapter 3, the likelihood approximation methods we will define in this thesis bear some similarity to composite likelihood methods (see Varin et al., 2011 and Cox and Read, 2004), in that we attempt to build our understanding of the overall system dynamics by considering what are essentially marginal densities, but in reality these methods are actually quite different. Otherwise, attempts to tackle the challenges we have discussed largely focus on ideas other than direct likelihood approximation.

1.4.1 Model Approximation

Model approximation is an area which has seen some focus. This involves consideration of an approximation to the model, under which the likelihood can be directly computed. One such simple approximation is the assumption of fixed length infectious periods. Under this assumption, all of the unknown infection times can be directly determined from the data, and the likelihood expression becomes tractable. Inference may then be performed using standard techniques for a completely observed outbreak (see e.g. Andersson and

Britton, 2000, Section 9).

Model approximation was also performed in Britton and Becker (2000), where they defined a two-level mixing model incorporating within-household local transmission and between-household global transmission. Non-independence of individuals residing in different households results in an intractable likelihood, but in replacing the global transmission dynamics with a fixed probability that each individual avoids infection from the global source (originally seen in Addy et al., 1991), Britton and Becker (2000) obtain a model with a tractable likelihood.

A further example is Filipe and Gibson (1998), who defined a spatio-temporal stochastic model for disease transmission by modelling individuals as the vertices of a square lattice (so that all individuals have four nearest neighbours with whom to interact, except those on the boundary who have two or three). They made the assumptions that the lattice was large enough that the boundary effects could be ignored, and that the initial distribution of infected individuals was from a spatially stationary process. This allowed them to make a deterministic approximation to the stochastic model, expressing the overall disease dynamics in terms of smaller cluster approximations. Cauchemez and Ferguson (2008), on the other hand, approximated a continuous-time SIR model by dividing the outbreak into a series of observation periods, and augmenting the data with the latent state of the system (i.e. the total number of infectives and susceptibles) at the beginning of each period. They then mimicked the SIR process with a diffusion process with known exact solution. Becker (1989) also includes numerous examples of simplified models used in practice.

As we have seen here, a variety of model approximation methods have been suggested in the literature, but often these have been only applied to a specific data set or model and hence lack general applicability. In contrast, we will explore likelihood approximation methods in this work. These will share similar themes with some of these model approximations, in making assumptions to

result in a tractable likelihood and especially in using small clusters to build up overall disease dynamics, but they will not require any simplifications of the actual models used.

1.4.2 Approximate Bayesian Computation (ABC)

One other considerable area of research has been Approximate Bayesian Computation, or ABC. Rather than approximating either the likelihood or model, ABC is what is known as a likelihood-free method. It was initially proposed in Rubin (1984), and then grew in popularity within the population genetics literature from Tavaré et al. (1997). It has since become much more widely applied within, for example, ecology, systems biology and evolutionary biology (e.g. Toni et al., 2009 and Csilléry et al., 2010), as well as infectious disease modelling (e.g. McKinley et al., 2009, Tanaka et al., 2006 and Blum and Tran, 2010). Its use in this field relies upon the fact that stochastic models for disease outbreaks are generally straightforward to simulate, and that this can be done very quickly for a given set of model parameters.

In its basic form, ABC is essentially a form of rejection algorithm. However, a benefit is that it can easily be incorporated into MCMC and Sequential Monte Carlo (SMC) algorithms. It is of particular use for likelihoods which are computationally intractable or of a high cost to evaluate, since it replaces likelihood calculations with comparisons between observed and simulated data. We review ABC methods and their implementation here, but a more detailed discussion on their use with stochastic epidemic models may be found in Kypraios et al. (2017).

If $\boldsymbol{\theta}$ contains the parameters to be estimated (for a disease outbreak we might have $\boldsymbol{\theta} = (\beta, \gamma)$ where β is the infection rate and γ the removal rate), as defined in Section 1.3.1, we wish to approximate the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{X})$, for data $\mathbf{X}$. Rather than explicitly calculating likelihood $\pi(\mathbf{X} \mid \boldsymbol{\theta})$, ABC methods take the general steps:

1. Sample a candidate $\boldsymbol{\theta}^*$ from the prior distribution $\pi(\boldsymbol{\theta})$
2. Simulate a data set (outbreak) $\mathbf{X}^*$ from the model with parameters $\boldsymbol{\theta}^*$.
3. Compare simulated data set $\mathbf{X}^*$ with observed data $\mathbf{X}$ using some distance measure d. If $d(\mathbf{X}^*, \mathbf{X}) \leq \epsilon$ for tolerance ϵ, we accept $\boldsymbol{\theta}^*$.

(1.4.1)

This process repeats until we have accepted a pre-determined number of $\boldsymbol{\theta}^*$ values. The result is a sample of parameters drawn from $\pi(\boldsymbol{\theta} \mid d(\mathbf{X}^*, \mathbf{X}) \leq \epsilon)$, which for suitable d and sufficiently small ϵ should well approximate $\pi(\boldsymbol{\theta} \mid \mathbf{X})$. ABC then requires no calculations of the likelihood. We simply require a choice of distance function d which measures the similarity between two outbreaks, and a tolerance ϵ which defines how close this distance must be to be accepted. In practice, this choice of d is by no means trivial, however. Often we instead measure the distance between some summary statistics of the data, such that we require $d(S(\mathbf{X}^*), S(\mathbf{X})) \leq \epsilon$. For infectious disease data, this might measure the difference in the removal curves via a sum of squared differences between the observed and simulated data, for example, as in McKinley et al. (2009).

ABC is frequently combined with MCMC and SMC since the rejection algorithm in (1.4.1) often suffers from very low acceptance rates. In ABC-MCMC, a Markov chain is generated with stationary distribution $\pi(\boldsymbol{\theta} \mid d(\mathbf{X}^*, \mathbf{X}) \leq \epsilon)$, so that parameters are usually sampled from the vicinity of their current values. This may suffer from similar problems to DA-MCMC however: that correlated parameters may cause slow convergence of the Markov chain.

ABC with SMC, as initially introduced in Sisson et al. (2007), seeks to avoid these convergence problems. Sequential Monte Carlo methods involve sampling from a series of proxy distributions which converge to the posterior, rather than the posterior itself. We define a set of N distributions $\pi_1, \pi_2, \ldots, \pi_N$, where $\pi_N = \pi(\boldsymbol{\theta} \mid \mathbf{X})$ is the posterior of interest and $\pi_1 > \pi_2 > \cdots > \pi_N$. We

then initially draw a large number of samples, called particles, from π_1 (which has been defined such that direct sampling is possible). The particles are then passed through a series of sequential importance sampling steps i which involve being weighted assuming they come from the corresponding sequential distribution π_i. The general idea is that the intermediary distributions π_i tend gradually towards the target distribution, so the method proceeds by moving and weighting the particles by how well they fit each successive distribution π_i. For ABC-SMC, these distributions π_i are simply defined as $\pi_i(\boldsymbol{\theta} \mid d(\mathbf{X}^*, \mathbf{X}) \leq \epsilon_i)$, for $i = 1, \ldots, N$ and tolerances $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_N$. This should, in principle, avoid the algorithm getting stuck in areas of low probability, decreasing the time to convergence. Certain choices of distance function d and tolerance ϵ can cause particle degradation, however, wherein after the particles have been passed through some of the intermediary distributions, only a few remain with non-zero weight. To combat this, a resampling stage is often introduced. In this, particles are resampled proportionately to their weight when degradation becomes high.

Overall, ABC and ABC-SMC methods are useful for epidemic models since they are widely applicable and may be used with models involving complex populations (see e.g. Brooks-Pollock et al., 2014 who used a spatial stochastic model for bovine tuberculosis, incorporating within-farm and between-farm transmission). However, they can still result in algorithms which are inefficient or slow, especially when the number of parameters to estimate is large. Choice of tolerance ϵ and distance function d can have considerable impact on the results found, and use of a summary statistic S also introduces additional bias into the method through loss of information. These must therefore be carefully selected. For further reading, Sunnåker et al. (2013) discuss many of the perceived drawbacks of ABC methods.

1.5 Structure of the Thesis

The remainder of this thesis is divided into two distinct parts. The first part, published in Stockdale et al. (2017), concerns a Bayesian analysis of the Abakaliki smallpox data. This data set has been much cited in the field of stochastic epidemic modelling, but never analysed in its full form using a true likelihood method. We seek to compare parameter estimates from our Bayesian analysis to estimates from Eichner and Dietz (2003), who did analyse the full data set but using an approximate likelihood. This will be presented as follows:

- **Chapter 2**. After introducing the Abakaliki smallpox data, we define the model to be used in our analysis. This is a variant of an SEIR model, and is the same as that used in Eichner and Dietz (2003) to ensure comparability. We outline the process for simulating from this model, since model assessment will be performed via simulation-based techniques. We also describe the Bayesian inference to be performed, and define the likelihood expression. We then detail the MCMC algorithm to be used, before concluding with the results of the analysis, and a discussion of these.

The second part of this thesis will focus on the development of likelihood approximation methods for the analysis of infectious disease data. As we have briefly discussed and will further explore in Chapter 3, current methods, such as the data augmented MCMC performed in Chapter 2, become computationally cumbersome for large populations or large amounts of missing data, as well as being burdened by correlation problems of this missing data. We seek to develop likelihood approximation methods which remove the need for data augmentation, allowing simpler MCMC or maximum likelihood estimation to be performed more easily.

- **Chapter 3**. This chapter will describe the development of two different likelihood approximation methods. Firstly, we introduce a generalised approximation based upon the method of Eichner and Dietz (2003) from

Chapter 2. We will then proceed to construct a new approximation; the Pair Based Likelihood Approximation (PBLA) method. We will define a number of different versions of this, for example for offering increased computational speed under more restrictive modelling assumptions.

- **Chapter 4**. The final chapter will involve a series of simulation studies, which compare parameter estimation using the likelihood approximation methods to standard DA-MCMC, followed by application to various real data sets. Specifically, we consider data from a respiratory disease outbreak on the Atlantic island of Tristan Da Cunha, from the West African Ebola epidemic of 2013-2016 and from the 2001 UK Foot and Mouth disease outbreak. These data sets will each have different requirements in terms of modelling, allowing us to analyse the performance of the PBLA method in different settings.

We will finally summarize the work presented in this thesis, in addition to possible areas of further research, in Chapter 5.

CHAPTER 2

Modelling and Bayesian Inference for the Abakaliki Smallpox Data

In this chapter we introduce the much-cited Abakaliki smallpox data set, and perform its first full Bayesian analysis to include all aspects of the data. This work is published in Stockdale et al. (2017).

We begin with an overview of the data and its relevance within the field of mathematical disease modelling, before defining the model and performing Bayesian inference to obtain estimates of model parameters. There exists one previous analysis of the full data set, but this relies upon an approximate likelihood. We instead use Markov Chain Monte Carlo methods with the true likelihood, which avoid the need for likelihood approximations. In addition to the basic model parameters, which will be compared to the results of the approximation method as well as interpreted in their real-world context, we estimate the path of infection and perform model assessment with simulation based methods.

2.1 Introduction

In 1967, an outbreak of Smallpox occurred in the Nigerian town of Abakaliki. The vast majority of cases were members of the Faith Tabernacle Church

(FTC), a religious organisation whose members refused vaccination. The outbreak was recorded in detail in a World Health Organisation (WHO) report by Thompson and Foege (1968), with information on not only the time series of case detections but also their place of dwelling, vaccination status, and FTC membership. The outbreak has inherent historical interest as it occurred during the WHO Smallpox eradication programme initiated in 1959. Although Smallpox was declared eradicated in 1979, it regained attention as a potential bioterrorism weapon in the early 2000s (see e.g. Gani and Leach, 2001, Meltzer et al., 2001, Halloran et al., 2002) and continues to be of interest due to concerns about its re-emergence or synthesis, (see e.g. Henderson and Arita, 2014, Eto et al., 2015, World Health Organisation, 2015 and references therein). Estimates of the parameters governing disease transmission are of considerable importance for public health planning, and thus being able to accurately obtain such quantities from available data is of considerable importance also.

Within the mathematical infectious disease modelling literature, the Abakaliki smallpox data set has been frequently cited, the first appearance being Bailey and Thomas (1971). The data set consists of a time series of symptom appearance (rash) times for the 32 individuals who were infected, along with other information on the composition of the population: FTC membership status, vaccination status and compound number (the affected individuals lived in a series of compounds; houses built around a central courtyard). Most analyses of the Abakaliki data, however, have used only the rash times, such as Shanmugan (2011) and Oh (2014), or indeed only the final size, as in Ball et al. (2002). In fact, to our knowledge, in all but one case the data set has been used as an example for new methodology; taking primarily the case detection times and not considering the other aspects of the data. Lau and Yip (2008), Huggins et al. (2004) and Yip (1989) used the data set to demonstrate Martingale-based methods for inference of the basic reproduction rate, the initial number of susceptibles and the infection rate, respectively. O'Neill and Becker (2001), McKinley et al. (2014), Boys and Giles (2007) and Golightly et al. (2014) introduced new MCMC techniques which the Abakaliki data rash times

were used to illustrate. In addition to incomplete use of the data, the suitability of the data to the method at hand is often not considered: in many cases, such as Clancy and O'Neill (2008), Kypraios (2009), Xiang and Neal (2014) and Becker (1976), an SIR (susceptible-infective-removed) model has been used for this outbreak despite smallpox being known to have an incubation period (see Ferguson et al., 2003). Ray and Marzouk (2008) used binomial random graphs to model inter- and intra-compound contacts, thus including the compounds feature of the data, though they did not consider the difference between FTC and non-FTC individuals or make use of the vaccination data. In fact, all of the aforementioned papers only considered a population of size 120, which is the number of FTC individuals inside the compounds; disregarding all non-FTC individuals and FTC individuals outside the affected compounds.

To our knowledge, the only paper which has analysed the full Abakaliki data set is Eichner and Dietz (2003). The authors use an individual-based stochastic transmission model which takes into account the natural disease progression of smallpox, as well as the introduction of control measures, the population structure and the vaccination statuses of individuals within it. Parameter estimation is then performed via maximum likelihood. However, one notable feature of this analysis is that it relies on an approximate likelihood, which in particular assumes that the likelihood contributions made by different infected individuals are mutually independent. The true likelihood of the observed data given the model parameters is practically intractable as it involves integrating over all possible unobserved events, and so Eichner and Dietz perform a back-calculation. This involves reconstructing the outbreak backwards (or forwards, in the case of removal times) from the data and assumed knowledge of the disease stages, for example the latent period or infectious period. This relies on approximation however, since calculations do not take into account that infectious pressure may vary during an individual's disease stage as other individuals are infected or recover. Particularly used for HIV analysis in the 1990s, similar methods can be found in Becchetti et al. (1993) and Brookmeyer (1991). The use of this approximation immediately raises the question

of how well it performs, and in particular how different the parameter estimates might be if the analysis was instead based on the true likelihood.

We seek to perform a full Bayesian analysis for the Abakaliki data, avoiding approximations by using data-augmentation to produce an analytically tractable (and correct) likelihood. In Section 2.2 we will introduce the data, before defining the model in Section 2.3. Section 2.4 concerns a simulation study, performed to confirm that the method is working as required. Section 2.5 will describe the Bayesian inference and calculations to obtain a full likelihood expression. Following this, Section 2.6 will describe the MCMC performed. We use the true likelihood to estimate the parameters of interest, using a Bayesian framework, data augmentation and MCMC. Section 2.7 then includes the results of this analysis and Section 2.8 discussion of these. We interpret the posterior estimates of the parameters in the model as well as a variety of reproduction numbers; directly comparing them to those obtained by Eichner and Dietz in their analysis. In addition we perform a sensitivity analysis, model checking, and consider the results of the simulation study. We also estimate quantities derived via data-augmentation, such as who-infected-whom and the time of infection for each individual, which do not feature in the analysis of Eichner and Dietz.

2.2 The Data

The data, as given in Thompson and Foege (1968), are structured as shown in Table 2.1, which displays information on the 32 cases of smallpox: when the infected individual's rash became apparent, their compound identifier, FTC membership status and vaccination status. We have defined our timescale by setting day 0 as the date of the first rash onset.

All of the infected individuals lived in compounds; these are typically one-storey dwellings housing several families and built around a central courtyard. The composition of the affected compounds is provided in Table 2.2,

where the total numbers of vaccinated and non-vaccinated FTC and non-FTC members within each compound are displayed. FTC members were known to mix frequently with one another, whilst remaining rather isolated from the rest of the community. FTC members also refused vaccination, although many of them were vaccinated prior to joining the organisation. For use in Table 2.2, we define quantities $n_{c,FTC}$ and $n_{c,non}$ as the total number of FTC and non-FTC individuals residing in compound c, respectively.

Note that on the 25th day after the rash of the initial infective became apparent, four FTC individuals from compound 1 (three vaccinated and one non-vaccinated) moved to compound 2. Table 2.1 and Table 2.2 show the composition of the compounds after the move. In addition, quarantine measures were put in place in Abakaliki, but not until part way through the outbreak. The exact time these measures were introduced was not recorded.

2.3 Model Structure

2.3.1 Population Structure

We wish to define a model for smallpox outbreaks in the town of Abakaliki, using the information provided in the WHO report by Thompson and Foege (1968). Beginning with notation, consider Abakaliki a closed population with $N = 31200$ individuals, labelled $0, \ldots, N-1$. The population is partitioned by compound: either one of the nine listed in the WHO report or 'outside' meaning within the town but not the affected compounds, and by confession: either belonging to the Faith Tabernacle Church (FTC) or not. Individuals $0, 1, \ldots, n_{com} - 1$ are those inside the compounds, where $n_{com} = 251$ is defined as the number of people living within the affected compounds. Any individual $k = 0, \ldots, N-1$ may be categorised as type (c_k, f_k), where $c_k = 1, \ldots, 9$ is the compound of k, or $c_k = 0$ indicates that k is outside the compounds. Similarly, f_k is k's confession (faith); FTC or non-FTC. These types may lead to

Table 2.1: Smallpox cases in Abakaliki during 1967, from Thompson and Foege (1968). Compounds are listed after the move of cases 7 and 8 and two non-infectives, on day 25 from compound 1 to 2.

Case number	Day of rash onset	Compound	Confession	Vaccination
0	0	1	FTC	No
1	13	1	FTC	No
2	20	1	FTC	No
3	22	1	FTC	No
4	25	1	FTC	No
5	25	1	FTC	No
6	25	1	FTC	No
7	26	2	FTC	Yes
8	30	2	FTC	Yes
9	35	1	FTC	No
10	28	4	FTC	No
11	40	5	FTC	No
12	40	1	FTC	No
13	42	1	FTC	No
14	42	1	FTC	No
15	47	1	FTC	No
16	50	5	FTC	No
17	51	2	FTC	No
18	55	1	FTC	No
19	55	2	FTC	No
20	56	6	Non	Yes
21	56	5	FTC	Yes
22	57	2	FTC	Yes
23	58	7	FTC	No
24	60	4	FTC	No
25	60	2	FTC	No
26	61	2	FTC	No
27	63	8	Non	Yes
28	66	3	FTC	No
29	66	9	FTC	No
30	71	5	FTC	No
31	76	2	FTC	Yes

Table 2.2: Composition of the compounds affected by smallpox in Abakaliki, Nigeria during 1967, from Eichner and Dietz (2003)

Compound	FTC			Non-FTC		
	Vaccinated	Nonvaccinated	$n_{c,FTC}$	Vaccinated	Nonvaccinated	$n_{c,non}$
1	18	15	33	0	0	0
2	9	5	14	1	0	1
3	2	8	10	0	0	0
4	2 - i_4	2 + i_4	4	28 + i_4	1 - i_4	29
5	4 - i_5	3 + i_5	7	13 + i_5	2 - i_5	15
6	0	0	0	40	3	43
7	4 - i_7	1 + i_7	5	12 + i_7	3 - i_7	15
8	0	0	0	37	5	42
9	0	1	1	26	6	32
Sum 1-9	35	39	74	161	16	177
Outside	46 x 35/74	46 x 39/74	46	30903 x 161/177	30903 x 16/177	30903
Total			120			31080

We do not have complete vaccination status for all compounds, and so we use i_4, i_5, i_7 to allow for different possible configurations. Since the total number of vaccinated individuals in each compound is known as well as the total number of FTC and non-FTC, but not the confession of all vaccinated individuals, we are able to derive a system of linear constraints. Namely, $i_4 + i_5 + i_7 = 4$, $i_4 \in \{0,1\}, i_5 \in \{0,1,2\}$ and $i_7 \in \{1,2,3\}$ (the first constraint as we know the total number of FTC and non-FTC individuals, and the latter three to ensure we do not allow negative numbers of people in any category). Note: this table displays the compound composition after the move of the four individuals from compound 1 to compound 2 on day 25.

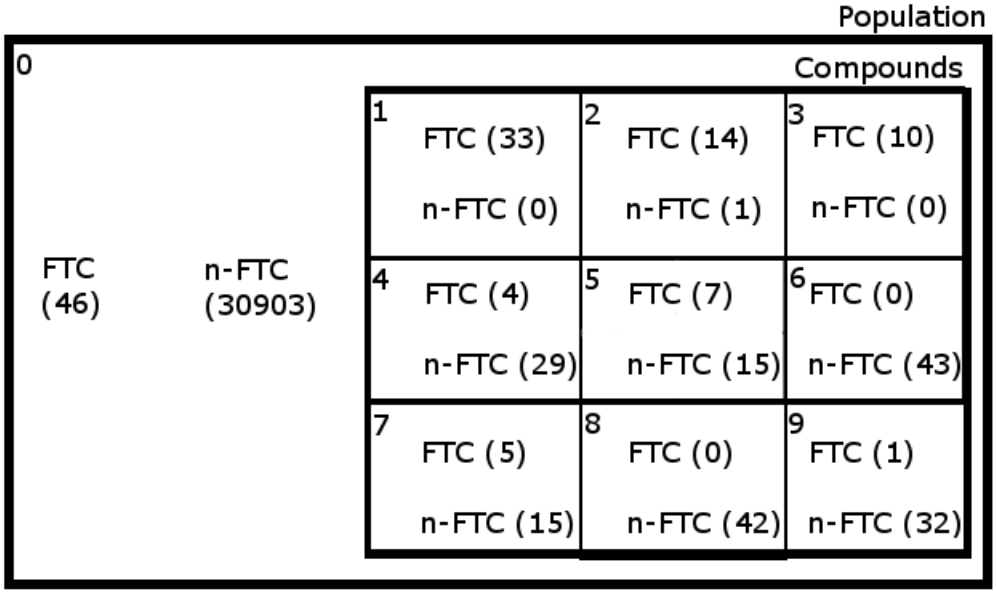


Figure 2.1: The structure of the population of Abakaliki as used in this study. FTC = member of the Faith Tabernacle Church, n-FTC = not a member of the FTC. Numbers in brackets represent the number of individuals within that category, after the move of four individuals on day 25 as detailed in Section 2.2.

differences in the mixing behaviour of individuals, though otherwise individuals are considered to be identical in their susceptibility to smallpox and their ability to infect others.

Figure 2.1 shows the population structure of the town of Abakaliki during the epidemic. Within the population are the compounds as described in the WHO report. There are nine of these compounds, and within any one may reside individuals of confessions FTC and non-FTC.

2.3.2 Transmission Model

We now describe a stochastic disease-transmission model for the spread of smallpox throughout the population of Abakaliki. This model is essentially the same as that described in Eichner and Dietz (2003), and is a variant of an SEIR (Susceptible-Exposed-Infective-Removed) model. Defining a time scale as in

Table 2.1, where time zero is the initial infective's rash time, at any given time t, each individual in the population will be in any one of six states: susceptible, exposed, with fever, with rash, quarantined or removed. Any susceptible may become exposed, as described below, and enter the incubation (or latent) period. During this stage individuals are not yet infectious. They will next enter the fever stage of the disease, at which point they become able to infect others. During the rash stage which follows, they will remain infectious but at a potentially different rate. We define the infectious period as the combined time spent in the fever and rash stages. Infectious individuals will either become removed (namely recovery or death, which we do not distinguish between) or isolated, this being the individual quarantined and henceforth unable to infect others. Quarantine procedures involve the removal of the individual from the population, these measures only being introduced part way through the outbreak at unknown time t_q.

Figure 2.2 is a visual representation of how any given individual may progress through the stages of susceptible, exposed, infectious with fever, infectious with rash (at the beginning of which their infectivity is changed) and finally either removal or quarantine; whichever comes first. For $j = 0, ..., N-1$, let e_j, i_j, r_j, q_j, τ_j denote, respectively, the times of exposure, fever, rash, quarantine and removal for individual j. If j never becomes infected, $e_j = i_j = r_j = q_j = \tau_j = \infty$. We assume that the epidemic is initiated by a single exposed individual, whom we label κ. We define the sets of these times as $\mathbf{e}$ (not including e_κ), $\mathbf{i}$, $\mathbf{r}$, $\mathbf{q}$ and $\boldsymbol{\tau}$.

The lengths of time spent in each disease progression stage for different individuals are assumed random with specified distributions, and mutually independent. The periods of time as identified in Figure 2.2 are distributed as follows:

- $i_j - e_j \backsim \Gamma\big((\frac{\mu_I}{\sigma_I})^2, \frac{\mu_I}{\sigma_I^2}\big)$,
- $r_j - i_j \backsim \Gamma\big((\frac{\mu_F}{\sigma_F})^2, \frac{\mu_F}{\sigma_F^2}\big)$,

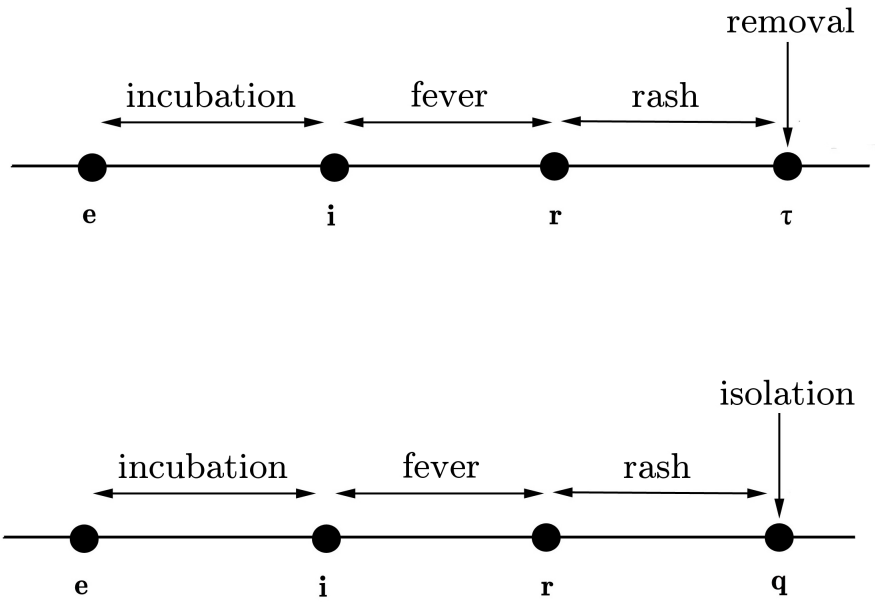


Figure 2.2: Disease progression in the smallpox model. The top image represents the infection of an individual who is removed through recovery or death, and the bottom shows an infection of someone who is quarantined. Isolation is only possible once quarantine measures have been introduced at time t_q.

- $\tau_j - r_j \backsim \Gamma\big((\frac{\mu_R}{\sigma_R})^2, \frac{\mu_R}{\sigma_R^2}\big)$,
- $q_j = \begin{cases} r_j + \Gamma(1, \frac{1}{4}) & \text{if } r_j \geq t_q, \\ t_q + \Gamma(1, \frac{1}{4}) & \text{if } r_j < t_q, \end{cases}$

where μ_A and σ_A represent the mean and standard deviation of the gamma distributed disease stage A for $A = I$, F and R, all of which are assumed known. Additionally, t_q denotes the time at which quarantine measures are introduced. The values of the means and standard deviations have been taken from Eichner and Dietz (2003), as shown in Table 2.3. Once quarantine measures are introduced, an individual may be put into isolation after a random delay following their rash onset date. Specifically, we define the quarantine time of individual j as $q_j = \max(r_j, t_q) + \Gamma(1, \frac{1}{4})$, as assumed by Eichner and

Table 2.3: Durations of periods in the infection process for Abakaliki smallpox outbreak. Time until quarantine determined by the maximum of rash time and time quarantine measures were introduced, t_q.

	Mean (days)	Standard deviation (days)
Period before fever	$\mu_I = 11.6$	$\sigma_I = 1.9$
From fever to rash	$\mu_F = 2.49$	$\sigma_F = 0.88$
From rash until recovery	$\mu_R = 16.0$	$\sigma_R = 2.83$
From rash to quarantine or from t_q to quarantine	$\mu_Q = 2.0$	$\sigma_Q = 2.00$

Dietz (2003). This means that no quarantining occurs prior to time t_q, after which it takes an average of two days for a detected individual to be placed in isolation, with standard deviation 2. Note also that an infected individual will have both a removal time and a quarantine time, for computational ease. Both quantities appear in the likelihood function, but in reality only the earlier of the two events takes place.

The epidemic begins at time e_κ with the exposure of the initial infective κ. Recall that the infectious period is defined in two parts: the fever period and the rash period, during each of which an individual will be infectious, but at potentially different rates. During their rash period, an individual j will have contacts with other members of their compound who are of the same confession at times given by a Poisson process of rate λ_h per day. Individuals outside of the nine compounds do not have such contacts. In addition, FTC individuals will have contacts at rate λ_f per day with other FTC individuals and contacts at rate λ_a per day with anybody in the population. Non-FTC individuals are assumed to have contacts with anybody in the population at rate $\lambda_a + \lambda_f$ per day since no information on their close contacts is available. During the fever period, the infections occur in exactly the same manner except that all rates are multiplied by factor b to account for the difference in infectiv-

ity. Typically, $b < 1$. In each case, the individual actually contacted is chosen uniformly at random from the pool of potential individuals in question. For example, contacts made with the entire population are drawn from the $N - 1$ other individuals. Note that this means that the individual-to-individual contact rate for such contacts is $\frac{\lambda_a}{N-1}$. Any contact from an infective to a susceptible results in immediate exposure of the susceptible. All of the Poisson processes describing contacts are assumed to be mutually independent

In addition, a proportion of the population is vaccinated. Vaccination status of all but a few individuals within the compounds is assumed known, and the proportions of FTC/non-FTC vaccinated individuals outside the compounds is assumed equivalent to inside. However, this vaccination is not necessarily effective: each recipient of the vaccine is completely protected with probability v, or remains completely susceptible with probability $1 - v$. Although the total number of vaccinated individuals is known, we do not have complete information on the composition of individuals with respect to vaccination status and FTC membership and so there are five potential configurations of twelve individuals with unknown details to consider, as shown in Table 2.4. For each individual in the population we have a vaccination status, assumed known for most individuals, and a protection status, unknown.

All individuals within the population remain living in their compounds, with the exception of four individuals who moved from compound 1 to compound 2 on the 25th day after the initial infection, two of whom later become infective.

The epidemic continues until there are no infectious or exposed individuals remaining in the population, at which point each person will either still be susceptible, or will have been quarantined/removed. We do not allow for reinfection.

Table 2.4: Possible combinations of twelve individuals, labelled 183, 213, 214, 215, 217, 231, 232, 233, 234, 236, 237, 250, with unknown vaccination status

Combination	Compound	FTC		Non-FTC	
		Vaccinated	Nonvaccinated	Vaccinated	Nonvaccinated
1	4		213	183	
	5	215, 214			217, 231
	7		232, 233, 234	236, 237, 250	
2	4	213			183
	5	215	214	217	231
	7		232, 233, 234	236, 237, 250	
3	4		213	183	
	5	215	214	217	231
	7	234	232, 233	236, 237	250
4	4	213			183
	5		214, 215	217, 231	
	7	234	232, 233	236, 237	250
5	4		213	183	
	5		214, 215	217, 231	
	7	234, 233	232	236	237, 250

2.3.3 Infectious Pressure

For an individual k who is susceptible at time t we define $\Lambda_k(t)$ as the infectious pressure acting upon them at time t, so that

$$\mathbb{P}\big(k \text{ is infected in } (t, t+\delta t] \mid k \text{ is susceptible at time } t\big) = \Lambda_k(t)\delta t + o(\delta t).$$

From the model definition in Section 2.3, $\Lambda_k(t)$ can be expressed as:

$$\Lambda_k(t) = \sum_{j \in \mathcal{N}_{inf}(t)} m(j,t) \times \begin{cases} \frac{\lambda_a}{N-1} + \frac{\lambda_f \delta_f(j,k)}{n-1} + \frac{\lambda_h \delta_c(j,k;t)}{n_{c,f_j}(t)-1} & \text{if } f_j = \text{FTC}, \\ \frac{\lambda_a+\lambda_f}{N-1} + \frac{\lambda_h \delta_c(j,k;t)}{n_{c,f_j}(t)-1} & \text{otherwise}, \end{cases} \tag{2.3.1}$$

where m is the fever/rash identifier:

$$m(j,t) = \begin{cases} b & \text{if } i_j \leq t < r_j, \\ 1 & \text{if } r_j \leq t < \min(\tau_j, q_j), \\ 0 & \text{otherwise,} \end{cases}$$

and $\delta_f(j,k) = 1$ if both k and j are FTC and 0 otherwise, $\delta_c(j,k;t) = 1$ if both k and j live in the same compound at time t and are of the same confession, equalling 0 otherwise. Recall that $N = 31,200$ is the total population size, $n = 120$ is the number of FTC individuals within the population, and $n_{c,f_j}(t)$ is the number of individuals in j's compound of the same faith as j at time t, including j themselves. Finally, $\mathcal{N}_{inf}(t)$ is the set of individuals infective at time t.

From the population diagram in Figure 2.1, there are four different types of susceptible when considering infectious pressure received: categorising over confession (FTC or non) and location (within the compounds or outside). Table 2.5 summarises the contact rates of the different types of susceptible, from all potential types of infector.

Susceptible k	Compound of infective j	Confession of infective j	Pressure from j to k
FTC, compound w	w	FTC	$\frac{\lambda_a}{N-1} + \frac{\lambda_f}{n-1} + \frac{\lambda_h}{n_{c,f_j}(t)-1}$
	w	Non-FTC	$\frac{\lambda_a+\lambda_f}{N-1}$
	w^c or outside	FTC	$\frac{\lambda_a}{N-1} + \frac{\lambda_f}{n-1}$
	w^c or outside	Non-FTC	$\frac{\lambda_a+\lambda_f}{N-1}$
non-FTC, compound w	w	FTC	$\frac{\lambda_a}{N-1}$
	w	Non-FTC	$\frac{\lambda_a+\lambda_f}{N-1} + \frac{\lambda_h}{n_{c,f_j}(t)-1}$
	w^c or outside	FTC	$\frac{\lambda_a}{N-1}$
	w^c or outside	Non-FTC	$\frac{\lambda_a+\lambda_f}{N-1}$
FTC, outside compounds	w	FTC	$\frac{\lambda_a}{N-1} + \frac{\lambda_f}{n-1}$
	w	Non-FTC	$\frac{\lambda_a+\lambda_f}{N-1}$
	Outside	FTC	$\frac{\lambda_a}{N-1} + \frac{\lambda_f}{n-1}$
	Outside	Non-FTC	$\frac{\lambda_a+\lambda_f}{N-1}$
non-FTC, outside compounds	w	FTC	$\frac{\lambda_a}{N-1}$
	w	Non-FTC	$\frac{\lambda_a+\lambda_f}{N-1}$
	Outside	FTC	$\frac{\lambda_a}{N-1}$
	Outside	Non-FTC	$\frac{\lambda_a+\lambda_f}{N-1}$

Table 2.5: Infectious pressure received by susceptible k from infective j. Here, $w \in \{1, ..., 9\}$ is any one of the affected compounds, and w^c denotes any affected compound other than w. In addition, N = size of the population, n = number of FTC individuals in the population (note this change in definition of n for this chapter alone) and $n_{c,f_j}(t)$ = number of individuals in the same compound and of the same faith as individual j at time t. Note: If j is in the fever stage, pressure is multiplied by the infectivity factor b.

2.4 Simulation

2.4.1 Simulation Process

We wish to simulate data from our model for smallpox outbreaks within the population of Abakaliki. This will allow us to perform a simulation study, as well as posterior predictive checking to assess model fit. The method is shown in Algorithm 4. Within this algorithm, the infectious pressure upon each susceptible k at given time t is calculated using the expression for $\Lambda_k(t)$ in Equation (2.3.1).

Algorithm 4 Simulation code for smallpox outbreaks in Abakaliki

---

Input: $\lambda_a, \lambda_f, \lambda_h, v, b, t_q$

Output: **e**, **i**, **r**, **q**, $\boldsymbol{\tau}$

1. Randomly generate protection status for vaccinated individuals (vaccination status known from the data), given vaccine efficacy $v \in (0, 1)$.

Inside compounds: generate statuses individually

for $i = 0$ **to** $n_{com} - 1$ **do**

Generate $U \sim U(0, 1)$

if $U < v$ **and** i is vaccinated **then**

Individual i is protected

end if

end for

Outside compounds: Generate the number of FTC/non-FTC protected individuals

Number FTC and protected $\sim$ Bin(number FTC and vaccinated, v)

Number non-FTC and protected $\sim$ Bin(number non-FTC and vaccinated, v)

2. Set initial infective: $\kappa = 0$.

Generate times for initial infective: e_κ, i_κ, r_κ=0, t_κ and q_κ.

Set time $t = i_\kappa$.

---

3. Loop to simulate infections

while number infectives or exposed > 0 **do**

for $i = 0$ **to** $N - 1$ **do**

inf(i) = sum of pressure from all infectives to susceptible i,

Potential infection time $t_p(i) \sim Exp(\frac{1}{\text{inf(i)}})$,

end for

Take minimum potential time $t_{\text{pm}} = \min(t_p(\mathbf{i}) : \mathbf{i}$ is the set of susceptible individuals), recording index j of the chosen individual.

Take minimum $t_{\text{min}} = \min(t_{\text{pm}}$ and other infection process events (infective moves to rash period, removal etc.)).

if $t_{\text{min}} = t_{\text{pm}}$ **then**

Infection takes place. Generate infection events $e_j = t_{\text{pm}}(j)$, i_j, r_j, t_j and q_j. Set $t = e_j$.

else

Infection does not take place, infection process event does. Set $t =$ time of process event.

end if

end while

To give some more detail on how the calculation of potential exposure times for those outside of the compounds is approached, only one time for FTC individuals outside and one time for non-FTC individuals outside is required to be calculated, since all outside individuals of the same confession receive the same infectious pressure. The mean of the exponentially distributed potential exposure time must be multiplied by $1/(\text{number of FTC individuals outside})$ or $1/(\text{number of non-FTC individuals outside})$ to include the required selection of which individual is to be infected. Since these outside individuals are arbitrarily numbered, apart from being categorised by confession, we define the first outside FTC individual to be infected as number $j = n_{com} + Fp + NFp + 1$, the second $j = n_{com} + Fp + NFp + 2$ and so on, with Fp and NFp defined as in Figure 2.3 as the numbers of FTC and non-FTC protected indi-

viduals, respectively. The case is similar for any non-FTC individuals outside who become infected, with the first being numbered $j = n_{com} + Fp + NFp + Fnp + 1$, for Fnp equal to the number of unprotected FTC individuals outside. This is more computationally efficient since it does not require calculation of thousands of identical infectious pressures.

This mechanism for outbreak simulation will be required in the later simulation study and model checking, but now we proceed to define the Bayesian inference and likelihood expressions to be used in our analysis.

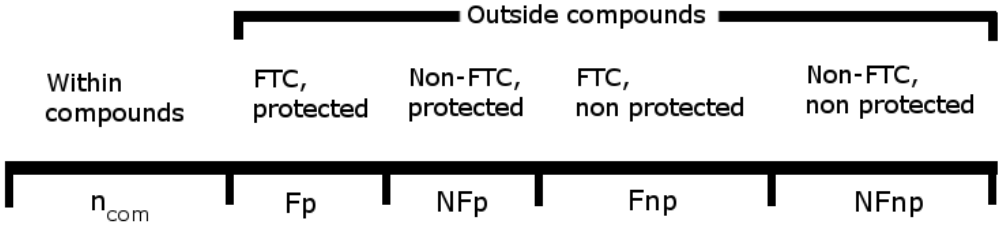


Figure 2.3: Numbering of the N individuals within the population of Abakaliki

2.5 Inference and Likelihood Expressions

With the data introduced and the model defined, we may perform Bayesian inference for the unknown model parameters, given the data and augmenting the unknown event times for each infective as well as the unknown protection and vaccination statuses. This is, to the best of our knowledge, the first Bayesian analysis of the Abakaliki data considering all aspects of the data.

2.5.1 Preliminaries

In order to derive an expression for the likelihood in our model, we first define some notation.

Recall that the vaccination statuses of individuals in Abakaliki are known, with a small number of exceptions as detailed in Table 2.4. For those known only, define $\mathbf{s} = \{s_i \mid i \in 0, .., N-1 \text{ and } s_i \text{ known}\}$ as the set of vaccination statuses of each individual, with s_i taking value 1 for vaccinated i and 0 for unvaccinated i. Then, let

$$\boldsymbol{\Phi} = (\kappa, e_\kappa, t_q, b, v, \lambda_a, \lambda_f, \lambda_h, \boldsymbol{\theta}, \mathbf{s})$$

where

$$\boldsymbol{\theta} = (\mu_I, \sigma_I, \mu_F, \sigma_F, \mu_R, \sigma_R, \mu_Q, \sigma_Q),$$

so that the components of $\boldsymbol{\theta}$, are parameters that are assumed known, and $\boldsymbol{\Phi}$ contains all of the model parameters, both those known and those to be estimated.

Recall that we defined $\mathbf{e}$, $\mathbf{i}$, $\mathbf{q}$ and $\boldsymbol{\tau}$ as the unknown sets of exposure (not including e_κ), infection, quarantine and removal times, respectively. The temporal data $\mathbf{r}$, as introduced in Section 2.2, consist of rash times for all infectives.

For those individuals with unknown vaccination status only, we define $\mathbf{s}^u = \{s_i^u \mid i \in 0, .., N-1 \text{ and } s_i^u \text{ unknown}\}$. The unknown protection status of each individual within the compounds is contained within $\tilde{\mathbf{p}} = (p_0, p_1, ..., p_{n_{com}-1})$; where $p_i = 1$ for successfully vaccinated (protected) or $p_i = 0$ for unsuccessfully vaccinated (unprotected) or not vaccinated at all. For individuals outside the compounds, define $(p_{n_{com}}, ..., p_{N-1})$ in the same way as for those inside. Finally, let $\mathbf{p} = (\tilde{\mathbf{p}}, p_{n_{com}}, ..., p_{N-1})$. For computational purposes, instead of separate protection statuses for each outside individual, we will in fact only require quantities x and y, where

$$x = \text{Number of FTC, vaccinated, unprotected but never infected individuals outside of the compounds,} \quad (2.5.1)$$

$$y = \text{Number of non-FTC, vaccinated, unprotected but never infected individuals outside of the compounds,} \quad (2.5.2)$$

neither of which are known from the data. We separate over FTC membership status since these individuals will have different mixing behaviours. Since the

likelihood contribution from such outside individuals can be described using x and y, this allows for storage of only these two numbers rather than a protection and a vaccination status for each of the 30,949 individuals in question.

Define

$$\gamma = (\mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \mathbf{s}^u, \mathbf{p}),$$

containing protection statuses for all N individuals, and

$$\tilde{\gamma} = (\mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \mathbf{s}^u, \tilde{\mathbf{p}}),$$

containing only the protection statuses for individuals within the compounds.

Next, define sets $\mathcal{N}$ of individuals with sub/superscripts as follows:

$$\begin{aligned}
inf &= \text{Becomes infected,} \\
n-inf &= \text{Never infected,} \\
sus &= \text{Initially susceptible, i.e. unvaccinated or vaccinated but not protected,} \\
FTC &= \text{Member of Faith Tabernacle Church,} \\
n-FTC &= \text{Not a member of Faith Tabernacle Church,} \\
oc &= \text{Outside the compounds,} \\
c &= \text{Inside the compounds.}
\end{aligned}$$

For example, $\mathcal{N}^c_{inf}$ denotes the set of individuals within the compounds who become infected.

Now, for $t \geq e_\kappa$ and $j = 0, ..., N-1$, define

$$\begin{aligned}
\Lambda_j(t) &= \text{infectious pressure acting on individual } j \text{ at time } t, \\
\Lambda(t) &= \text{infectious pressure acting on all individuals who are susceptible at time } t \\
&= \sum_{\substack{j \in \mathcal{N}_{sus} \\ j\,:\,e_j > t}} \Lambda_j(t),
\end{aligned}$$

where $\Lambda(t)$ can be subdivided into $\Lambda(t) = \Lambda_{OC}(t) + \Lambda_{CN}(t) + \Lambda_{CC}(t)$ with each term in the sum representing the overall infection pressure at time t

to those outside the compounds, to those inside the compounds who never become infected, and to those inside who do become infected, respectively. These Λ terms can hence be defined as

$$\begin{aligned}
\Lambda_{CN}(t) &= \sum_{\substack{j \in \mathcal{N}^c_{sus,n-inf} \\ j\,:\,e_j > t}} \Lambda_j(t), \\
\Lambda_{CC}(t) &= \sum_{\substack{j \in \mathcal{N}^c_{sus,inf} \\ j\,:\,e_j > t}} \Lambda_j(t), \\
\Lambda_{OC}(t) &= \sum_{\substack{j \in \mathcal{N}^{oc}_{sus} \\ j\,:\,e_j > t}} \Lambda_j(t).
\end{aligned}$$

With this notation, we denote the likelihood of the data $\mathbf{r}$ given the model parameters $\mathbf{\Phi}$ as $\pi(\mathbf{r} \mid \mathbf{\Phi})$. This is practically intractable in all but trivial cases however, as we do not observe the complete infection process and so it is infeasible to integrate over all possible infectious period parameters as the number of infectives increases. However augmenting the data $\mathbf{r}$ with $\boldsymbol{\gamma}$ we obtain instead the tractable likelihood $\pi(\mathbf{r}, \boldsymbol{\gamma} \mid \mathbf{\Phi})$. An extension of the standard SIR model likelihood defined in Section 1.3.5, this likelihood is given by

$$\begin{aligned}
\pi(\mathbf{r}, \boldsymbol{\gamma} \mid \mathbf{\Phi}) = &\left(\prod_{j \in \mathcal{N}_{inf}} \Lambda_j(e_{j-}) \right) \times \mathrm{e}^{-\int_{e_\kappa}^{T} \Lambda(t)dt} \\
&\times \prod_{j \in \mathcal{N}_{inf}} f_I(i_j - e_j) f_F(r_j - i_j) f_R(\tau_j - r_j) f_Q(q_j - \max(r_j, t_q)) \\
&\times v^{\sum_{r=0}^{N-1} p_r \mathbb{1}_{\{s_r = 1 \text{ or } s^u_r = 1\}}} (1-v)^{\sum_{r=0}^{N-1} (1-p_r) \mathbb{1}_{\{s_r = 1 \text{ or } s^u_r = 1\}}},
\end{aligned} \tag{2.5.3}$$

where T is the end of the epidemic, defined as the first time at which no infectives or exposed individuals remain in the population. In addition, $\Lambda_j(e_{j-}) = \lim_{t \uparrow e_j} \Lambda_j(t)$, where e_{j-} represents the time just before exposure of j. The first product term $\Lambda_j(e_{j-})$ defines the pressure on each infective just before their exposure and the exponential term $\mathrm{e}^{-\int_{e_\kappa}^{T} \Lambda(t)dt}$ represents the pressure on susceptibles over the entirety of the epidemic. Next are the densities $f_A(\cdot)$ of the exposure, fever, rash and time-to-quarantine periods, where f_A, for $A = (I, F,$

R, Q), is defined as the pdf of a $\Gamma(\mu_A, \sigma_A)$ distribution. Lastly, the final term is the likelihood of the protection status arrangement.

We wish to find $\pi(\mathbf{\Phi} \mid \mathbf{r})$, but in order to make use of Equation (2.5.3), we use the augmented posterior density $\pi(\mathbf{\Phi}, \boldsymbol{\gamma} \mid \mathbf{r})$. By Bayes' theorem, see that

$$\begin{aligned} \pi(\mathbf{\Phi}, \boldsymbol{\gamma} \mid \mathbf{r}) &= \frac{\pi(\mathbf{r}, \boldsymbol{\gamma} \mid \mathbf{\Phi})\pi(\mathbf{\Phi})}{\pi(\mathbf{r})} \\ &\propto \pi(\mathbf{r}, \boldsymbol{\gamma} \mid \mathbf{\Phi})\pi(\mathbf{\Phi}), \end{aligned}$$

so that the posterior density is the product of the tractable, augmented likelihood and the prior density $\pi(\mathbf{\Phi})$, divided by a normalising constant.

Assuming independence a priori of the components of $\mathbf{\Phi}$,

$$\begin{aligned} \pi(\mathbf{\Phi}) &= \pi(\kappa, e_\kappa, t_q, b, v, \lambda_a, \lambda_f, \lambda_h, \boldsymbol{\theta}, \mathbf{s}) \\ &= \pi(\kappa)\pi(e_\kappa)\pi(t_q)\pi(b)\pi(v)\pi(\lambda_a)\pi(\lambda_f)\pi(\lambda_h)\pi(\boldsymbol{\theta})\pi(\mathbf{s}). \end{aligned}$$

We assume that λ_a, λ_f and λ_h have gamma distributed priors, v has a uniform prior on $(0, 1)$ and b and t_q have improper, uniform priors on $(0, \infty)$. We set a discrete uniform prior for κ over the number of infected individuals. In addition, e_κ has an improper uniform prior on $(-\infty, i_\kappa)$. Since $\boldsymbol{\theta}$ and $\mathbf{s}$ are assumed known, $\pi(\boldsymbol{\theta})$ and $\pi(\mathbf{s})$ are just point masses.

Before continuing to the likelihood calculations, the notation required in this chapter is summarized in Table 2.6.

2.5.2 Integrating out Parameters x and y

In its current form, our data augmentation scheme results in a likelihood involving the protection status of each of $N = 31,200$ individuals. It is possible, however, to integrate out these parameters for all individuals outside of the compounds. This is essentially because the number of protected individuals is Binomially distributed, and also arises from the fact that individuals outside of the compounds do not contribute compound mixing terms to the likelihood; they may only differ in their FTC membership. Since this removes

Table 2.6: Principal notation.

Parameter	Interpretation
N	Population size
n_{com}	Number of individuals within the compounds
n	Number of FTC individuals
$\mathcal{N}_{inf}$	Set of ever infected individuals
$\mathcal{N}_b^a$	Set of individuals with location a and status b
	(such as within the compounds and ever infected)
λ_a	Global infection rate
λ_f	FTC infection rate
λ_h	Household infection rate
b	Infectivity factor for fever period
v	Vaccine efficacy
t_q	Time quarantine measures introduced
$\boldsymbol{\theta}$	Fixed parameters for disease stage lengths
κ	Identity of initial infective
e_κ	Exposure time of initial infective
$\mathbf{s}$	Vector of vaccination statuses (for all individuals)
$\mathbf{s}^u$	Vector of unknown vaccination statuses
$\mathbf{p}$	Vector of protection statuses (for all individuals)
$\tilde{\mathbf{p}}$	Vector of protection statuses (compound individuals only)
$\boldsymbol{\Phi}$	$(\kappa, e_\kappa, t_q, b, v, \lambda_a, \lambda_f, \lambda_h, \boldsymbol{\theta}, \mathbf{s})$
$\mathbf{e}$	Vector of exposure times
$\mathbf{i}$	Vector of fever times
$\mathbf{r}$	Vector of rash times
$\boldsymbol{\tau}$	Vector of removal times
$\mathbf{q}$	Vector of quarantine times
γ	$(\mathbf{e}, \mathbf{i}, \boldsymbol{\tau}, \mathbf{q}, \mathbf{s}^u, \mathbf{p})$
$\tilde{\gamma}$	$(\mathbf{e}, \mathbf{i}, \boldsymbol{\tau}, \mathbf{q}, \mathbf{s}^u, \tilde{\mathbf{p}})$

almost 31,000 parameters, this provides a likelihood which is much faster to compute.

In this section we will show that

$$\pi(\mathbf{r}, \boldsymbol{\gamma} \mid \boldsymbol{\Phi}) = \pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}}, x, y \mid \boldsymbol{\Phi}),$$

where x and y are defined as in Equations (2.5.1) and (2.5.2). This first step shows that the separate protection statuses for individuals outside the compounds can be summarised by just the total numbers of FTC and non-FTC vaccinated, unprotected but never infected individuals. We will then integrate out x and y as follows. We begin by expressing the likelihood as

$$\pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}}, x, y \mid \boldsymbol{\Phi}) = \pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}} \mid x, y, \boldsymbol{\Phi})\pi(x, y \mid \boldsymbol{\Phi}) \tag{2.5.4}$$

and hence, by Bayes' Theorem,

$$\pi(\boldsymbol{\Phi}, \tilde{\boldsymbol{\gamma}}, x, y \mid \mathbf{r}) = \frac{\pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}} \mid x, y, \boldsymbol{\Phi})\pi(x, y \mid \boldsymbol{\Phi})\pi(\boldsymbol{\Phi})}{\pi(\mathbf{r})}.$$

Integrating out x and y, equivalent to summing in this case since they take discrete values, we find

$$\sum_{x,y} \pi(\boldsymbol{\Phi}, \tilde{\boldsymbol{\gamma}}, x, y \mid \mathbf{r}) = \frac{\pi(\boldsymbol{\Phi})}{\pi(\mathbf{r})} \sum_{x,y} \pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}} \mid x, y, \boldsymbol{\Phi})\pi(x, y \mid \boldsymbol{\Phi}), \tag{2.5.5}$$

implying that

$$\pi(\boldsymbol{\Phi}, \tilde{\boldsymbol{\gamma}} \mid \mathbf{r}) = \frac{\pi(\boldsymbol{\Phi})}{\pi(\mathbf{r})} \pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}} \mid \boldsymbol{\Phi}).$$

This gives new target density $\pi(\boldsymbol{\Phi}, \tilde{\boldsymbol{\gamma}} \mid \mathbf{r})$, independent of x and y as desired.

2.5.2.1 Removing protection statuses for individuals outside the compounds

We now must prove that

$$\pi(\mathbf{r}, \boldsymbol{\gamma} \mid \boldsymbol{\Phi}) = \pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}}, \mathrm{x}, \mathrm{y} \mid \boldsymbol{\Phi}),$$

in order to perform the integration detailed above.

Recall that

$$\begin{aligned}\pi(\mathbf{r},\boldsymbol{\gamma}\mid\boldsymbol{\Phi}) = &\prod_{j\in\mathcal{N}_{inf}}\Lambda_j(e_j-)\times \mathrm{e}^{-\int_{e_K}^{T}\Lambda(t;x,y)dt}\\ &\times\prod_{j\in\mathcal{N}_{inf}} f_I(i_j-e_j)f_F(r_j-i_j)f_R(\tau_j-r_j)f_Q(q_j-\max(r_j,t_q))\\ &\times v^{\sum_{r=0}^{N-1} p_r\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}(1-v)^{\sum_{r=0}^{N-1}(1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}},\end{aligned} \tag{2.5.6}$$

where we may write $\Lambda(t;x,y) = \Lambda_{OC}(t;x,y)+\Lambda_{CN}(t)+\Lambda_{CC}(t)$ as defined in Section 2.5.1 but now with dependence on x and y. Then the only terms in Equation (2.5.6) involving x and y are $\Lambda_{OC}(t;x,y)$ and $v^{\sum_{r=0}^{N-1}(p_r\mid\{s_r=1 \text{ or } s_r^u=1\})}(1-v)^{\sum_{r=0}^{N-1}(1-p_r\mid\{s_r=1 \text{ or } s_r^u=1\})}$, since x and y are determined by the protection status of individuals outside of the compounds. It is possible to subdivide the likelihood of the protection status over whether the individual is inside/outside and by whether they do or do not become infected as follows:

$$\begin{aligned}&v^{\sum_{r=0}^{N-1} p_r\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}(1-v)^{\sum_{r=0}^{N-1}(1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}\\ &\quad= v^{\sum_{r=0}^{n_{com}-1} p_r\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}(1-v)^{\sum_{r=0}^{n_{com}-1}(1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}\\ &\qquad\times v^{\sum_{r=n_{com}}^{N-1} p_r\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}(1-v)^{\sum_{r=n_{com}}^{N-1}(1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}\\ &\quad= v^{\sum_{r=0}^{n_{com}-1} p_r\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}(1-v)^{\sum_{r=0}^{n_{com}-1}(1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}\\ &\qquad\times v^{\sum_{r=n_{com}}^{N-1} p_r\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}(1-v)^{\sum_{r=n_{com}}^{N-1}(1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}_{r\in\mathcal{N}_{inf}}\\ &\qquad\times(1-v)^{\sum_{r=n_{com}}^{N-1}(1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}}}_{r\in\mathcal{N}_{n-inf}}.\end{aligned}$$

The terms corresponding to individuals inside the compounds do not depend on x and y, and neither does the term concerning outside infectives, since their protection status is known. Also, there are only unknown vaccination statuses for individuals inside of the compounds and so we may disregard the s_r^u terms.

Let us then define the expression dependent on x and y as

$$L_{xy} = \mathrm{e}^{-\int_{e_\kappa}^{T} \Lambda_{OC}(t;x,y)dt} v^{\sum_{r=n_{com}}^{N-1} p_r s_r} (1-v)^{\sum_{r=n_{com}, r\in\mathcal{N}_{n-inf}}^{N-1} (1-p_r)s_r}, \tag{2.5.7}$$

where, with s_r^u disregarded, the indicator functions collapse to just s_r as it is itself an indicator function, with value 1 for r vaccinated and 0 for r non-vaccinated.

The integral of Λ_{OC} is equal to the sum over all infectives j of the pressure from j to any given FTC or non-FTC individual outside of the compounds (throughout all time (e_κ, T)), summed over all the initially susceptible FTC and non-FTC individuals. Thus,

$$\mathrm{e}^{-\int_{e_\kappa}^{T} \Lambda_{OC}(t;x,y)dt} = \exp\Big(-\sum_{j\in\mathcal{N}_{inf}} \sum_{k\in\mathcal{N}_{sus}^{oc}} \Psi_{jk}\Big) \tag{2.5.8}$$

with Ψ_{jk} = total infectious pressure from j to susceptible k during the time interval (e_κ, T).

We wish to express Equation (2.5.8) in terms of x and y and so, partitioning according to whether susceptibles become infected or not, see that

$$\begin{aligned}
\exp\Big(-\sum_{j\in\mathcal{N}_{inf}} \sum_{k\in\mathcal{N}_{sus}^{oc}} \Psi_{jk}\Big) &= \exp\left(-\sum_{j\in\mathcal{N}_{inf}} \Big(\sum_{k\in\mathcal{N}_{inf}^{oc}} \Psi_{jk} + \sum_{k\in\mathcal{N}_{n-inf}^{oc}} \Psi_{jk}\Big)\right) \\
&= \exp\left(-\sum_{j\in\mathcal{N}_{inf}} \Big(\sum_{k\in\mathcal{N}_{inf,FTC}^{oc}} \Psi_{jk} + \sum_{k\in\mathcal{N}_{inf,n-FTC}^{oc}} \Psi_{jk} \right. \right. \\
&\qquad \left. \left. + \sum_{k\in\mathcal{N}_{n-inf,FTC}^{oc}} \Psi_{jk} + \sum_{k\in\mathcal{N}_{n-inf,n-FTC}^{oc}} \Psi_{jk}\Big)\right),
\end{aligned} \tag{2.5.9}$$

which has also been partitioned over the confession of the susceptibles: whether FTC or not.

However, the data only indicate if individuals are vaccinated, not whether they are protected, since the vaccine may not have been effective. Of the four Ψ_{jk} values in Equation (2.5.9), the numbers of initially susceptible FTC and non-FTC individuals outside that do become infective are known, but not the

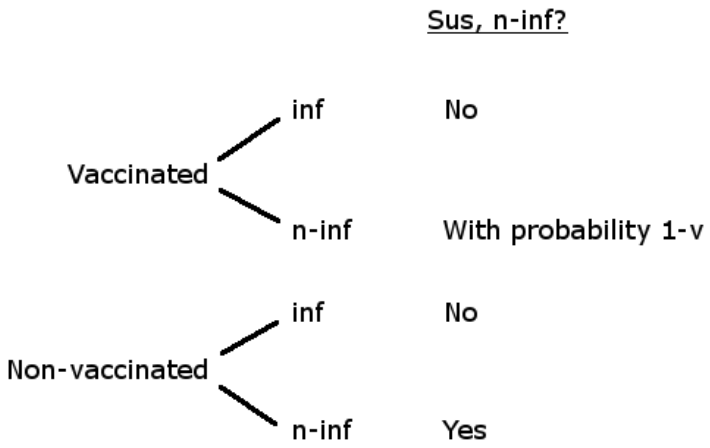


Figure 2.4: Tree diagram for protection status of individuals outside the compounds. Here inf = infected, n-inf = non-infected and sus= susceptible.

number of initial susceptibles that are never infected. Figure 2.4 displays the possible combinations of individuals in the latter two categories.

So, the total number of initially susceptible, never infected individuals outside the compounds is given by

$$|\mathcal{N}^{oc}_{n-inf}| = a^{oc}_{n-inf} + \text{Bin}(b^{oc}_{n-inf}, 1-v),$$

where a^{oc}_{n-inf} is the known number of non-vaccinated, never infective individuals outside the compounds and b^{oc}_{n-inf} is the known number of vaccinated, never infective individuals outside, each one of whom is susceptible with probability $1-v$. $\text{Bin}(n,p)$ represents a binomial distribution with number of trials n and success probability p. The number of b^{oc}_{n-inf} that are susceptible is equal to $x+y$, specifically

$$x \sim \text{Bin}(b^{oc}_{n-inf,FTC}, 1-v),$$
$$y \sim \text{Bin}(b^{oc}_{n-inf,n-FTC}, 1-v),$$

given v, where $b^{oc}_{n-inf,FTC}$ is the number of FTC individuals outside who are vaccinated but not infected and $b^{oc}_{n-inf,n-FTC}$ is the number of non-FTC individuals outside who are vaccinated but not infected.

Therefore Equation (2.5.7) for L_{xy} becomes

$$\begin{aligned} L_{xy} = \quad & \exp\Big(-\sum_{j\in\mathcal{N}_{inf}}\Big(\sum_{k\in\mathcal{N}^{oc}_{inf,FTC}}\Psi_{jk} + \sum_{k\in\mathcal{N}^{oc}_{inf,n-FTC}}\Psi_{jk} \\ & +\chi_F(j)(a^{oc}_{n-inf,FTC}+x)+\chi_{NF}(j)(a^{oc}_{n-inf,n-FTC}+y)\Big)\Big) \\ & \times v^{\sum_{r=n_{com}}^{N-1} p_r s_r}(1-v)^{\sum_{r\in\mathcal{N}_{n-inf}}^{N-1}(1-p_r)s_r}, \end{aligned} \tag{2.5.10}$$

where

$$\chi_F(j) = \big(b(r_j - i_j) + (\min(q_j,\tau_j) - r_j)\big) \times \begin{cases} \frac{\lambda_a}{N-1} + \frac{\lambda_f}{n-1} & \text{if } f_j = \text{FTC} \\ \frac{\lambda_a+\lambda_f}{N-1} & \text{otherwise} \end{cases}$$

and

$$\chi_{NF}(j) = \big(b(r_j - i_j) + (\min(q_j,\tau_j) - r_j)\big) \times \begin{cases} \frac{\lambda_a}{N-1} & \text{if } f_j = \text{FTC} \\ \frac{\lambda_a+\lambda_f}{N-1} & \text{otherwise} \end{cases}$$

represent the contribution from infective j to a never infected FTC/non-FTC susceptible outside the compounds over all time (e_κ, T). This contribution is equal for all susceptibles k.

Considering the protection status likelihood parts of Equation (2.5.10), note that

$$(1-v)^{\sum_{r\in\mathcal{N}_{n-inf}}^{N-1}(1-p_r)s_r} = (1-v)^{x+y},$$

since the sum is equal to the number of vaccinated but unprotected individuals outside who do not become infected. Hence

$$v^{\sum_{r=n_{com}}^{N-1} p_r s_r} = v^{b^{oc}_{n-inf}-x-y},$$

as the sum is equal to the number of vaccinated and protected individuals outside, which can be seen as equivalent to the number of vaccinated never infected individuals outside minus those who are unprotected.

Splitting up the terms involving x and y, Equation (2.5.10) can be written in the form

$$\begin{aligned} L_{xy} = \exp\Bigg(- \sum_{j \in \mathcal{N}_{inf}} \Bigg(\sum_{k \in \mathcal{N}^{oc}_{inf,FTC}} \Psi_{jk} + \sum_{k \in \mathcal{N}^{oc}_{inf,n-FTC}} \Psi_{jk} \\ + a^{oc}_{n-inf,FTC}\chi_F(j) + a^{oc}_{n-inf,n-FTC}\chi_{NF}(j) \Bigg) \Bigg) \\ \times \exp\left(- \sum_{j \in \mathcal{N}_{inf}} \chi_F(j) \times x \right) \times (1-v)^x v^{b^{oc}_{n-inf,FTC} - x} \\ \times \exp\left(- \sum_{j \in \mathcal{N}_{inf}} \chi_{NF}(j) \times y \right) \times (1-v)^y v^{b^{oc}_{n-inf,n-FTC} - y}, \quad (2.5.11) \end{aligned}$$

which demonstrates that $\pi(\mathbf{r}, \boldsymbol{\gamma} \mid \boldsymbol{\Phi}) = \pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}}, \mathrm{x}, \mathrm{y} \mid \boldsymbol{\Phi})$ as claimed. None of the unknown protection statuses for outside individuals are now explicitly required, resulting in improved computational speed, and we may now sum this expression over x and y to obtain the overall likelihood $\pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}} \mid \boldsymbol{\Phi})$, which is faster still to compute.

2.5.2.2 Sum over x and y

Considering Equation (2.5.11), the first exponential term does not depend upon x or y and so we may disregard it for now. For the rest of the expression, recognise that the sum takes the form of a moment generating function (MGF) for the binomial distribution and hence use the fact that, for $W \sim \mathrm{Bin}(n, p)$,

$$\mathbb{E}(\mathrm{e}^{-\theta W}) = (p\mathrm{e}^{-\theta} + (1-p))^n, \ \theta \geq 0.$$

Setting $\sum_{j \in \mathcal{N}_{inf}} \chi_F(j) = \chi_F$ and $\sum_{j \in \mathcal{N}_{inf}} \chi_{NF}(j) = \chi_{NF}$ as θ, as well as $b^{oc}_{n-inf,FTC}$ and $b^{oc}_{n-inf,n-FTC}$ as n and $1-v$ as p we obtain

$$\begin{aligned} L = \exp\Big(- \sum_{j \in \mathcal{N}_{inf}} \Big(\sum_{k \in \mathcal{N}^{oc}_{inf,FTC}} \Psi_{jk} + \sum_{k \in \mathcal{N}^{oc}_{inf,n-FTC}} \Psi_{jk} \Big) \\ - a^{oc}_{n-inf,FTC}\chi_F - a^{oc}_{n-inf,n-FTC}\chi_{NF} \Big) \\ \times \Big((1-v)\mathrm{e}^{-\chi_F} + v \Big)^{b^{oc}_{n-inf,FTC}} \Big((1-v)\mathrm{e}^{-\chi_{NF}} + v \Big)^{b^{oc}_{n-inf,n-FTC}}, \end{aligned}$$

for which taking logs yields

$$\begin{aligned}\log(L) = &- \sum_{j\in\mathcal{N}_{inf}} \Big(\sum_{k\in\mathcal{N}^{oc}_{inf}} \Psi_{jk} \Big) - a^{oc}_{n-inf,FTC}\chi_F - a^{oc}_{n-inf,n-FTC}\chi_{NF} \\ &+ b^{oc}_{n-inf,FTC} \log\Big(v + (1-v)\mathrm{e}^{-\chi_F}\Big) \\ &+ b^{oc}_{n-inf,n-FTC} \log\Big(v + (1-v)\mathrm{e}^{-\chi_{NF}}\Big). \end{aligned} \tag{2.5.12}$$

2.5.3 Likelihood

To obtain the full, tractable likelihood expression, combine the section from Equation (2.5.12) with the remaining parts of the original likelihood from Equation (2.5.6), resulting in an overall log likelihood of

$$\begin{aligned}\log(\pi(\mathbf{r}, \tilde{\boldsymbol{\gamma}} \mid \boldsymbol{\Phi})) \;=\; & \\ & \log\Big(\prod_{\substack{j\in\mathcal{N}_{inf}\\ j\neq\kappa}} \Lambda_j(e_j-)\Big) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) - \Lambda_{CC}(t)dt \\ & - \sum_{j\in\mathcal{N}_{inf}} \Big(\sum_{k\in\mathcal{N}^{oc}_{inf}} \Psi_{jk}\Big) - a^{oc}_{n-inf,FTC}\chi_F - a^{oc}_{n-inf,n-FTC}\chi_{NF} \\ & + b^{oc}_{n-inf,FTC} \log\big(v + (1-v)\mathrm{e}^{-\chi_F}\big) \\ & + b^{oc}_{n-inf,n-FTC} \log\big(v + (1-v)\mathrm{e}^{-\chi_{NF}}\big) \\ & + \log\Big(\prod_{j\in\mathcal{N}_{inf}} f_I(i_j - e_j) f_F(r_j - i_j) f_R(\tau_j - r_j) f_Q(q_j - \max(r_j, t_q))\Big) \\ & + \sum_{r=0}^{n_{com}-1} p_r \mathbb{1}_{\{s_r=1 \text{ or } s^u_r=1\}} \log(v) \\ & + \sum_{r=0}^{n_{com}-1} (1-p_r) \mathbb{1}_{\{s_r=1 \text{ or } s^u_r=1\}} \log(1-v) \\ & + \sum_{\substack{r=n_{com}\\ r\in\mathcal{N}_{inf}}}^{N-1} (1-p_r) s_r \log(1-v), \end{aligned}$$

where

$$\prod_{\substack{j\in\mathcal{N}_{inf}\\ j\neq\kappa}}\Lambda_j(e_j-) = \prod_{\substack{j\in\mathcal{N}_{inf}\\ j\neq\kappa}}\sum_{\substack{i\in\mathcal{N}_{inf}\\ i\neq j}} m(i,e_j-) \times \begin{cases} \frac{\lambda_a}{N-1}+\frac{\lambda_f\delta_f(i,j)}{n-1}+\frac{\lambda_h\delta_c(i,j;e_j-)}{n_{c,f_i}(e_j-)-1} & \text{if } f_i=\text{FTC},\\ \frac{\lambda_a+\lambda_f}{N-1}+\frac{\lambda_h\delta_c(i,j;e_j-)}{n_{c,f_i}(e_j-)-1} & \text{otherwise,}\end{cases}$$

with

$$m(i,e_j-)=\begin{cases} b & \text{if } i_i<e_j-<r_i\\ 1 & \text{if } r_i<e_j-<\min(\tau_i,q_i)\\ 0 & \text{otherwise.}\end{cases}$$

The log posterior density of interest is thus

$$\begin{aligned}\log(\pi(\boldsymbol{\Phi},\tilde{\boldsymbol{\gamma}}\mid\mathbf{r},\boldsymbol{\theta})) &\propto \log(\pi(\mathbf{r},\tilde{\boldsymbol{\gamma}}\mid\boldsymbol{\Phi})\pi(\boldsymbol{\Phi}))\\ &= \log(\pi(\mathbf{r},\tilde{\boldsymbol{\gamma}}\mid\boldsymbol{\Phi}))+\log(\pi(\kappa))+\log(\pi(e_\kappa))+\log(\pi(t_q))\\ &\quad+\log(\pi(b))+\log(\pi(v))+\log(\pi(\lambda_a))\\ &\quad+\log(\pi(\lambda_f))+\log(\pi(\lambda_h)).\end{aligned}$$

We have obtained a tractable likelihood expression which is sufficiently fast to compute, and we may now perform MCMC; sampling from the posterior density to obtain estimates of model parameters.

2.6 MCMC

In this section we will detail the MCMC algorithm used to update the model parameters and the augmented data.

Within the MCMC algorithm, all of the 12 parameters are updated singly in a systematic order using Metropolis-Hastings updates, with the exception of the exposure, infection, quarantine and removal times which are updated in pairs. More complex updates could be considered, but since these single updates work well and lead to sufficient mixing, it does not appear necessary.

Proposed values outside of the possible range for each parameter are immediately rejected (for example, all the infection rates must be positive). The acceptance probability is calculated using the full conditional distributions, as fully detailed in Appendix A, since calculation of the full likelihood is computationally demanding.

We obtain 100,000 samples in all cases, taken after an initial burn-in of 10,000 iterations and thinning to record every 10th iteration for sufficient independence in these samples. In each loop of the algorithm, first the λ values are individually updated, followed by v (vaccine efficacy), b (infectivity factor of fever infectious period) and t_q (days until isolation procedures begin). For a number of randomly selected individuals, the pair of exposure and fever times, followed by the pair of quarantine and removal times, are then updated. This is followed by the protection status of a small number of individuals within the compounds proposed to be changed. Finally, vaccination statuses $\mathbf{s}^u$ must be updated. Whereas to update $\tilde{\mathbf{p}}$ we randomly select any element of $\tilde{\mathbf{p}}$ and propose a change to it, with $\mathbf{s}^u$ we consider the limited possible number of unknown vaccination configurations for several individuals within the compounds, as shown in Table 2.4. The total number of vaccinated people in each compound is known, but not necessarily whether those vaccinated are FTC or non-FTC, and so to update $\mathbf{s}^u$ we randomly select one of the five potential vaccination configurations from $\mathbf{c} = \{c_i \mid i = 1, ..., 5\}$.

The MCMC was coded in C, and takes on average 2.5 hours for 100,000 samples to be obtained. A summary of the MCMC process used is given in Algorithm 5. The number of event times and protection statuses updated per iteration in the algorithm (5 and 15, respectively) were tuned to provide good mixing.

We now explain each of the MCMC updates in more detail. To update any parameter α, a candidate value is drawn from the proposal density. Any positive candidate $\tilde{\alpha}$ value is accepted with probability

$$\frac{\pi(\tilde{\alpha})q(\alpha|\tilde{\alpha})}{\pi(\alpha)q(\tilde{\alpha}|\alpha)} \wedge 1,$$

Algorithm 5 MCMC process for obtaining samples of the Abakaliki outbreak parameters.

Input: **r**, it = number of iterations, b = length of burn-in, th = degree of thinning

Output: $\lambda_a, \lambda_f, \lambda_h, v, b, t_q$

1. Establish initial values of output parameters as well as event times, protection statuses and unknown vaccination statuses.

2.

for $i = -b$ **to** it **do**

for $inner = 0$ **to** th **do**

a) Update λ_a

b) Update λ_h

c) Update λ_f

d) Update v

e) Update b

f) Update t_q

g)

for $j = 1$ **to** 5 **do**

Randomly select an infected individual

Update their exposure and infection times as a pair

Update their quarantine and removal times as a pair

end for

h)

for $j = 1$ **to** 15 **do**

Randomly select an individual within the compounds

Update their protection status

end for

i) Update the unknown vaccination status configuration, and corresponding protection statuses

end for

Record current output parameter values

end for

where $q(\cdot|\cdot)$ denotes the proposal density/mass function and $\pi(\cdot)$ is the α-dependent full conditional distribution, provided in Appendix A. Considering each parameter individually:

λ values updated - infection rates: Candidate value

$$\tilde{\lambda}_a = \lambda_a + x$$

is proposed, where $x \sim N(0, \sigma^2_{\lambda_a})$, i.e. x is a Gaussian distributed random variable with mean 0 and fixed variance $\sigma^2_{\lambda_a}$. In practice, $\sigma^2_{\lambda_a}$ is tuned to provide reasonable mixing of the Markov chain. The probability of acceptance is given by:

$$\frac{\pi(\tilde{\lambda}_a \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, b, \lambda_f, \lambda_h, \mathbf{s})}{\pi(\lambda_a \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, b, \lambda_f, \lambda_h, \mathbf{s})} \wedge 1,$$

with similar expressions for λ_f and λ_h.

Using a randomly generated, uniform distributed number U between 0 and 1, the proposed value is accepted or rejected under the given probability. Since we operate on a log scale, the candidate is accepted if

$$\begin{aligned} \log(U) \;<\; & \log(\pi(\tilde{\lambda}_a \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, b, \lambda_f, \lambda_h, \mathbf{s})) \\ & - \log(\pi(\lambda_a \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, b, \lambda_f, \lambda_h, \mathbf{s})). \end{aligned}$$

This process is carried out for λ_a, λ_h and λ_f in turn.

v value updated - vaccine efficacy: A very similar procedure is used to update v, with a candidate value suggested as a random Gaussian distributed variable added to the current v value. Provided this candidate lies between 0 and 1, it is accepted with probability

$$\frac{\pi(\tilde{v} \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, b, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})}{\pi(v \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, b, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})} \wedge 1.$$

b value updated - infectivity factor: Next, b is updated using the same candidate selection process as for v. In this case, the probability of acceptance is

$$\frac{\pi(\tilde{b} \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})}{\pi(b \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})} \wedge 1.$$

t_q value updated - time quarantine procedures introduced: In much the same way as before, a new value for t_q is proposed, based on a random Gaussian distributed amount added to its previous value, and accepted with probability

$$\frac{\pi(\tilde{t}_q \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, b, v, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})}{\pi(t_q \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, b, v, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})} \wedge 1.$$

e, i, q, t values updated - exposure, infection, quarantine and removal times: A number, typically in the range 5-10, of individuals per iteration of the algorithm are randomly selected to have their exposure, infection, quarantine and removal values updated. Candidate values are selected, with exposure and infection being accepted/rejected as a unit followed, by the pair of quarantine and removal since we expect these quantities to be correlated.

Candidates for the period of time between start of infectivity and known rash start time, then exposure and start of infectivity are proposed as random gamma distributed values. They are accepted/rejected as a unit, before candidate values for the time between rash and quarantine and also rash and removal are proposed and similarly judged as a unit. The process for each individual is as follows

1. Select an individual j uniformly at random from the set $\mathcal{N}_{inf}$,
2. Simulate F $\backsim \Gamma(\mu_F, \sigma_F)$ and set $\tilde{i}_i = r_i - F$,
3. Simulate I $\backsim \Gamma(\mu_I, \sigma_I)$ and set $\tilde{e}_i = \tilde{i}_i - I$,
4. Simulate R $\backsim \Gamma(\mu_R, \sigma_R)$ and set $\tilde{\tau}_i = r_i + R$,
5. Simulate X $\backsim \Gamma(\mu_Q, \sigma_Q)$ and set $\tilde{q}_i = \max(r_i, t_q) + X$.

Making sure to keep track of the initial infective, the candidates for e_i and i_i are accepted with probability

$$\frac{\pi(\tilde{e}_i, \tilde{i}_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}_{-i}, \mathbf{i}_{-i}, \mathbf{q}, \boldsymbol{\tau}, \tilde{\mathbf{p}}, \mathbf{s}^u)}{\pi(e_i, i_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}_{-i}, \mathbf{i}_{-i}, \mathbf{q}, \boldsymbol{\tau}, \tilde{\mathbf{p}}, \mathbf{s}^u)} \wedge 1.$$

After this, and regardless of the acceptance of the exposure and infection times, candidate times for τ_i and q_i are considered similarly. The process is repeated for each of the selected infective individuals.

Protection status updated: Similarly, a number of people in the compounds, again typically 5-10, are randomly selected to have their protection status changed. The change for each is accepted with probability

$$\frac{\pi(\tilde{p}_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \mathbf{p}_{-i}, \mathbf{s}^u)}{\pi(p_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \mathbf{p}_{-i}, \mathbf{s}^u)} \wedge 1.$$

Only those individuals who are vaccinated and not infected are eligible to have their protection status changed. These updates are accepted/rejected separately rather than as a unit.

Unknown vaccination status individuals updated: Finally, there are those people within the compound whose vaccination statuses are unknown. A new combination of these is proposed, by selection of one of the aforementioned c values from 1 through 5 which represent the five potential configurations of unknown vaccination status. Protection statuses of those affected are then updated accordingly; set to 0 if an individual is proposed as non-vaccinated, and set to 1 independently with probability v if an individual is proposed vaccinated.

Selecting one of five possible configurations of unknown vaccination status uniformly at random, any given one is accepted with probability

$$\frac{\pi(\tilde{c}_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \tilde{\mathbf{p}})}{\pi(c_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \tilde{\mathbf{p}})} \wedge 1.$$

2.7 Results

2.7.1 Abakaliki Data

With the full likelihood expressions obtained and MCMC scheme defined, we may now analyse the Abakaliki data. We seek to compare the results from our MCMC to those of Eichner and Dietz, and both can be found in Table 2.7. The posterior means, medians and credible intervals from MCMC are given

alongside the results of Eichner and Dietz. Figure 2.5 also contains density plots for the parameters of interest as well as the basic reproduction number, compared to Eichner and Dietz' maximum likelihood estimates (MLEs). In general, the results from MCMC appear very similar to Eichner and Dietz', particularly for the six model parameters. This is perhaps unexpected, and indicates that Eichner and Dietz's method may be fairly accurate despite its approximations.

Our mean estimate of the basic reproduction number $R_0 = (\mu_R + b\mu_F)(\lambda_a + \lambda_f + \lambda_h)$ is 7.96 for the whole infectious period, which means that in an entirely susceptible population an infected person will on average infect 7.96 others. This is slightly higher than Eichner and Dietz's estimate of 6.87. Similarly, our estimate of the reproduction number for the fever period $R_F = b\mu_F(\lambda_a + \lambda_f + \lambda_h)$ is 0.53 compared to Eichner and Dietz's 0.164. In this case the difference can be explained by our larger estimate for infectivity factor b, which has a highly skewed posterior density.

Table 2.8 gives estimates of a selection of reproduction numbers. Figure 2.6 contains density plots for these reproduction numbers. Defining R_Q as the reproduction number once quarantine measures are in place (i.e. R_0 with $\mu_R = \mu_Q = 2.0$) it is estimated at a mean of 1.459, interestingly meaning the epidemic is still super-critical. Defining pairs of pre- and post- quarantine measure reproduction numbers for spread only within compounds, for between FTC individuals and for in the wider population (i.e. $R_{Qa} = (\mu_Q + b\mu_F)\lambda_a$ and so on), we can see the impact of these different types of transmission. As we would expect, all reproduction numbers are greatly lowered post-quarantine compared to their pre-quarantine counterparts. However, we also see that (i) within compounds, the epidemic is super-critical both before and after t_q; (ii) within the FTC membership, the epidemic changes from super- to sub-critical and (iii) in the wider population, the epidemic is always sub-critical. This would imply that what stops the outbreak spreading further is a combination of a depletion of susceptibles in the compounds and the fact that the global

	Posterior mean	Posterior median	Credible interval	Eichner and Dietz MLE	Eichner and Dietz confidence interval
λ_a	0.041	0.035	(0, 0.093)	0.0281	(0.00447, 0.101)
λ_f	0.063	0.059	(0.009, 0.010)	0.0562	(0.0187, 0.127)
λ_h	0.358	0.349	(0.150, 0.565)	0.335	(0.192, 0.527)
v	0.808	0.817	(0.668, 0.947)	0.816	(0.644, 0.922)
b	0.522	0.374	(0.0, 1.500)	0.157	(0, 1.89)
t_q	50.4	50.2	(42.4, 58.3)	51.5	(44.7, 59.6)
R_0	7.96	7.79	(4.33, 11.59)	6.87	(4.52, 10.1)
R_F	0.531	0.431	(0.0, 1.364)	0.164	(0.0, 1.31)

Table 2.7: Parameter estimates and equal-tailed 95% credible intervals for the Abakaliki smallpox outbreak from the true likelihood approach, alongside the results of Eichner and Dietz (2003) for comparison. 100,000 MCMC samples were obtained. $R_0 = (\mu_R + b\mu_F)(\lambda_a + \lambda_f + \lambda_h)$ and $R_F = b\mu_F(\lambda_a + \lambda_f + \lambda_h)$.

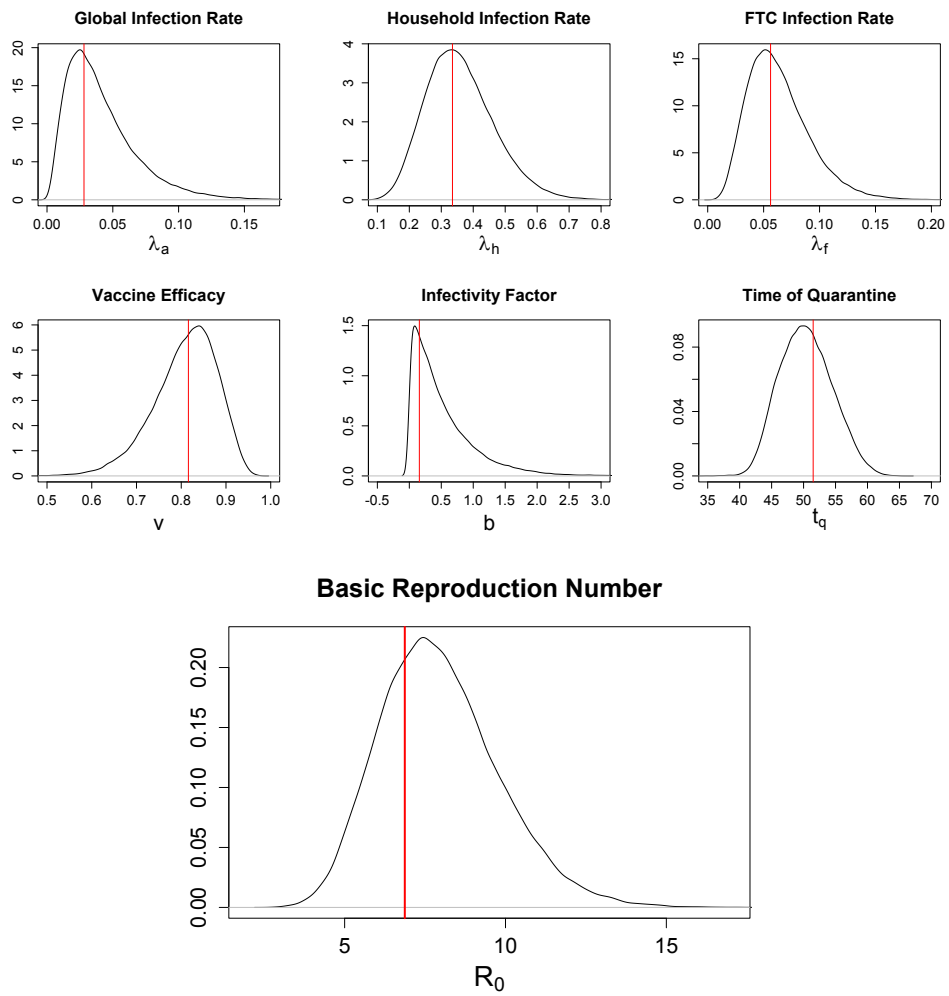


Figure 2.5: Posterior densities of the six parameters of interest and the basic reproduction number, from the Abakaliki outbreak data. Red lines represent Eichner and Dietz' MLEs. Shown are 100,000 samples from an MCMC run.

spread is sub-critical, rather than the quarantine procedure itself.

We also consider the correlation between the model parameters. Figure 2.7 displays this, with a scatter plot and Pearson's correlation coefficient for each pair of parameters. The lack of correlation seen in the plot suggests that the six basic model parameters can indeed be individually estimated from the data, and our model is not over-parameterised.

We lastly consider the posterior distribution of the exposure times for each infected individual, by taking the estimated exposure time for each infective at

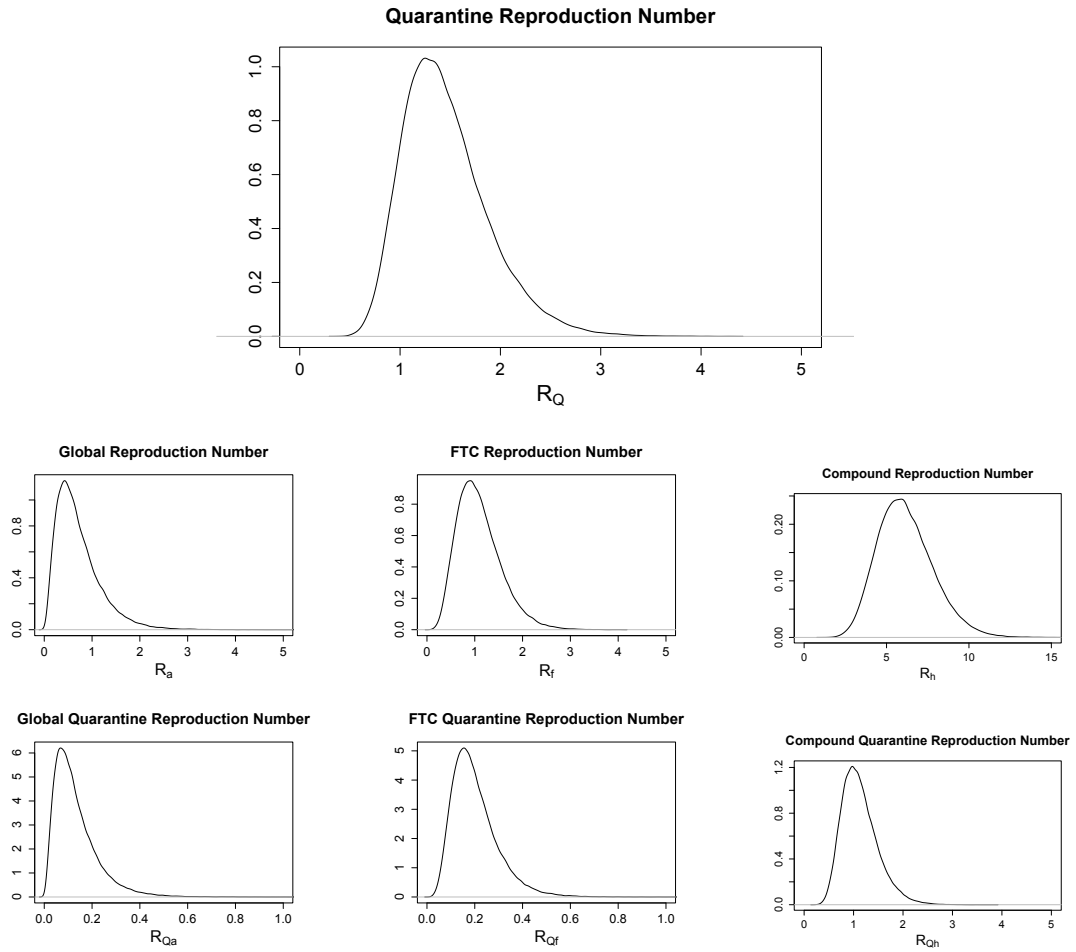


Figure 2.6: Posterior densities of the reproduction numbers contained in Table 2.8, from the Abakaliki outbreak data. Shown are 100,000 samples from an MCMC run.

each iteration of the MCMC algorithm. These are shown in a heat map in Figure 2.8. We see that generally there is small uncertainty in the exposure times, most of them following the same ordering as the rash times. This is likely due to the small variances assumed for the disease stage lengths. This plot also allows us to consider temporal features of the outbreak, such as the generations of infectives. We see two easily discernible generations at the start of the outbreak (largely corresponding to those the initial infective infects, and then that generations' cases), and then two less discernible generations from around day 30 onwards. Visible are some clustered groups of individuals with very similar exposure times (and many estimated as infected by the same individual,

	Posterior mean	Posterior median	Credible interval
$R_Q = (\mu_Q + b\mu_F)(\lambda_a + \lambda_f + \lambda_h)$	1.46	1.40	(0.62, 2.30)
$R_a = (\mu_R + b\mu_F)(\lambda_a)$	0.712	0.606	(0.0,1.63)
$R_f = (\mu_R + b\mu_F)(\lambda_f)$	1.09	1.03	(0.192,1.99)
$R_h = (\mu_R + b\mu_F)(\lambda_h)$	6.15	6.00	(2.85,9.45)
$R_{Qa} = (\mu_Q + b\mu_F)(\lambda_a)$	0.132	0.109	(0.0,0.313)
$R_{Qf} = (\mu_Q + b\mu_F)(\lambda_f)$	0.201	0.183	(0.014,0.388)
$R_{Qh} = (\mu_Q + b\mu_F)(\lambda_h)$	1.13	1.08	(0.411,1.84)

Table 2.8: Parameter estimates and equal-tailed 95% credible intervals for various reproduction numbers, where 100,000 MCMC samples are used. R_Q is the reproduction number for once quarantine measures are introduced. R_x is the reproduction number corresponding to the infection rate λ_x, where $x = a$, f or h, and R_{Qx} is equivalent, but once quarantine measures are in place.

as will be seen in Figure 2.9), which is more akin to a point-source outbreak where individuals are exposed to a highly infectious source for a short time, causing a sharp peak in cases. This highlights the high transmission potential of smallpox.

2.7.2 Source of Infection

In addition to the results analysed so far, we are also able to estimate the most likely path of smallpox transmission for the Abakaliki outbreak, i.e. who infected whom. This is a novel analysis for the Abakaliki data, since Eichner and Dietz' maximum likelihood approach does not allow for it.

Using our MCMC algorithm, we obtain samples from the posterior distribution of the estimated infector of each infective. If an individual j receives infectious pressure $\Lambda_j(t) = \sum_{k=1}^{m} a_k(t)$ at the time of their exposure, where $a_k(t)$ is the pressure from the kth of m infectives at time t, then the probability that

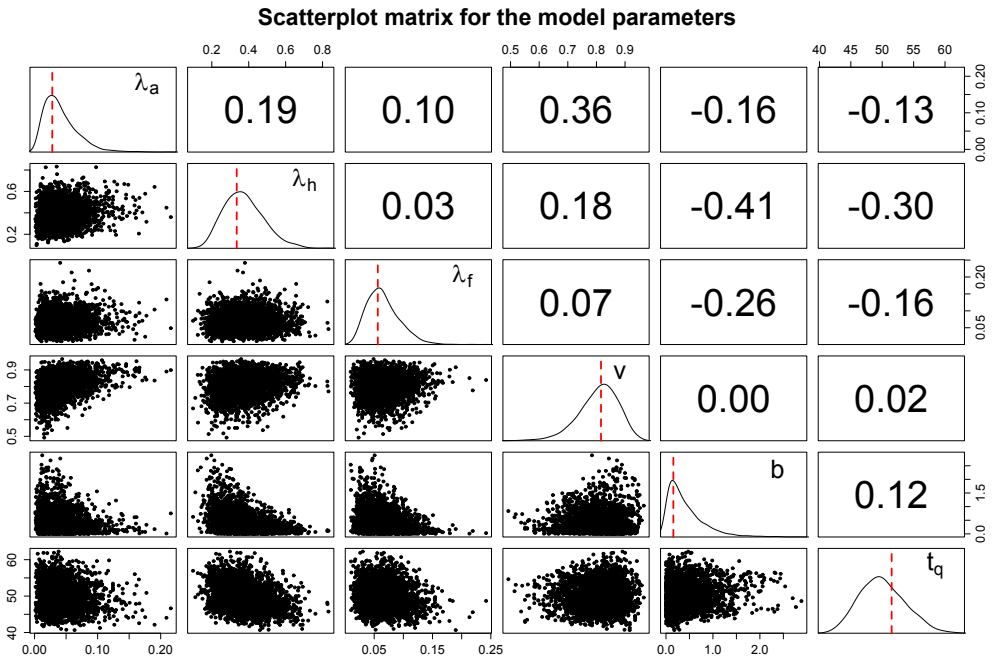


Figure 2.7: Scatterplot matrix of the model parameters, including Pearson's correlation coefficient for each pair and, on the diagonal, the posterior densities of the parameters.

individual k actually infected j is $\frac{a_k(t)}{\Lambda_j(t)}$. These samples can then be combined to find the estimated probability that any given individual infected any other.

Taking the individual estimated to have infected each person with the highest posterior probability, we obtain the estimated transmission pathway in Figure 2.9. We see that compound 1 acts very much as the root of the outbreak, with initial infective 0 infecting many individuals within this compound. Later generations of compound 1 infectives then lead to the spread of infection into other areas; note that infective 8 was one of the individuals who moved from compound 1 to 2, and is estimated to have been the one to introduce the disease into that compound. These findings agree with those in Thompson and Foege (1968), who stated that all of the first cases identified in compounds 2 through 9 except one could be traced to personal contact with a compound 1 infective.

Figure 2.10 displays the uncertainty around the most likely infector of each individual, by plotting the posterior probabilities of each infective having in-

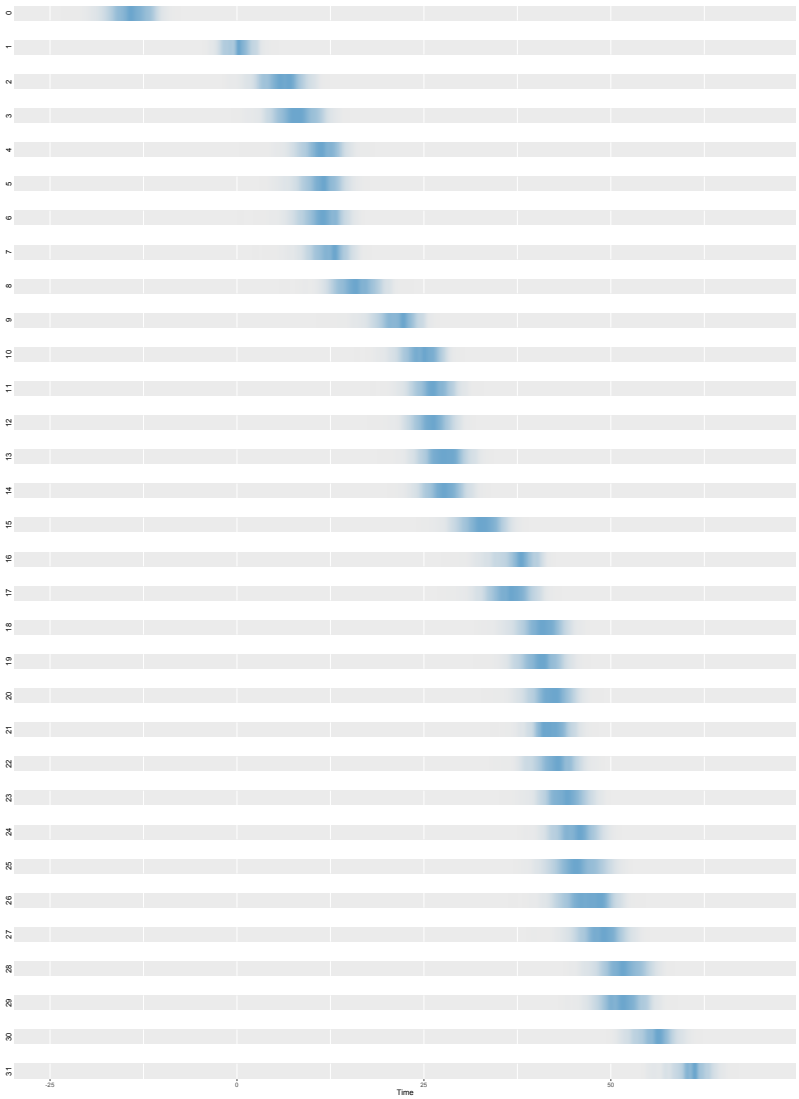


Figure 2.8: Heat map of the estimated exposure times of each infective, from 100,000 MCMC samples.

fected every other. We see that for some infectives, the earlier ones especially, we can be far more certain of their source of infection, whereas in the middle of the outbreak when new infections were being more rapidly discovered, the path of transmission is less easy to infer. This is as we might expect, and matches the analysis of the uncertainty around the exposure times.

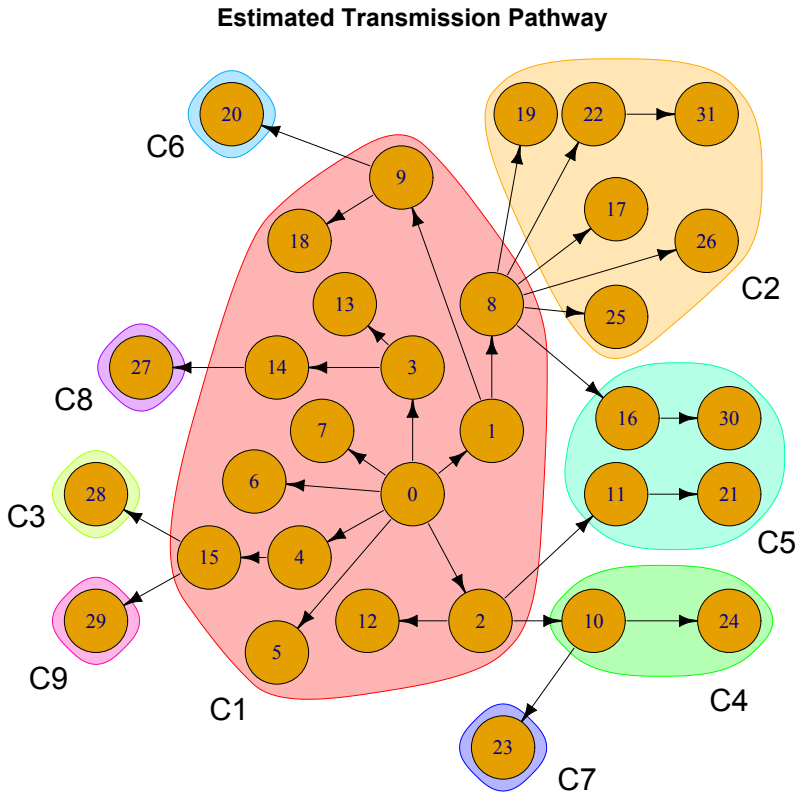


Figure 2.9: The estimated transmission pathway for the Abakaliki outbreak. Nodes represent infected individuals and the edge pointing to them represents the highest posterior probability among all possible infectors. Individuals are clustered by compound. Note that individuals 7 and 8 moved from compound 1 to compound 2 during the outbreak.

2.7.3 Simulation Study

To assess the performance of the MCMC algorithm, we now perform a simulation study. Following the structure outlined in Section 2.4, we simulate a number of outbreaks (discounting any of final size 1, which do occur with relative frequency) using both the Eichner and Dietz parameter estimates and two alternatives. We then perform MCMC on this simulated data, to see how well it is able to recover the true values. We simulate 30 data sets for each set of parameter values, as simulating a larger number was found to have no

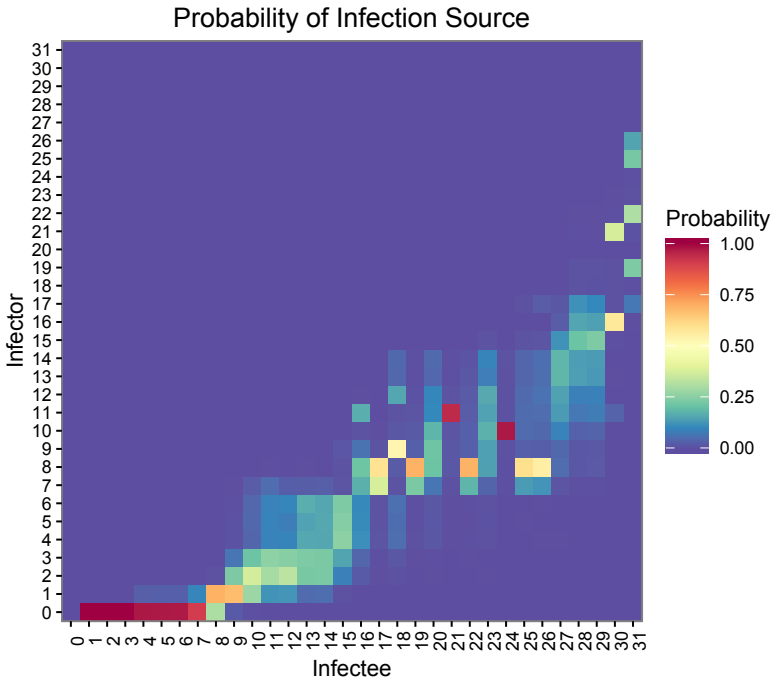


Figure 2.10: Heat map showing the posterior probabilities of individuals having infected others in the population of Abakaliki.

considerable impact on the results, and run the MCMC algorithm for each. This results in 30 posterior means per set of parameter values, for which we calculate the mean of these means as well as 95% probability intervals.

We compare the results from MCMC to the true values in Table 2.9. Density plots for the posterior mean values are given in Figures 2.11, 2.12 and 2.13. We see that the estimates are generally close to the true values, though with some overestimation. As we would expect, the estimates when $\lambda_a = 0.4$ are closer to the truth than for the Eichner and Dietz parameter values, since this causes larger outbreaks leading to more available information on the parameters. This is also the case, to some extent, when we increase t_q to cause larger outbreaks, as we might expect.

Table 2.10 shows the results of a single outbreak with much larger parameter values, yielding a large final size. In this, we did not update the event times or protection statuses, but fixed them to the values from the simulated data since otherwise computation was very slow. With a large outbreak we would hope

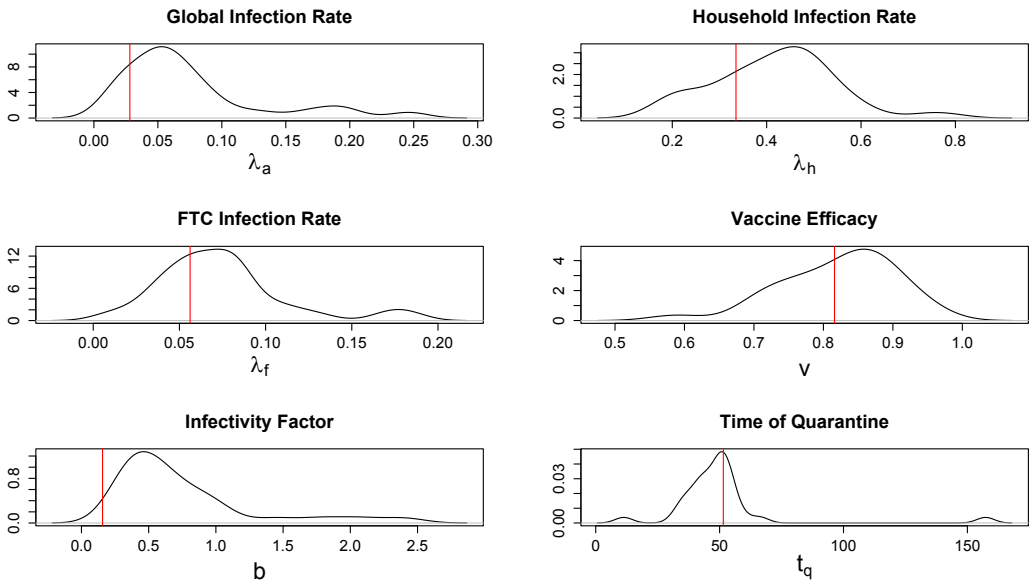


Figure 2.11: Density plots for means of the posterior estimates of 30 simulations for the Eichner and Dietz parameter values. Red lines represent the true values used in the simulations.

for estimates very close to the true values, which largely is the case except for some overestimation of λ_h. We expect that this might be due to a limit on the amount of information that can be gained about λ_h. Whereas for λ_a and λ_f we can learn more and more with larger outbreaks, since λ_h interactions only occur between compound infectives we are limited on what we can learn. For a number of large outbreaks simulated, only around 30 out of around 700 infectives resided inside the compounds, and it was found that as the number of compound infectives increases, the estimation of λ_h does improve.

Overall, these simulations suggest that the MCMC algorithm is performing well, as it is able to recover true parameter values from simulations with relative accuracy.

2.7.4 Sensitivity Analysis

We also perform a sensitivity analysis for model checking. This will assess the susceptibility of the results to changes in the underlying model assump-

Table 2.9: Simulation study results. 30 simulations per set of parameter values were created, and MCMC run on this simulated data. We provide the mean estimate of the 30 posterior mean values and a 95% probability interval, where 100,000 MCMC samples were obtained.

	Parameter	True Value	Mean of Posterior Means	95% Probability Interval	Range of Final Sizes Simulated
Eichner and Dietz	λ_a	0.0281	0.073	(0.0, 0.178)	15–42 (mean: 25)
	λ_f	0.0562	0.074	(0.001, 0.146)	
	λ_h	0.335	0.417	(0.181, 0.652)	
	v	0.816	0.822	(0.651, 0.992)	
	b	0.157	0.792	(0.0, 2.00)	
	t_q	51.5	52.1	(0.0, 105)	
Modified λ_a	λ_a	0.4	0.365	(0.188, 0.541)	65–109 (mean: 82)
	λ_f	0.0562	0.065	(0.016, 0.114)	
	λ_h	0.335	0.321	(0.081, 0.560)	
	v	0.816	0.787	(0.663, 0.910)	
	b	0.157	0.589	(0.0, 1.59)	
	t_q	51.5	50.0	(43.0, 57.0)	
Modified t_q and λ_f	λ_a	0.0281	0.031	(0.0,0.070)	12–133 (mean: 56)
	λ_f	0.20	0.193	(0.114,0.271)	
	λ_h	0.335	0.305	(0.142,0.467)	
	v	0.816	0.816	(0.710,0.921)	
	b	0.157	0.735	(0.0,1.78)	
	t_q	150	181	(70.4,291)	

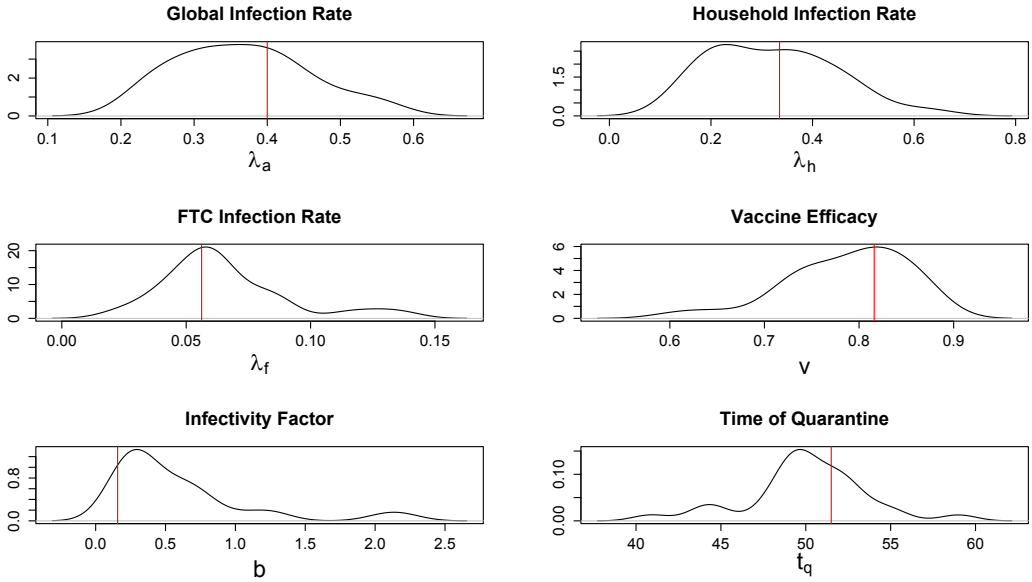


Figure 2.12: Density plots for means of the posterior estimates of 30 simulations for modified $\lambda_a = 0.4$. Red lines represent the true values used in the simulations.

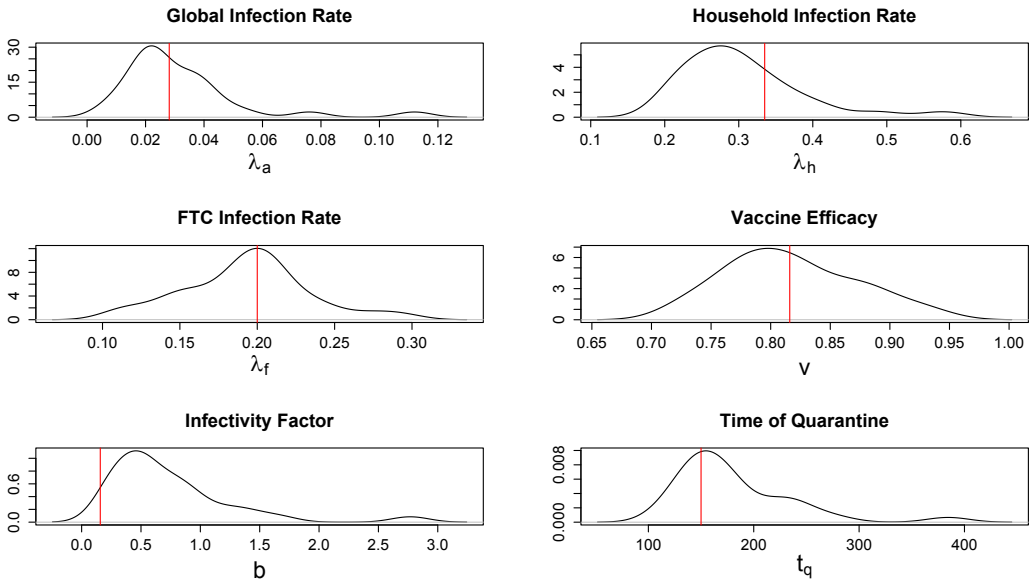


Figure 2.13: Density plots for means of the posterior estimates of 30 simulations for modified $t_q = 150$ and $\lambda_f = 0.2$. Red lines represent the true values used in the simulations.

Table 2.10: Simulation study results for a single large outbreak of final size 734. We provide the mean estimate over a single MCMC run of length 10,000, and the equal-tailed 95% credible interval.

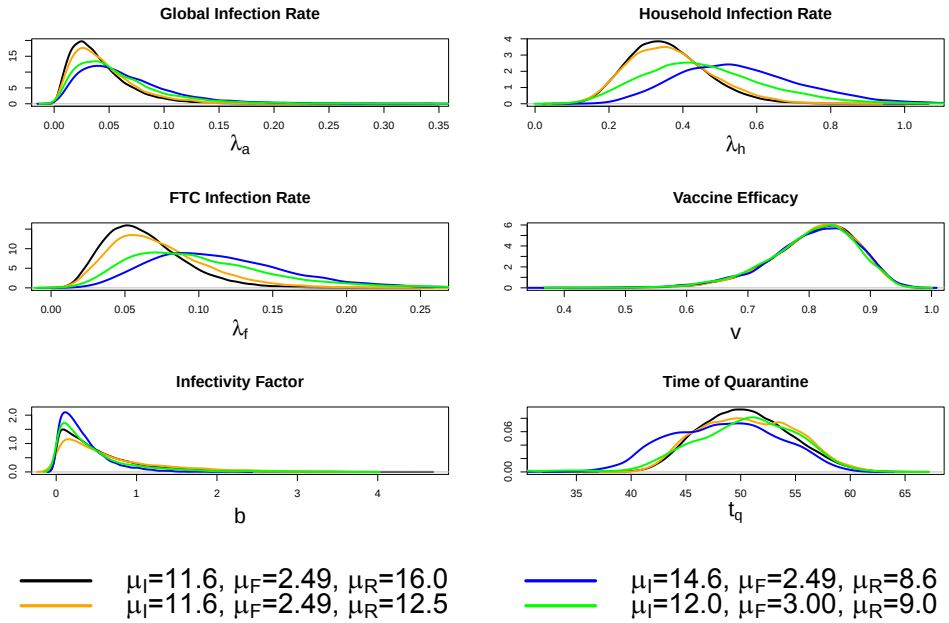
Parameter	True Value	Posterior mean	95% Credible Interval
λ_a	0.85	0.868	(0.762, 0.973)
λ_f	0.2562	0.265	(0.192, 0.337)
λ_h	0.535	0.711	(0.475, 0.946)
v	0.816	0.819	(0.648, 0.989)
b	0.157	0.132	(0, 1.335)
t_q	51.5	52.049	(-0.582, 104.680)

tions; in this case the infectious period length parameters μ_I, μ_R and so on. We will vary these parameters governing the length of time spent in the different disease stages, and examine the impact on parameter estimation from MCMC.

Figure 2.14 displays posterior densities for our parameters of interest over a range of infectious period mean durations. We vary μ_I, μ_F and μ_R and examine the effect on the shape of the posteriors. As would be expected, when μ_R, the length of the rash period before removal, is reduced to make shorter average infectious periods, the estimates of the infection rates increase to compensate. The estimates of the other three main parameters of interest are largely unchanged. Note that the estimation of R_0 is somewhat sensitive to the choice of μ_R, likely an artefact of the relatively small number of cases, the quarantine procedure and the population structure. In a large and uninterrupted outbreak we would typically expect R_0 to be determined by the outbreak size, but that is not the case here due to these extra complexities of population structure and control measures.

The values of μ_I, μ_F and μ_R have been informed by the literature, but we are perhaps less certain about the values of μ_Q and σ_Q as these are data-specific and not recorded in the Thompson and Foege (1968) report. We see the results of varying μ_Q and σ_Q, affecting the time taken to quarantine an infective, in

Figure 2.15, and note that this change has very little impact on estimation. This similarity is reassuring since these values are those we are least certain of, and implies that even if the time taken to quarantine infectives is different to that which we assume, the effect on our results is minimal.



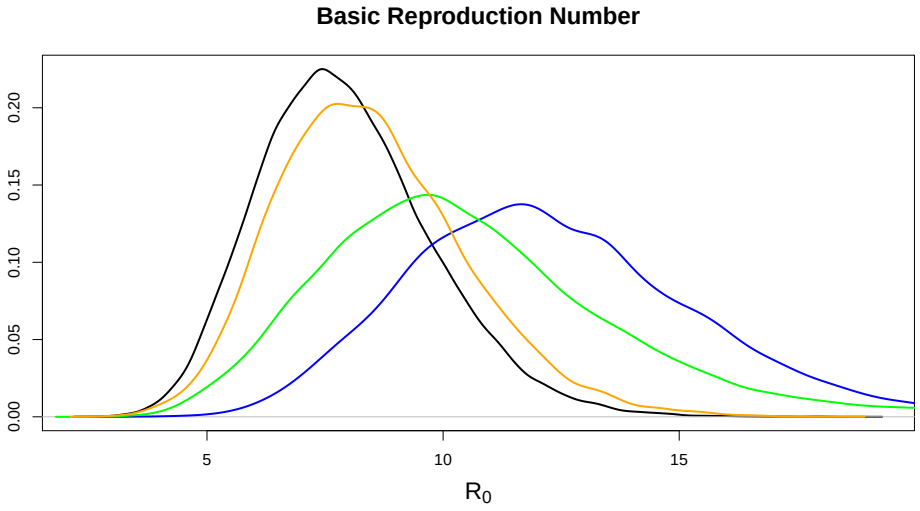


Figure 2.14: Posterior densities of the six parameters of interest and R_0, when different mean durations of the infectious periods are used.

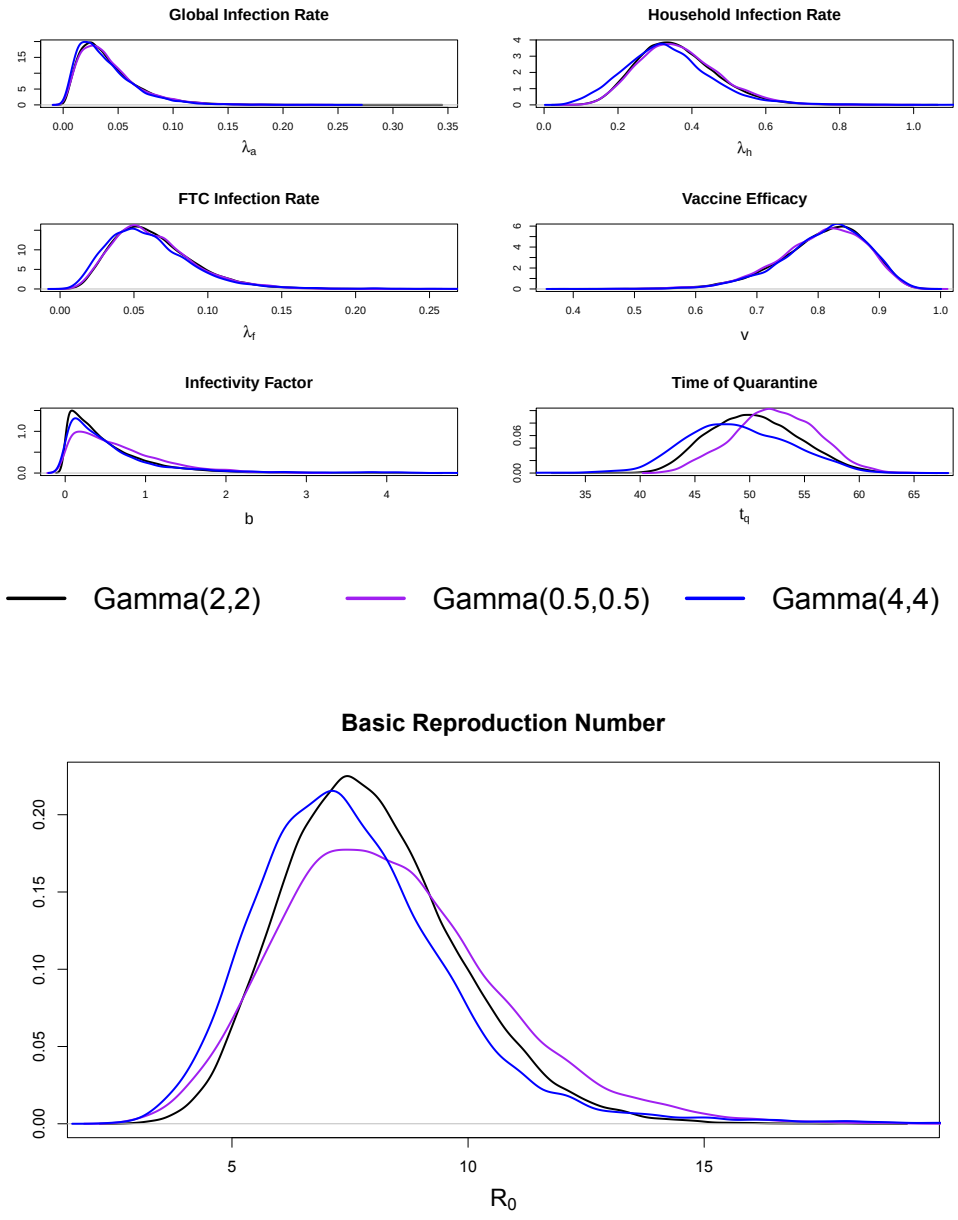


Figure 2.15: Posterior densities of the six parameters of interest and R_0, when the time taken to quarantine an infective, μ_Q, along with σ_Q is varied.

2.7.5 Posterior Predictive Checking

As a final piece of analysis, in order to assess how well our model fits the data we will perform posterior predictive checking. This involves taking samples of the basic model parameters from the posterior density, making sure to use a well mixed and thinned chain, and using our model to simulate a smallpox outbreak forwards in time from each. A well-performing model will lead to simulations similar to the Abakaliki outbreak. We use a number of measures to judge this similarity, namely comparing final size, epidemic duration and incidence curves.

We begin with final size; the total number of infectives in each outbreak. Taking 5000 sets of posterior estimates from a well mixed MCMC chain, we simulate a smallpox outbreak from each and record the final size. Figure 2.16 shows a histogram of these final sizes where we see that, in this respect, the simulations are fairly similar to the true Abakaliki outbreak, but generally of slightly smaller final size. The mean final size is 23.5 compared to the Abakaliki outbreak size of 32. This is somewhat surprising; since we allow for infections outside of the compounds, which were not seen in the data, we might expect the final size of simulations to be larger on average.

In the Abakaliki outbreak, however, it is important to note that two of the four individuals who moved compound were infective at the time of the move, and that these individuals were the first cases seen outside of compound 1. If we consider only those simulations in which at least one of the moving individuals was infective, we see an increase in the mean final size to 29.27. Figure 2.17 compares this subset of the simulations to all of the simulations as a whole. We see that when we only consider simulations similar to the data in this respect, the final size is much closer to that observed. This also supports our previous comment in Section 2.7.2 that the move of these two infected individuals was key in transmitting the disease outside of compound 1.

We next consider epidemic duration, which we define as the length of time be-

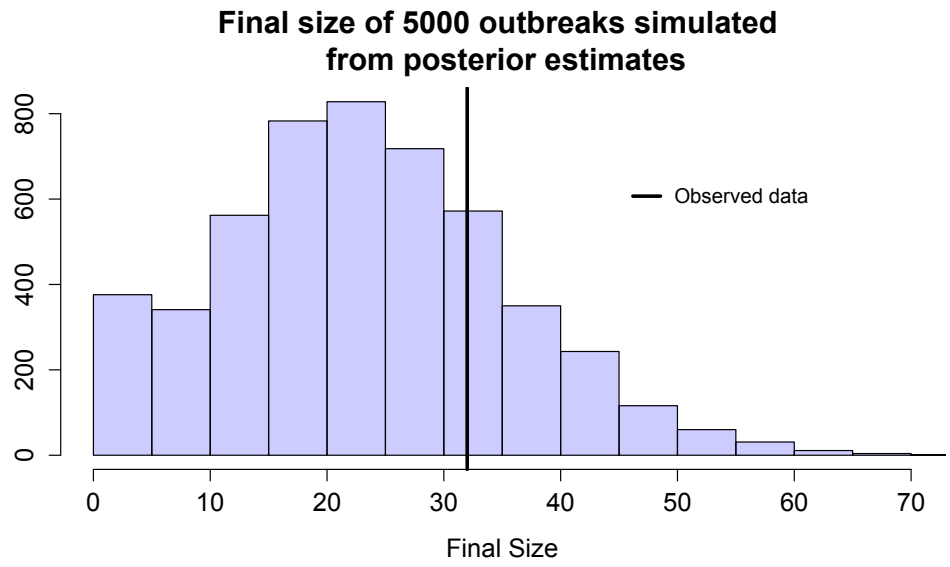


Figure 2.16: Using 5000 posterior samples of the parameter estimates to simulate outbreaks, a histogram displaying the final size of each, compared to the observed Abakaliki data.

tween the first case detection (rash) time and the last. Figure 2.18 displays the duration of 5000 simulated outbreaks, compared to the observed data duration. With a mean of 76.75 days, we see our simulated outbreaks are generally very similar in duration to the Abakaliki outbreak (of length 76 days), most likely due to the good estimation of t_q. Note the slight peak for short outbreaks of length 0-10, caused largely by those which immediately went extinct.

In the same manner as before, comparing the epidemic duration of the subset of outbreaks where infected individuals moved compound, we see a small increase in the mean. Figure 2.19 shows this, and it is interesting that in this case the subset of simulations provides a worse estimate of the observed epidemic duration. However, we would indeed expect an increase in epidemic duration from the increased final size of these simulations compared to all of the simulations as a whole.

Figure 2.20 allows for examination of the correlation between final size and epidemic duration for the simulated outbreaks. As would be expected, longer durations tend to be seen when there are larger outbreaks, with the two having

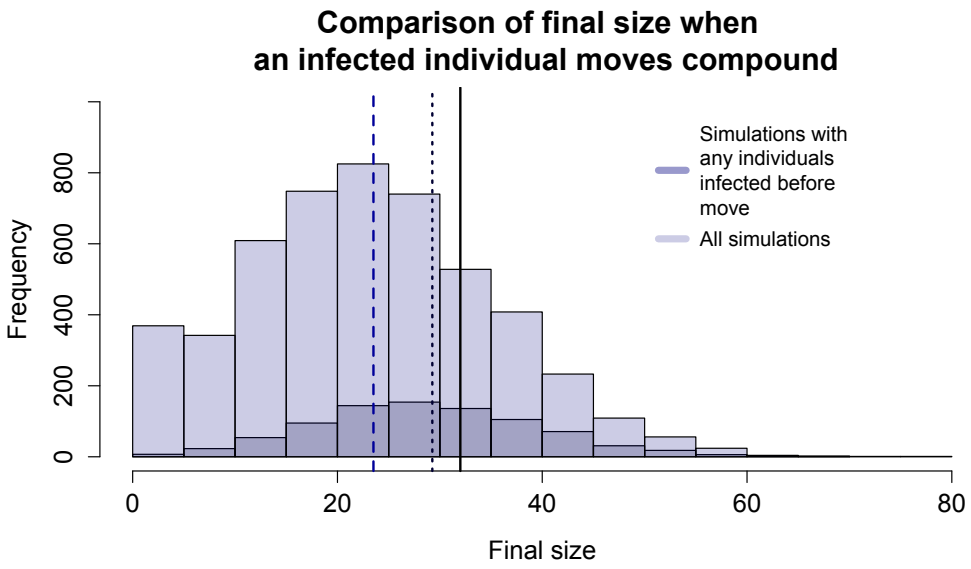


Figure 2.17: A comparison of final size between all 5000 simulations and the subset of simulations where an individual who moves compound is infected during their move. The dashed line represents the mean final size of all simulations, the dotted line represents the mean final size of only those outbreaks where an infected individual moves compound, and the solid black line represents the observed data.

correlation of 0.57. The result for the true Abakaliki data does not appear distinctly different to that of the simulations.

Table 2.11 provides a brief set of statistics for the simulated data sets compared to the Abakaliki outbreak, summarising the plots discussed so far. Note the difference in the percentage of outside infectives seen, which was not recorded in the Abakaliki outbreak but invariably occurs in simulations. Although the simulations do not appear similar to the data in this respect, it is noteworthy that Thompson and Foege (1968) claimed that 'there must have been some deaths which were very well concealed' within Abakaliki, and so it is certainly not unquestionable that there were indeed infectives outside of the compounds who were just not recorded.

Lastly, we wish to compare the cumulative number of cases at any given time in simulations to the Abakaliki data. However, this is difficult to do with out-

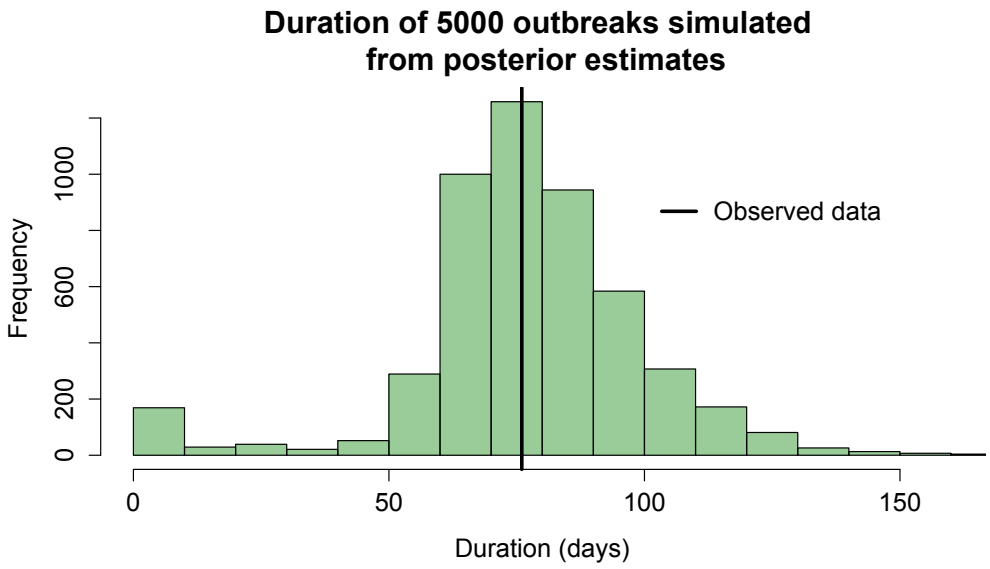


Figure 2.18: Using 5000 posterior samples of the parameter estimates to simulate outbreaks, a histogram displaying the duration from the first rash time to the last of each, compared to the observed Abakaliki data.

breaks of different sizes, and so we take instead only those simulations of the same final size as the data (32). Figure 2.21 displays the incidence curves of 4000 simulations of size 32, which we see are generally similar in shape to the data, perhaps with some right skew. This implies that the observed data are reasonably well captured by the behaviour of the model. To quantify this more exactly, we calculate a posterior predictive p-value for the discrepancy between simulations and the data, defined as the probability that a simulation $\mathbf{R}^{\text{rep}}$ is more extreme than the data $\mathbf{R}^{\text{obs}}$. Here, element R_j is equal to the j^{th} rash time (where rash times are chronologically ordered as usual). We would wish for simulations to be more extreme around 50% of the time, so a p-value of 0.5 is optimum. In order to calculate the p-value, we must select a discrepancy measure $D(\mathbf{R}, \mathbf{\Phi})$; a function of data $\mathbf{R}$ and model parameters $\mathbf{\Phi}$. We use a chi-squared measure as detailed in Gelman et al. (1996), of the form

$$D(\mathbf{R}, \mathbf{\Phi}) = \sum_j \frac{\left(R_j - \mathbb{E}(R_j \mid \mathbf{\Phi})\right)^2}{\text{Var}(R_j \mid \mathbf{\Phi})}.$$

Note that neither the mean nor the variance term are available analytically,

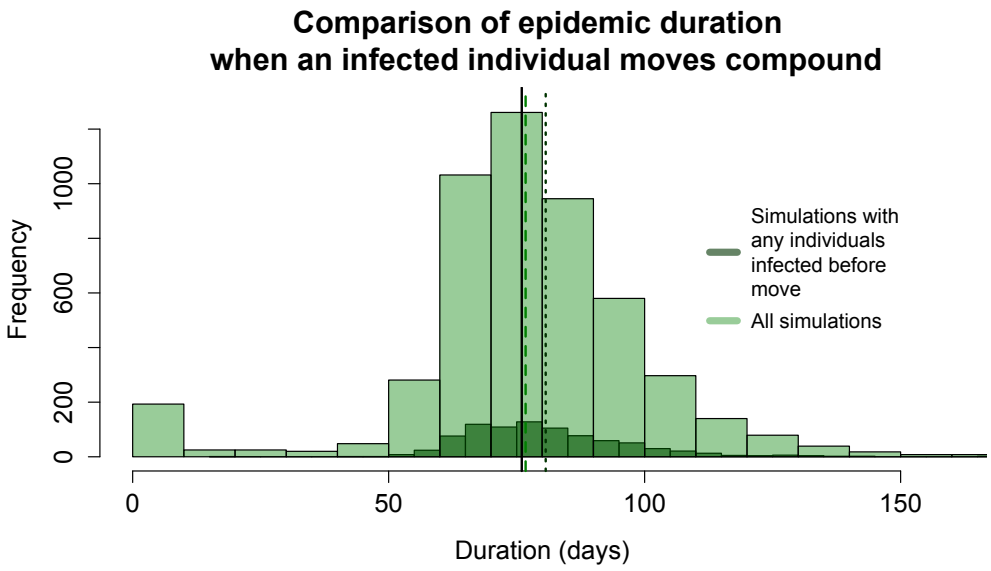


Figure 2.19: A comparison of epidemic duration between all 5000 simulations and only those simulations where an individual who moves compound is infected before the move. The dashed line represents the mean duration of all simulations, the dotted line represents the mean duration of only those outbreaks where an infected individual moved compound and the solid black line represents the observed data.

and so we obtain these via simulation. Given $\mathbf{\Phi}$, we simulate until we have a suitably sized sample of outbreaks with 32 cases. The mean and variance of the jth rash time is then estimated directly from this sample. Next suppose that we have M samples from the posterior, labelled $\mathbf{\Phi}^{(1)}, \ldots, \mathbf{\Phi}^{(M)}$. Repeatedly simulating until we obtain an outbreak with 32 cases, we use the ith sample of $\mathbf{\Phi}$ to obtain a simulated epidemic with rash times R^{rep_i}.

Then the posterior predictive p-value is defined as

$$\begin{aligned} \text{ppp-value} &= \mathbb{P}\Big(D(R^{\text{rep}}, \mathbf{\Phi}) \geq D(R^{\text{obs}}, \mathbf{\Phi}) \mid R^{\text{obs}}\Big) \\ &\approx \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{D(R^{\text{rep}_i}, \mathbf{\Phi}^{(i)}) \geq D(R^{\text{obs}}, \mathbf{\Phi}^{(i)})}. \end{aligned}$$

A p-value of 0.42 is obtained (for $M = 100$), which is sufficiently close to the optimum value of 0.5 to be accepted. Hence we conclude that the simulated

Table 2.11: Comparison of 4000 simulated outbreaks from posterior estimates, and the Abakaliki data over a range of criteria. * This value is calculated considering only outbreaks where at least one of the individuals who moves compound is infected by the time of their move

	Mean of Simulations	Abakaliki Data
Outbreak duration (days)	76.75	76
Final size	23.51 (29.27*)	32
Percentage of outside infectives	19.99%	0%
Percentage of FTC infectives	90.71%	93.8%

outbreaks of size 32 are similar in this respect to the data, indicative of a good model fit. A more accurate value could be obtained with larger values of M, but the procedure is highly time-consuming in practice due to the restriction on the final size of simulations.

Overall, posterior predictive checking has shown a good model fit, indicating that our MCMC results are reliable. It has also highlighted some important aspects of the data, such as further confirming the importance of the individuals who moved compound to the spread of the outbreak, and showing the large proportion of infectives outside the compounds in simulations, which may be cause for further work.

2.8 Discussion

2.8.1 Parameter Estimates

Our estimates of the infection rates show clearly that the dominant mode of transmission within Abakaliki was between individuals in the same compound, in agreement with the findings of Eichner and Dietz (2003). This is supported by the estimated disease transmission pathway of Section 2.7.2 where

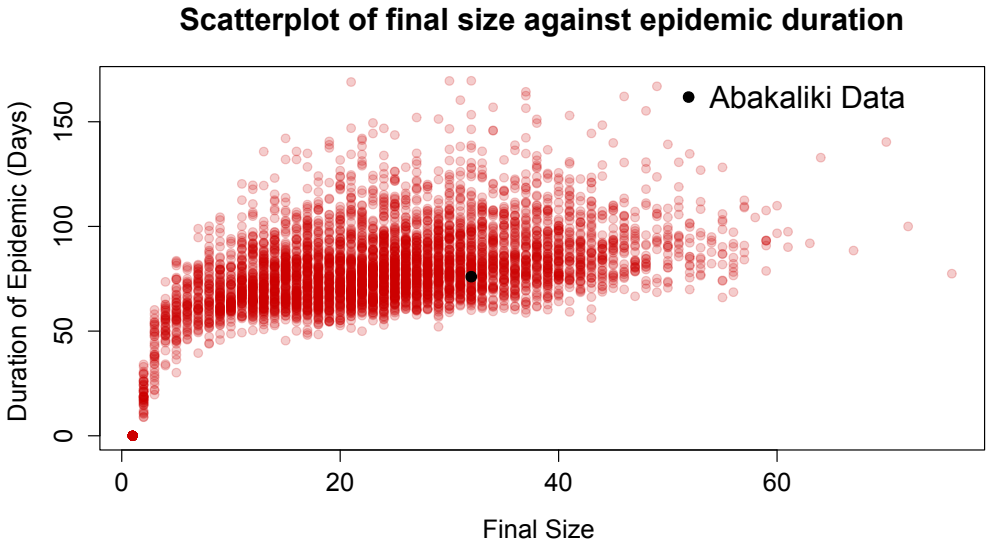


Figure 2.20: For 1000 simulated outbreaks, a scatterplot of final size against the duration of the epidemic. The black point provides the values from the Abakaliki outbreak.

we see the majority of transmission events taking place within compounds. This is also in agreement with the original WHO report by Thompson and Foege (1968) who found that FTC membership did not appear to be the major transmission mechanism but rather that compound links and particularly family membership were dominant.

Again similarly to Eichner and Dietz, we found the vaccine to have had around 81% efficacy. Our estimate for b, the factor for the change in infectivity during the fever period, was 0.5. Higher than that of Eichner of Dietz, this is likely due to the skewed shape of the posterior density but shows evidence that the modelling of smallpox with an SIR model that has been frequently seen in those citing the Abakaliki data may not be appropriate.

The time quarantine measures were introduced was estimated to be between day 50 and 51. From our investigation, it seems that the introduction of control procedures was somewhat important in preventing a much larger scale outbreak. Under the model assumptions alone, the quarantine reduced the average time spent in the rash period from 16 days to 2 days. However, inves-

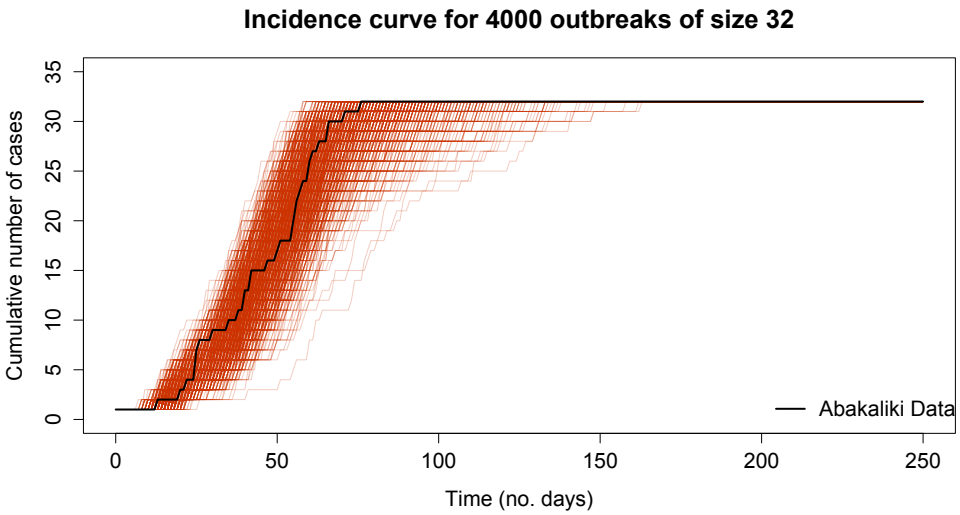


Figure 2.21: The cumulative number of smallpox cases observed by each day is shown, for 4000 outbreaks of size 32 only. The black curve shows the incidence curve for the true Abakaliki outbreak.

tigation of reproduction numbers for the outbreak led to further discoveries about this.

2.8.2 Reproduction Numbers

We obtained a posterior mean estimate for R_0 of 7.96, higher than that found by Eichner and Dietz (2003) of 6.87 and even more so than other estimates for an SIR model where just FTC individuals are considered (see O'Neill and Roberts (1999), who estimate R_0 around 1). This highlights the high infectivity of smallpox as well as the importance of a detailed model taking population structure into account.

Although R_0 is interpreted as the average number of secondary cases any given infective will cause in a fully susceptible population, in our case this is hard to analogize since the majority of transmission was within-compound and the pool of susceptible individuals in an infected compound depleted rapidly.

To further consider the impact of control measures in a more relevant way,

we considered the reproduction numbers both pre- and post-quarantine procedures being put in place. We obtained an R_Q (reproduction number when quarantine measures are in place) value of 1.46; far lower than pre-quarantine but interestingly insufficient alone to stop large scale spread. We concluded that important must also have been the depletion of susceptible individuals within the affected compounds (which had a large proportion of FTC inhabitants), and a sub-critical epidemic elsewhere in the population. With respect to contacts between FTC individuals, it appears that the quarantine measures were key in minimising the disease spread, but for compound contacts it was the depletion of susceptible individuals which slowed the epidemic rather than quarantine. Despite this, introducing quarantine procedures at a later day in simulations was found to increase outbreak size somewhat, with average final sizes 24, 44 and 64 for $t_q = 50$, 100 and 200 respectively. With no quarantine procedures at all, we found the average final size to be 86, highlighting the sub-critical epidemic in the wider population and the considerable impact of the depletion of susceptible individuals within the compounds. As in Thompson and Foege (1968), we see that small pockets of poorly protected individuals who mix frequently together can facilitate outbreaks of smallpox even in a generally well-protected population.

2.8.3 Model Fit

The tests we have performed indicate that the model fits the data fairly well. The model does invariably predict cases outside of the compounds, likely because of the rather unrealistic assumption of homogeneous mixing in the entire population, and especially the assumption of homogeneous mixing of non-FTC individuals which there was no data available to inform. Thompson and Foege (1968) stated that the FTC community was largely isolated from the rest of the population with the exception of a few traders, and so a model in which just some fraction of FTC members had contact with the outside community might be more applicable, although more complex and not directly informed

by any data.

Something that might also be useful to consider, although again there is not sufficient data to support this, is the inclusion of age categories. Thompson and Foege (1968) state that there appeared to be a much larger proportion of susceptibles among children, and consequently a higher attack rate. Pre-school children in particular were largely susceptible, as many older children had been vaccinated in school despite their parents' beliefs. Even with available data, however, it seems probable that a model accounting for age of the individual would be over-parameterised.

2.8.4 Accuracy of the Eichner and Dietz Likelihood Approximation

As we have seen, the results of our full Bayesian analysis are fairly similar to those of Eichner and Dietz (2003), indicating that their approximation method may be of use in other situations.

Investigation of the Eichner and Dietz method reveals that the likelihood function obtained is numerically but not analytically tractable, specifically since it involves integrals which must be numerically evaluated. Although this suffices for maximum likelihood methods, as used by Eichner and Dietz for the Abakaliki outbreak, it is prohibitive for use within MCMC algorithms as the likelihood must be repeatedly evaluated at large computational cost. We also note that the distributions for the length of time in each disease stage used here have relatively small variances, meaning that the model is closely comparable to one in which the event times are assumed known. In this case, Eichner and Dietz's method provides the true likelihood since the distributions used to approximate unknown event times collapse to point masses around the true values. It is of interest, therefore, to develop approximate likelihood functions which are both useful for non-constant infectious periods and which are analytically tractable.

2.9 Conclusions

In this chapter we have completed a full Bayesian analysis of the Abakaliki smallpox data, and compared the results of this to those of Eichner and Dietz (2003). The parameter estimates found highlight the dominance of within-compound smallpox transmission as well as the impact on the end of the outbreak of susceptible depletion within the compounds, rather than quarantine procedures alone. Novel results include estimates of the transmission pathway as well as analysis of the uncertainty around exposure times, and model checking has confirmed that the model fits well. Overall, we have seen that our parameter estimates are very similar to those of Eichner and Dietz. This indicates that analytically tractable approximate likelihood functions are of interest to investigate, in particular for situations where current methods struggle such as large populations and multi-level mixing.

CHAPTER 3

Likelihood Approximation Methods

3.1 Introduction and Motivation

As our investigation in Chapter 2 has shown, inference for disease outbreak data often deals with complex models and large amounts of missing data, requiring problem-specific analysis and computation. Although MCMC methods have become considered somewhat the 'gold standard' for analysis of this kind of data, there are many problems associated with this (see e.g. De Angelis et al., 2015, O'Neill, 2010, and references therein). Specifically, there are issues with data dependency as well as more realistic models leading to more difficult analysis, especially in terms of computational burden. In this chapter, we begin by exploring a number of these problems, and then suggest new methodology which seeks to address them.

A primary problem with the analysis of infectious disease data is that these data may be highly dependent. MCMC algorithms which require imputation of large amounts of missing data then often mix very slowly. Details of this are discussed in Kypraios (2007), where high posterior correlations between the infection times and the infectious period parameter are shown. Specifically, as the outbreak size n increases, infectious period parameter γ (assum-

ing exponential infectious periods for ease) and the sum of the infectious periods $B = \sum_{j=1}^{n}(r_j - i_j)$ become more highly correlated *a posteriori*. If γ and B were our parameters of interest, a two-state Gibbs sampler may suffer mixing problems due to this dependency (see e.g. Roberts and Sahu, 1997). It becomes increasingly difficult to update γ and B separately, since one essentially determines the other. This is complicated even further by our DA-MCMC algorithm, which may only update a subset of the infection times per iteration rather than the whole sum B.

To understand why this high correlation occurs, we may consider the full conditional distribution of γ, as defined in Equation (1.3.10). Assuming a low-rate exponential prior, this full conditional distribution is roughly $\Gamma(n+1, B)$. This therefore has variance $\frac{n+1}{B^2}$, which tends to zero as n goes to infinity, since B is of order n. In other words, the larger n gets, the more γ is determined by the value of B, and the correlation is stronger. Methods have been developed to combat this problem, such as partially (or completely) non-centered parameterisations (see e.g. Neal and Roberts, 2005) which can lead to faster convergence of the Markov chain by, for example, proposing new infection times when updating γ rather than performing these independently. However, these are not easily applicable to all MCMC samplers with all models (Papaspiliopoulos et al., 2003).

As well as the potential introduction of mixing challenges in the MCMC, large amounts of missing data resulting from large populations/outbreaks require many, possibly costly, evaluations of the likelihood. A large population size also increases the time to calculate this likelihood since it includes a product over individuals, and DA-MCMC hence becomes highly computationally intensive. With the increasing ability to easily collect and store large data sets, as well as the growing interconnectedness of communities, there is a rising demand for realistic analyses of large scale outbreaks. Also, particularly in cases of real-time forecasting as is becoming more commonplace, the ability to perform estimation quickly and efficiently will be key. These problems are

hence becoming more significant, and there is a real need for the development of methodology which might solve them.

Options for improving the computational speed of MCMC algorithms do exist, such as the use of parallel computing, but since we require a record of the states visited by the Markov chain this is not particularly straightforward. Other methods (e.g. Rambaut et al., 2008) have partitioned the data and analysed each section independently, but of course this ignores any correlation between the different sections. We hence propose that methods which include the entirety of the data whilst avoiding the need for data augmentation, or indeed MCMC entirely, could become useful tools for analysis.

In Chapter 2, one key finding was the relative accuracy of Eichner and Dietz' approximation method compared to standard MCMC methods, despite the lack of data augmentation. This indicates that the development of likelihood approximation methods may be useful for addressing our computational problems. DA-MCMC allows us to sample from the high dimensional probability densities often involved in the analysis of infectious disease data, even though these may not be analytically written down. Conversely, these approximation methods will seek to reduce this dimensionality by eliminating the artificial parameters that are the augmented data, so that we may sample from the posterior directly.

In this chapter we will introduce a series of likelihood approximations which attempt to tackle the problems we have discussed. The first will be a generalised version of Eichner and Dietz' method from Chapter 2, and the remainder will be a new series of approximation methods based on assuming independence in interactions between pairs of individuals. We name these Pair-Based Likelihood Approximations (PBLA). As we have seen in the previous chapter, the true likelihood in these analyses is equal to the integral over all unknown event times of the augmented likelihood, which is itself a product over all individuals. In assuming that all pairs of individuals make independent contributions to the likelihood, we are essentially able to reduce

this high dimensional integral to a product of two dimensional integrals, each of which may be analytically calculated. The likelihood expression is then entirely tractable and we may approximately sample from it using standard MCMC, or even simply obtain parameter estimates with maximum likelihood estimation methods. This is similar in spirit to composite likelihood methods (see e.g. Varin et al., 2011 and particularly the pairwise likelihood in Cox and Read, 2004) in that we write the likelihood as a product over constituent parts, but as we will explain more thoroughly in Section 3.4.1, the similarity does not extend much beyond this. These Eichner and Dietz and Pair-Based Likelihood Approximation methods may then be used in situations requiring complex models with many parameters, for large populations and for large amounts of missing data, to combat the computational issues we have discussed.

The chapter will proceed as follows. We begin by defining the general model and notation to be used for the likelihoods in this chapter in Section 3.2. The first approximation, which will be introduced in Section 3.3, will be the generalization of Eichner and Dietz' approach from Chapter 2. This framework will be applied to the special cases of exponentially distributed infectious periods (Section 3.3.2) and gamma distributed infectious periods (with only positive integer valued shape parameters) (Section 3.3.3). In Section 3.4 we will then introduce and define the series of Pair-Based Likelihood Approximations. For each approximation we include the general framework as well as specific calculations for given infectious periods. We conclude the chapter with a discussion of a numerical drawback of the PBLA method in Section 3.4.10, as well as an extension of the method to SEIR models in Section 3.4.11.

3.2 Model and Likelihood

We begin by defining the stochastic epidemic model and basic approximate likelihood structure to be used in this chapter.

For simplicity, we will at first restrict our attention to the SIR model. At

any given time, every individual in a closed population of size N will be in one of three states: susceptible, infected or removed, and individuals will progress through the states in that order. Individuals in the population are labelled $1, 2, \ldots, N$, and we label the cases as all individuals $j = 1, 2, \ldots, n$ who become infected (where n is the final size of the outbreak). For these individuals, i_j denotes their infection time and r_j their removal time. The infection times $\mathbf{i} = \{i_j : j = 1, 2, \ldots, \kappa - 1, \kappa + 1, \ldots, n\}$, where κ is the unknown initial infective, are unknown, and the data consist of removal times $\mathbf{r} = \{r_j : j = 1, 2, \ldots, n, \text{ where } r_1 < r_2 < \cdots < r_n\}$. Individuals are therefore ordered such that $1, 2, \ldots, n$ are those who will eventually be infected ($n \leq N$ necessarily). Then $n, n + 1, \ldots, N$ are the individuals who remain susceptible at the end of the outbreak. Note that we require $r_j < r_{j+1}$ strictly for all $j = 1, 2, \ldots, n - 1$, which will be discussed further in Section 3.4.10.

The outbreak begins with the infection of the initial infective κ, at time i_κ, and continues until no infectious individuals remain. The infectious periods of different infectives are assumed independent and identically distributed, with probability density (or mass) function $f_I(\cdot \mid \boldsymbol{\theta})$, where f_I has parameter vector $\boldsymbol{\theta}$. We do not allow for reinfection, so any individual who reaches the removed stage will remain there for the duration of the outbreak. During any individual i's infectious period, they will have contact with any other individual j at a time given by the point of a Poisson process of rate β_{ij}, where all such Poisson processes are assumed mutually independent. If a contact occurs with a susceptible individual, this results in their immediate infection. Then we define $\boldsymbol{\beta} = \{\beta_{ij} : i, j \in \{1, 2, \ldots N\}\}$ as a matrix of these contact rates. This allows for a wide range of possibilities for population structure: homogeneous mixing, multi-level mixing and network structures may all be incorporated. We may then also define the concept of infectious pressure on any susceptible j, as the sum over all current infectives i of β_{ij}. A higher infectious pressure essentially represents an increased probability of individual j being infected, since more infectives may have potential contacts with them.

Augmenting the observed removal times with the unobserved infection times and the identity of the initial infective, as in Chapter 2, we define the augmented likelihood as a product over individuals, but with new notation, as

$$\pi(\mathbf{i}, \mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \kappa, i_\kappa) = \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \phi_j f_I(r_j - i_j) \right) \phi_\kappa f_I(r_\kappa - i_\kappa)$$

where

$$\begin{aligned}
\chi_j &= \text{Infectious pressure acting on } j \text{ at time of infection} \\
&= \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} \mathbb{1}_{\{k \text{ infective at } i_j\}}, \\
\psi_j &= \mathbb{P}(j \text{ avoids infection until time } i_j) \\
&= \exp\Big(- \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} (r_k \wedge i_j - i_k \wedge i_j) \Big) \\
\phi_j &= \mathbb{P}(j \text{ fails to infect all non-infected individuals, labelled } n+1, \ldots, N), \\
&= \exp\Big(- \sum_{k=n+1}^{N} \beta_{jk} (r_j - i_j) \Big). \qquad (3.2.1)
\end{aligned}$$

We recall $a \wedge b$ is the minimum of a and b, and hence $r_k \wedge i_j - i_k \wedge i_j$ represents the total length of time for which there is infectious pressure between individuals j and k. This new notation will be of practical use for the introduction of the likelihood approximation methods.

Since the infection times and κ are unobserved, we obtain the target likelihood by integrating over the infection times and the identity of the initial infective so that

$$\pi(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \int \pi(\mathbf{i}, \mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \kappa, i_\kappa) \pi(i_\kappa, \kappa) \, d\mathbf{i} \, di_\kappa \, d\kappa,$$

where we have assumed that i_κ and κ are independent of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ *a priori*. Next,

see that

$$\begin{aligned}
\pi(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) &= \int \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \phi_j f_I(r_j - i_j) \right) \phi_\kappa f_I(r_\kappa - i_\kappa) \\
&\quad \times \pi(i_\kappa, \kappa) \, d\mathbf{i} \, di_\kappa \, d\kappa \qquad (3.2.2) \\
&= \sum_{\kappa=1}^{n} \pi(\kappa) \int \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \phi_j \right) \phi_\kappa f_I(r_\kappa - i_\kappa) \\
&\quad \times \pi(i_\kappa \mid \kappa) \prod_{\substack{j=1 \\ j \neq \kappa}}^{n} f_I(r_j - i_j) \, d\mathbf{i} \, di_\kappa \\
&= \sum_{\kappa=1}^{n} \pi(\kappa) \mathop{\mathbb{E}}_{\mathbf{i}, i_\kappa} \left[\left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \phi_j \right) \phi_\kappa \pi(i_\kappa \mid \kappa) \right], \qquad (3.2.3)
\end{aligned}$$

so that we take expectations over all of the infection times (including i_κ), which are independent and identically distributed from $f_I(r_j - i_j \mid \boldsymbol{\theta})$. However, it is difficult to evaluate the likelihood in this form using a Monte Carlo scheme, since in order to avoid the likelihood being equal to zero we require $\chi_j > 0$ for all individuals j except the initial infective. This means positive infectious pressure on all individuals at the moment they become infective. In practice, a large number of Monte Carlo simulations would lead to 'impossible' outbreaks, i.e. those which are inconsistent with the observed data since there is not a potential infector for all infectees.

One way to proceed is to assume independence over j (our first approximation), so that

$$\mathop{\mathbb{E}}_{\mathbf{i}, i_\kappa} \left[\left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \phi_j \right) \phi_\kappa \pi(i_\kappa \mid \kappa) \right] \approx \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \mathop{\mathbb{E}}_{\mathbf{i}, i_\kappa} \left[\chi_j \psi_j \phi_j \right] \right) \mathop{\mathbb{E}}_{i_\kappa} \left[\phi_\kappa \pi(i_\kappa \mid \kappa) \right]. \qquad (3.2.4)$$

We are hence assuming that each individual's contribution to the likelihood is independent, for instance that the infectious pressure on some individual j when they are infected is independent of the infectious pressure on k when they are infected. This is not strictly true, since each quantity may be influenced by, for example, whether a third individual l was able to place infectious pressure upon each of them. We might expect this approximation to be

most accurate when the infection times are relatively spread out compared to the length of the infectious periods, since there will be less uncertainty as to which individuals were able to infect others.

There are a number of possible ways to make further approximations to the likelihood given in Equation (3.2.4), which we will explore in the remainder of this chapter. The first will be a generalized version of Eichner and Dietz' method from the Abakaliki data analysis of Chapter 2.

3.3 The Eichner and Dietz Approximation

3.3.1 General Framework

In the analysis of the Abakaliki smallpox data in Chapter 2, we compared our results from DA-MCMC to those of Eichner and Dietz (2003), who used a likelihood approximation to perform maximum likelihood estimation. We found that the results using the true likelihood were very similar to those from the approximation, and this has motivated this chapter which aims to further develop these kind of approaches. Here, we will provide a full derivation and description of a generalized Eichner and Dietz (ED) likelihood, restricting our attention to a homogeneously mixing population and SIR model. Under this model, the contact rate between any pair of individuals is given by $\frac{\beta}{N}$. Although Eichner and Dietz did not do this, we also assume that the initial infective is the individual whose removal time is first (r_1) for simplicity. We do not include their infection process in the likelihood since they are assumed to have been infected before the start of the outbreak, though of course they are able to infect others.

Considering the likelihood in the format which Eichner and Dietz use, we begin by considering the force of infection $\Lambda_j(t)$ to which any individual j is exposed at time t. In the SIR case this is equal to the infection rate $\frac{\beta}{N}$ multiplied by the number of currently infective individuals. With the infection

times unobserved, the number of infectives at any time is unknown also, and so we incorporate the probability that each individual is infective at a given time. The force of infection is then given by

$$\Lambda_j(t) = \sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{\beta}{N} I(r_k, t),$$

where $I(r_k, t)$ represents the probability that individual k, who is known to be removed at time r_k, was infective at time t, and may therefore be expressed as

$$I(r_k, t) = 1 - \int_t^{r_k} f_I(r_k - u)du \qquad \text{for } t < r_k,$$

and 0 otherwise. Recall that $f_I(\cdot \mid \boldsymbol{\theta})$ represents the probability density (mass) function of the infectious period distribution.

Then, the likelihood for any given case j with removal time r_j is given by

$$L_{\text{case}}(r_j) = \int_{-\infty}^{r_j} \Lambda_j(t) \exp\Big(- \int_{-\infty}^{t} \Lambda_j(u)du \Big) f_I(r_j - t)dt.$$

The first term $\Lambda_j(t)$ represents this likelihood of infection events, the second $\exp\Big(- \int_{-\infty}^{t} \Lambda_j(u)du \Big)$ represents the likelihood of the avoidance of infection and the terms $f_I(r_j - t)$ provide the densities of the infectious periods, in much the same way as the general likelihood format introduced in Section 1.3.5.

Similarly, the likelihood for any non-case j is given by

$$L_{\text{non}}(j) = \exp\Big(- \int_{-\infty}^{\infty} \Lambda_j(u)du \Big).$$

This is since all non-cases still received infectious pressure from infectives throughout the outbreak, although they were never infected. If $u > r_n$, this integrand will be zero.

Recalling that we do not include the infection of the initial infective, the combined ED likelihood of all observations is therefore given by

$$\pi_{\text{ED}}(\mathbf{r} \mid \beta, \boldsymbol{\theta}) = \left(\prod_{j=2}^{n} L_{\text{case}}(r_j) \right) \left(\prod_{j=n+1}^{N} L_{\text{non}}(j) \right).$$

This may easily be converted to the notation introduced in Section 3.2. Recalling the quantity $\chi_j = \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} \mathbb{1}_{\{k \text{ infective at } i_j\}}$ defined in Equation (3.2.1),

we may also more generally define $\chi_j(t) = \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} \mathbb{1}_{\{k \text{ infective at } t\}}$. (Note that $\chi_j(i_j) = \chi_j$). Then the quantity $\Lambda_j(t)$ is equivalent to $\mathbb{E}[\chi_j(t)]$: the expected infectious pressure acting on individual j at time t. Therefore,

$$\begin{aligned}
& \pi_{\text{ED}}(\mathbf{r} \mid \beta, \boldsymbol{\theta}) \\
= & \left(\prod_{j=2}^{n} \int_{-\infty}^{r_j} \mathbb{E}[\chi_j(t)] \exp\Big(- \int_{-\infty}^{t} \mathbb{E}[\chi_j(u)] du \Big) f_I(r_j - t) dt \right) \left(\prod_{j=n+1}^{N} L_{\text{non}}(j) \right) \\
= & \left(\prod_{j=2}^{n} \int_{-\infty}^{r_j} \mathbb{E}[\chi_j] \exp\Big(- \int_{-\infty}^{i_j} \mathbb{E}[\chi_j(u)] du \Big) f_I(r_j - i_j) di_j \right) \left(\prod_{j=n+1}^{N} L_{\text{non}}(j) \right).
\end{aligned}$$

Then, the term $\exp\Big(- \int_{-\infty}^{i_j} \mathbb{E}[\chi_j(u)] du \Big)$ can be approximated to the ψ_j term from Equation (3.2.1). Eichner and Dietz have made the approximation that

$$\begin{aligned}
\mathbb{E}[\psi_j \mid i_j] &= \mathbb{E}\left[\exp\left(- \sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{\beta}{N} (r_k \wedge i_j - i_k \wedge i_j) \right) \mid i_j \right] \\
&\approx \exp\left(-\mathbb{E}\left[\sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{\beta}{N} (r_k \wedge i_j - i_k \wedge i_j) \mid i_j \right] \right) \\
&= \exp\left(-\mathbb{E}\left[\int_{-\infty}^{i_j} \sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{\beta}{N} \mathbb{1}_{\{k \text{ infective at } u\}} \, du \right] \right) \\
&= \exp\left(- \int_{-\infty}^{i_j} \mathbb{E}[\chi_j(u)] \, du \right),
\end{aligned}$$

where instead of considering the value of $r_k \wedge i_j - i_k \wedge i_j$ (the amount of time there was infectious pressure between k and j) conditional on the value of i_j, we integrate over all time u up to i_j, considering if k is infective at each time. This may be considered as different ways of building the same quantity.

The final point to note is that, in term L_{non}, Eichner and Dietz consider the likelihood contributions from the perspective of the non-infected individuals. As in, the product over all individuals who do not become infected of the infectious pressure that was placed upon them. Under the notation of Section 3.2, we used $\mathbb{E}[\phi_j]$ where we considered the likelihood contribution from all infectives failing to infect non-cases instead. However, this is again two ways of expressing the same quantity.

So rather than writing this expression in terms of ϕ_j or L_{non} we define, for any individual l who avoids infection,

$$\begin{aligned}\rho_l &= \mathbb{P}(l \text{ avoids infection}) \\ &= \mathrm{e}^{-\frac{\beta}{N}(P_1+\cdots+P_n)} \\ &= \mathrm{e}^{-\frac{\beta}{N}\sum_{j=1}^{n} P_j}\end{aligned}$$

where $P_1, \ldots, P_n$ are the infectious periods $r_i - i_i$ of infectives $i = 1, \ldots, n$. Eichner and Dietz have then used the approximation

$$\mathbb{E}[\rho_l] = \mathbb{E}\big[\mathrm{e}^{-\frac{\beta}{N}\sum_{j=1}^{n} P_j}\big] \approx \mathrm{e}^{-\frac{\beta}{N}\sum_{j=1}^{n} \mathbb{E}[P_j]}.$$

Then

$$\prod_{j=1}^{n} \mathbb{E}[\phi_j] = \prod_{l=n+1}^{N} \mathbb{E}[\rho_l] = \prod_{l=n+1}^{N} L_{\text{non}}(j).$$

Overall, we obtain the Eichner and Dietz approximate likelihood

$$\pi_{\text{ED}}(\mathbf{r} \mid \beta, \boldsymbol{\theta}) = \Bigg(\prod_{j=2}^{n} \int_{-\infty}^{r_j} \mathbb{E}[\chi_j] \exp\Big(-\int_{-\infty}^{i_j} \mathbb{E}[\chi_j(u)]\, du\Big) f_I(r_j - i_j)\, di_j\Bigg)\Bigg(\prod_{l=n+1}^{N} \mathbb{E}[\rho_l]\Bigg), \tag{3.3.1}$$

where the first product represents the likelihood contribution from infected individuals, and the second product represents the contribution from those who were not infected. The two integrals contained in this likelihood might need to be evaluated numerically, depending on the choice of infectious period. It is important to note that this method is exact for constant infectious periods, but for any other choice of infectious periods it is an approximation, despite this not being explicitly mentioned in Eichner and Dietz (2003).

We will now explore the use of this method with both exponential and gamma distributed infectious periods, deriving likelihood expressions in each case.

3.3.2 Exponential Infectious Periods

We consider the special case of exponentially distributed infectious periods within the general Eichner and Dietz framework. As before, we use infection

rate $\frac{\beta}{N}$ between individuals, where now the infectious period is exponentially distributed with rate γ, and so has probability density function $f(x) = \gamma e^{-\gamma x}$, $x > 0$. All other aspects of the model are as defined in Section 3.2.

We define l_j as the contribution to the likelihood from any infected individual j (note that $l_j = L_{\text{case}}(r_j)$), so that

$$\pi_{\text{ED}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \Big(\prod_{j=2}^{n} l_j \Big) \Big(\prod_{l=n+1}^{N} \mathbb{E}[\rho_l] \Big).$$

Then l_j is given by

$$\begin{aligned} l_j &= \int_{-\infty}^{r_j} \mathbb{E}[\chi_j] e^{-\int_{-\infty}^{i_j} \mathbb{E}[\chi_j(u)]\, du} f_I(r_j - i_j)\, di_j \\ &= \int_{-\infty}^{r_j} \frac{\beta}{N} \sum_{\substack{k=1 \\ k \neq j}}^{n} (e^{-\gamma(r_k - t)} \mathbb{1}_{t < r_k}) \exp\Big(- \int_{-\infty}^{t} \frac{\beta}{N} \sum_{\substack{i=1 \\ i \neq j}}^{n} e^{-\gamma(r_i - u)} \mathbb{1}_{u < r_i} du \Big) \\ &\quad \times \gamma e^{-\gamma(r_j - t)} \mathbb{1}_{\{t < r_j\}} dt. \end{aligned}$$

Taking the central exponent alone, we may simplify to

$$\begin{aligned} \int_{-\infty}^{t} \frac{\beta}{N} \sum_{\substack{i=1 \\ i \neq j}}^{n} e^{-\gamma(r_i - u)} \mathbb{1}_{\{u < r_i\}} du &= \frac{\beta}{N} \sum_{\substack{i=1 \\ i \neq j}}^{n} e^{-\gamma r_i} \int_{-\infty}^{t} e^{\gamma u} \mathbb{1}_{\{u < r_i\}} du \\ &= \frac{\beta}{N} \sum_{\substack{i=1 \\ i \neq j}}^{n} e^{-\gamma r_i} \int_{-\infty}^{t \wedge r_i} e^{\gamma u} du \\ &= \frac{\beta}{\gamma N} \sum_{\substack{i=1 \\ i \neq j}}^{n} e^{-\gamma(r_i - t \wedge r_i)} \\ &= C_j(t), \text{ say.} \end{aligned}$$

Then,

$$\begin{aligned} l_j &= \frac{\beta\gamma}{N} \sum_{\substack{k=1 \\ k \neq j}}^{n} \int_{-\infty}^{r_j} e^{-\gamma(r_k - t)} \mathbb{1}_{\{t < r_k\}} e^{-C_j(t)} e^{-\gamma(r_j - t)} dt \\ &= \frac{\beta\gamma}{N} \sum_{\substack{k=1 \\ k \neq j}}^{n} e^{-\gamma(r_j + r_k)} \int_{-\infty}^{r_k \wedge r_j} e^{2\gamma t - C_j(t)} dt. \end{aligned}$$

To avoid evaluation of a large number of integrals, we reverse the order of the integral and summation for computational efficiency:

$$l_j = \frac{\beta\gamma}{N} \int_{-\infty}^{T} \sum_{\substack{k=1 \\ k\neq j}}^{n} \mathrm{e}^{-\gamma(r_j+r_k)+2\gamma t-C_j(t)} \mathbb{1}_{\{t<(r_k \wedge r_j)\}} dt, \tag{3.3.2}$$

where $T = r_n$ is the end of the outbreak, after which no infectious pressure is applied. This provides an expression for the likelihood contribution from an infective j which requires only one integration, although this is not analytically tractable and so must be calculated numerically.

Moving on to non-infectives, the probability that any non-infected individual j avoids infection is given by

$$\rho_j = \mathrm{e}^{-\frac{\beta}{N}\sum_{k=1}^{n} P_k},$$

where $P_k = r_k - i_k$ defines the infectious period of infective k. Then the likelihood contribution from uninfected j is given by

$$\hat{l}_j = \mathbb{E}[\rho_j] = \mathbb{E}\left[\mathrm{e}^{-\frac{\beta}{N}\sum_{k=1}^{n} P_k}\right] \approx \mathrm{e}^{-\frac{\beta}{N}\sum_{k=1}^{n} \mathbb{E}[P_k]} = \mathrm{e}^{-\frac{\beta n}{\gamma N}} \tag{3.3.3}$$

since $P_k \sim \mathrm{Exp}(\gamma)$ results in a mean infectious period of length $\frac{1}{\gamma}$ for any infective k, of which there are n.

Combining Equations (3.3.2) and (3.3.3), the overall likelihood expression using this approximation method is given by

$$\begin{aligned}
\pi_{\mathrm{ED}}(\mathbf{r} \mid \beta, \gamma) &= \left(\prod_{j=1}^{n} l_j\right)\left(\prod_{j=n+1}^{N} \hat{l}_j\right) \\
&= \left(\prod_{j=1}^{n} \frac{\beta\gamma}{N} \int_{-\infty}^{T} \sum_{\substack{k=1 \\ k\neq j}}^{n} \mathrm{e}^{-\gamma(r_j+r_k)+2\gamma t-C_j(t)} \mathbb{1}_{\{t<(r_k \wedge r_j)\}} dt\right) \\
&\quad \times \left(\prod_{j=n+1}^{N} \mathrm{e}^{-\frac{\beta n}{\gamma N}}\right) \\
&= \left(\prod_{j=1}^{n} \frac{\beta\gamma}{N} \int_{-\infty}^{T} \sum_{\substack{k=1 \\ k\neq j}}^{n} \mathrm{e}^{-\gamma(r_j+r_k)+2\gamma t-C_j(t)} \mathbb{1}_{\{t<(r_k \wedge r_j)\}} dt\right) \\
&\quad \times \left(\mathrm{e}^{-\frac{\beta n}{\gamma N}(N-n)}\right)
\end{aligned} \tag{3.3.4}$$

where

$$C_j(t) = \frac{\beta}{\gamma N} \sum_{\substack{i=1 \\ i \neq j}}^{n} \mathrm{e}^{-\gamma(r_i - (t \wedge r_i))}.$$

As discussed, the integral in the likelihood expression must be computed numerically, using a numerical integration technique of choice. In Chapter 4 we will perform a simulation study to assess the accuracy of the Eichner and Dietz method, for which we will use a simple trapezium rule method since this is found to provide sufficiently accurate results when compared with more complex techniques.

3.3.3 Gamma Infectious Periods

As we will explore in Chapter 4, the ED method with exponential infectious periods struggles for small outbreaks or those with large numbers of uninfected individuals, where the method does not appear to perform all that well compared to standard DA-MCMC. Hence, we also consider gamma distributed infectious periods. As the shape parameter increases we would expect the approximation to perform better, since we recall the method is exact for constant infectious periods.

We continue with infection rate $\frac{\beta}{N}$ between individuals, but now the infectious periods are gamma distributed with shape m and rate γ, and so have probability density function $f(x) = \frac{\gamma^m}{\Gamma(m)} x^{m-1} \mathrm{e}^{-\gamma x}$, $x > 0$. We restrict our attention to positive integer valued m, so that the distribution is in fact Erlang. Again, the remainder of the model is defined as in Section 3.2.

We begin with the likelihood expression from Equation (3.3.1). Substituting in the probability density function of the gamma distribution, the contribution to

the likelihood from any infected individual j is given by

$$\begin{aligned} l_j &= \int_{-\infty}^{r_j} \mathbb{E}[\chi_j] \mathrm{e}^{-\int_{-\infty}^{i_j} \mathbb{E}[\chi_j(u)]\, du} f_I(r_j - i_j)\, di_j \\ &= \int_{-\infty}^{r_j} \frac{\beta}{N} \sum_{\substack{k=1 \\ k\neq j}}^{n} \left(\mathrm{e}^{-\gamma(r_k - t)} \sum_{p=0}^{m-1} \frac{(\gamma(r_k - t))^p}{p!} \mathbb{1}_{\{t<r_k\}} \right) \frac{(r_j - t)^{m-1}}{\Gamma(m)} \gamma^m \mathrm{e}^{-\gamma(r_j - t)} \\ &\quad \times \exp\left(- \int_{-\infty}^{t} \frac{\beta}{N} \sum_{\substack{i=1 \\ i\neq j}}^{n} \mathrm{e}^{-\gamma(r_i - u)} \sum_{p=0}^{m-1} \frac{(\gamma(r_i - t))^p}{p!} \mathbb{1}_{\{u<r_i\}} du \right) dt. \end{aligned} \tag{3.3.5}$$

Taking the inner integral and changing the order of sums and integration, we have

$$\begin{aligned} &- \int_{-\infty}^{t} \frac{\beta}{N} \sum_{\substack{i=1 \\ i\neq j}}^{n} \mathrm{e}^{-\gamma(r_i - u)} \sum_{p=0}^{m-1} \frac{(\gamma(r_i - u))^p}{p!} \mathbb{1}_{\{u<r_i\}} du \\ &\qquad = -\frac{\beta}{N} \sum_{\substack{i=1 \\ i\neq j}}^{n} \mathrm{e}^{-\gamma r_i} \sum_{p=0}^{m-1} \frac{\gamma^p}{p!} \int_{-\infty}^{t\wedge r_i} \mathrm{e}^{\gamma u} (r_i - u)^p\, du. \end{aligned} \tag{3.3.6}$$

A change of variable on the inner integral shows that

$$\begin{aligned} \int_{-\infty}^{t\wedge r_i} \mathrm{e}^{\gamma u} (r_i - u)^p\, du &= \frac{\mathrm{e}^{\gamma r_i}}{\gamma^{p+1}} \int_{\gamma(r_i - (t\wedge r_i))}^{\infty} \mathrm{e}^{-y} y^p\, dy \\ &= \frac{\mathrm{e}^{\gamma r_i}}{\gamma^{p+1}} \Big(1 - \int_{-\infty}^{\gamma(r_i - (t\wedge r_i))} \mathrm{e}^{-y} y^p\, dy \Big), \end{aligned} \tag{3.3.7}$$

where $y = \gamma(r_i - u)$. The integral is in the form of a gamma cumulative distribution function (CDF) $F_{a,b}(y)$, where $a = p + 1$ and $b = 1$, evaluated at $t = \gamma(r_i - (t \wedge r_i))$. For $Y \sim \Gamma(a, b)$ it is known that for integer $a \geq 1$,

$$F_{a,b}(y) = 1 - \mathrm{e}^{-by} \sum_{q=0}^{a-1} \frac{1}{q!} (by)^q,$$

which may be applied here since p takes integer values only. Hence, Equation (3.3.7) becomes

$$\frac{\mathrm{e}^{\gamma r_i}}{\gamma^{p+1}} \Big(1 - \int_{-\infty}^{\gamma(r_i - (t\wedge r_i))} \mathrm{e}^{-y} y^p\, dy \Big) = \frac{\mathrm{e}^{\gamma r_i}}{\gamma^{p+1}} p!\, \mathrm{e}^{-\gamma(r_i - (t\wedge r_i))} \sum_{q=0}^{p} \frac{(\gamma(r_i - (t \wedge r_i)))^q}{q!},$$

and Equation (3.3.6) can be expressed as

$$
\begin{aligned}
& -\frac{\beta}{N}\sum_{\substack{i=1\\ i\neq j}}^{n} \mathrm{e}^{-\gamma r_i} \sum_{p=0}^{m-1} \frac{\gamma^p}{p!} \frac{\mathrm{e}^{\gamma r_i}}{\gamma^{p+1}} p!\, \mathrm{e}^{-\gamma(r_i-(t\wedge r_i))} \sum_{q=0}^{p} \frac{(\gamma(r_i-(t\wedge r_i)))^q}{q!} \\
= \; & -\frac{\beta}{\gamma N}\sum_{\substack{i=1\\ i\neq j}}^{n} \mathrm{e}^{-\gamma(r_i-(t\wedge r_i))} \sum_{p=0}^{m-1}\sum_{q=0}^{p} \frac{(\gamma(r_i-(t\wedge r_i)))^q}{q!} \\
= \; & -\frac{\beta}{\gamma N}\sum_{\substack{i=1\\ i\neq j}}^{n} \mathrm{e}^{-\gamma(r_i-(t\wedge r_i))} \sum_{p=0}^{m-1} \frac{(\gamma(r_i-(t\wedge r_i)))^p}{p!}(m-p) \\
= \; & -D_j(t), \text{ say.}
\end{aligned}
$$

Overall, we obtain for Equation (3.3.5)

$$
\begin{aligned}
l_j \; = \; & \frac{\beta\gamma^m}{N\Gamma(m)} \int_{-\infty}^{r_j} \sum_{\substack{i=1\\ i\neq j}}^{n} \left(\mathrm{e}^{-\gamma(r_i-t)} \sum_{p=0}^{m-1} \frac{(\gamma(r_i-t))^p}{p!} \mathbb{1}_{\{t<r_i\}} \right) \mathrm{e}^{-D_j(t)} (r_j-t)^{m-1} \\
& \times \mathrm{e}^{-\gamma(r_j-t)} \; dt \\
= \; & \int_{-\infty}^{T} \sum_{p=0}^{m-1} \frac{\beta\gamma^{m+p}}{N\Gamma(m)p!} \sum_{\substack{i=1\\ i\neq j}}^{n} \mathrm{e}^{-\gamma(r_j+r_i)+2\gamma t-D_j(t)} (r_i-t)^p (r_j-t)^{m-1} \\
& \times \mathbb{1}_{\{t<(r_j\wedge r_i)\}} \; dt,
\end{aligned}
$$

where the integral and sums have been ordered for computational efficiency. As in the exponential case, this integral cannot be analytically evaluated.

The likelihood contribution for uninfected individuals is given by

$$
\begin{aligned}
\hat{l}_j \; &= \; \mathbb{E}[\rho_j] = \mathbb{E}[\mathrm{e}^{-\frac{\beta}{N}\sum_{k=1}^{n} P_k}] \\
&\approx \; \mathrm{e}^{\frac{\beta}{N}\sum_{k=1}^{n} \mathbb{E}[P_k]} = \mathrm{e}^{\frac{\beta}{N}\sum_{k=1}^{n} \frac{m}{\gamma}} \\
&= \; \mathrm{e}^{\frac{-\beta n m}{N\gamma}},
\end{aligned}
$$

since $P_k \sim \Gamma(m,\gamma)$ results in a mean infectious period of length $\frac{m}{\gamma}$ for any infective k, of which there are n .

The overall likelihood expression with the ED approximation method for gamma

distributed infectious periods is given by

$$\begin{aligned}
\pi_{\mathrm{ED}}(\mathbf{r} \mid \beta, \gamma) &= \left(\prod_{j=1}^{n} l_j\right)\left(\prod_{j=n+1}^{N} \hat{l}_j\right) \\
&= \left(\prod_{j=1}^{n} \int_{-\infty}^{T} \sum_{p=0}^{m-1} \frac{\beta\gamma^{m+p}}{N\Gamma(m)p!} \sum_{\substack{i=1 \\ i\neq j}}^{n} \mathrm{e}^{-\gamma(r_j+r_i)+2\gamma t-D_j(t)}(r_i-t)^p(r_j-t)^{m-1}\right. \\
&\quad \left.\times \mathbb{1}_{\{t<(r_j\wedge r_i)\}}\ dt\right)\left(\prod_{j=n+1}^{N} \mathrm{e}^{-\frac{\beta n m}{N\gamma}}\right) \\
&= \left(\prod_{j=1}^{n} \int_{-\infty}^{T} \sum_{p=0}^{m-1} \frac{\beta\gamma^{m+p}}{N\Gamma(m)p!} \sum_{\substack{i=1 \\ i\neq j}}^{n} \mathrm{e}^{-\gamma(r_j+r_i)+2\gamma t-D_j(t)}(r_i-t)^p(r_j-t)^{m-1}\right. \\
&\quad \left.\times \mathbb{1}_{\{t<(r_j\wedge r_i)\}}\ dt\right)\left(\mathrm{e}^{-\frac{\beta n m}{N\gamma}(N-n)}\right),
\end{aligned}$$

where

$$D_j(t) = \frac{\beta}{\gamma N} \sum_{\substack{i=1 \\ i\neq j}}^{n} \mathrm{e}^{-\gamma(r_i-(t\wedge r_i))} \sum_{l=0}^{m-1} \frac{(\gamma(r_i-(t\wedge r_i)))^l}{l!}(m-l).$$

As in the exponential case, the integral in the likelihood must be calculated numerically. We again will use the trapezium rule for this, since more complex methods were found to provide only very limited improvement in accuracy.

3.3.4 Heterogeneous mixing and non-identically distributed infectious periods

So far we have focused on populations which are assumed to be homogeneously mixing and contain only individuals with identically distributed infectious periods. However, the Eichner and Dietz approximation method may be extended beyond this. In this section, we will provide likelihood expressions for the Eichner and Dietz approximation which allow for a heterogeneously mixing population, as well as individuals with different infectious period parameters.

We assume now that the contact rate from any individual j to any individual k is given by β_{jk}. We define $\boldsymbol{\beta} = \{\beta_{jk} : j,k \in 1,\ldots,N\}$ as the complete set of these contact rates. This allows for a heterogeneously mixing population model with any structure desired. We assume that each infected individual has the same infectious period distribution, but now allowing for different parameters. For example, $r_j - i_j \sim \text{Exp}(\gamma_j)$ or $r_j - i_j \sim \Gamma(m_j, \gamma_j)$ for all infected individuals j.

We will not repeat the likelihood calculations here since they are very similar to the homogeneous case, but the resulting likelihood expressions are as follows.

Exponential Infectious Periods

We assume that all infectious individuals j have infectious periods $r_j - i_j \sim \text{Exp}(\gamma_j)$, and define $\boldsymbol{\theta} = \{\gamma_j : j \in 1,\ldots,n\}$. The ED likelihood is then given by

$$\begin{aligned}\pi_{\text{ED}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) &= \left(\prod_{j=1}^{n} l_j\right)\left(\prod_{j=n+1}^{N} \hat{l}_j\right) \\ &= \left(\prod_{j=1}^{n}\int_{-\infty}^{T}\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\gamma_j \mathrm{e}^{-\gamma_k(r_k-t)-\gamma_j(r_j-t)-C_j(t)}\mathbb{1}_{\{t<(r_k\wedge r_j)\}}dt\right) \\ &\quad\times\left(\prod_{j=n+1}^{N}\exp\Big(-\sum_{k=1}^{n}\frac{\beta_{kj}}{\gamma_k}\Big)\right),\end{aligned}$$

where

$$C_j(t) = \sum_{\substack{i=1\\i\neq j}}^{n}\frac{\beta_{ij}}{\gamma_i}\mathrm{e}^{-\gamma_i(r_i-(t\wedge r_i))}.$$

Gamma Infectious Periods

For the Gamma case, we assume that all infectious individuals j have infectious periods $r_j - i_j \sim \Gamma(m_j, \gamma_j)$, and define $\boldsymbol{\theta} = \{\gamma_j, m_j : j \in 1,\ldots,n\}$. The ED likelihood is given by

$$\pi_{\mathrm{ED}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \left(\prod_{j=1}^{n} l_j \right) \left(\prod_{j=n+1}^{N} \hat{l}_j \right)$$
$$= \left(\prod_{j=1}^{n} \int_{-\infty}^{T} \sum_{\substack{k=1 \\ k \neq j}}^{n} \sum_{l=0}^{m_k - 1} \frac{\beta_{kj} \gamma_j^{m_j + l}}{\Gamma(m_j) l!} \mathrm{e}^{-\gamma_k (r_k - t) - (\gamma_j (r_j - t)) - D_j(t)} (r_k - t)^l (r_j - t)^{m_j - 1} \right.$$
$$\left. \times \mathbb{1}_{\{t < (r_j \wedge r_k)\}} \, dt \right) \left(\prod_{j=n+1}^{N} \exp\left(\sum_{k=1}^{n} \beta_{kj} \frac{m_k}{\gamma_k} \right) \right),$$

where

$$D_j(t) = \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{\beta_{ij}}{\gamma_i} \mathrm{e}^{-\gamma_i (r_i - (t \wedge r_i))} \sum_{l=0}^{m_i - 1} \frac{(\gamma_i (r_i - (t \wedge r_i)))^l}{l!} (m_i - l).$$

Although we will not explore the computational implementation of this extension to the method, the integrals in these likelihoods may be numerically integrated as before and then used, for example, for maximum likelihood estimation.

3.3.5 Conclusions

We have defined the Eichner and Dietz likelihood approximation for both exponential and gamma infectious periods, with a particular focus on homogeneously mixing populations with identically distributed infectious periods but also extending the theory to heterogeneously mixing populations with non-identical infectious periods. Chapter 4 will include analysis of the method through simulation studies, and we will find that the method performs fairly well for gamma distributed infectious periods, though less well for exponential periods or outbreaks in large populations with only a very small or very large proportion of infectives. The method also relies on numerical integration as the likelihood expressions cannot be analytically calculated, and this can be relatively slow to compute. Although the ED method would be useful in some situations, it would be beneficial to develop further approximation

methods which are more widely applicable as well as offering an increase in computational speed.

3.4 Pair-Based Likelihood Approximations

In this section we develop a series of new approximation methods which seek to avoid numerical integration as well as offer increased performance in a wider range of situations than the Eichner and Dietz method. We term these Pair-Based Likelihood Approximations (PBLA) since they will essentially consider the contribution to the likelihood from different pairs of individuals as independent, resulting in an approximate (but tractable) likelihood. These likelihoods will require no data augmentation or numerical integration to calculate: MCMC or maximum likelihood estimation may be performed directly, without the need for these further, potentially computationally costly, steps.

We will derive the first PBLA method in Section 3.4.1, which further versions will extend upon. In Section 3.4.2 we will apply PBLA I to exponential infectious periods and analytically derive the resulting likelihood expressions, and then in Section 3.4.3 we will do the same for gamma infectious periods. Section 3.4.4 similarly obtains these likelihood expressions, but using probabilistic arguments which provide more insight to the calculations. We then proceed to define and derive further PBLA methods which we number accordingly: PBLA II in Section 3.4.5 and PBLA III in Section 3.4.6. Following this, we define two further PBLA versions which offer increased computational speed, but may only be used in more specific situations. Section 3.4.8 describes a PBLA method which uses a central limit theorem to make further approximations to the likelihood, but which requires homogeneous mixing and exponentially distributed infectious periods. Section 3.4.9 then describes the PBLA V method, which takes a further step in grouping the infectious pressure between individuals, but will require that the pressure from any individual j to k is equal to the pressure from k to j. After describing a numerical limitation

of the PBLA approach in Section 3.4.10, we then describe an extension of the PBLA approach to the SEIR model in Section 3.4.11.

Since the PBLA methods as well as the ED approximation involve considerable amounts of notation, refer to Table 3.1 for a summary of this. Some of these quantities have already been defined, and some are to come in the following sections.

3.4.1 PBLA I: General Framework

For the PBLA method, we will work again under the general model and likelihood format defined in Section 3.2. Recall that the population is of size N with individuals labelled $1, 2, \ldots, N$, of which $1, 2, \ldots, n$ become infected where $n \leq N$. The contact rate from individual i to individual j is given by β_{ij}, contained in $\boldsymbol{\beta}$ and allowing for heterogeneous mixing, and the infectious periods have length with distribution $f_I(\cdot \mid \boldsymbol{\theta})$. Infection times $\mathbf{i} = \{i_j : j = 1, 2, \ldots, \kappa - 1, \kappa + 1, \ldots, n\}$ are unknown, and removal times $\mathbf{r} = \{r_j : j = 1, 2, \ldots, n, \text{ where } r_1 < r_2 < \cdots < r_n\}$, form the data. We work with labelled cases, such that i_j and r_j are the infection and removal time for individual j, respectively, for all infectives j.

We will now derive the PBLA I likelihood expression, beginning with the approximate likelihood expression from Equation (3.2.3);

$$\pi(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) \approx \sum_{\kappa=1}^{n} \pi(\kappa) \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \mathop{\mathbb{E}}_{\mathbf{i}, i_\kappa} \left[\chi_j \psi_j \phi_j \right] \right) \mathop{\mathbb{E}}_{i_\kappa} \left[\phi_\kappa \pi(i_\kappa \mid \kappa) \right],$$

where we recall that the expectations are with respect to infection times $\mathbf{i}$ and i_κ, with κ being the initial infective.

As in the ED method, we make approximations to this likelihood in order to find a tractable expression. The first key assumption is that

$$\mathbb{E}[\chi_j \psi_j \phi_j] \approx \mathbb{E}[\chi_j \phi_j] \mathbb{E}[\psi_j].$$

Table 3.1: Table of commonly-used notation for the likelihood approximation methods, as used in chapters 3 and 4.

Quantity	Definition	Description
χ_j	$\sum_{\substack{k=1 \\ k\neq j}}^{n} \beta_{kj} \mathbb{1}_{\{k \text{ infective at } i_j\}}$	Infectious pressure acting on j at their time of infection
ψ_j	$\exp\Big(-\sum_{\substack{k=1 \\ k\neq j}}^{n} \beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)\Big)$	$\mathbb{P}(j$ avoids infection until $i_j)$
ϕ_j	$\exp\Big(-\sum_{k=n+1}^{N} \beta_{jk}(r_j - i_j)\Big)$	$\mathbb{P}(j$ fails to infect all non-infected individuals)
B_j	$\sum_{l=n+1}^{N} \beta_{jl}$	Sum of infection rates from j to all non-infectives
δ_j	$\gamma + B_j$	Change-of-variables quantity for PBLA III
$F_{k,\theta}(x)$	$1 - \sum_{l=0}^{k-1} \frac{1}{l!}(\theta x)^l \mathrm{e}^{-\theta x}$	
$\mathbb{E}\big[(r+\mathrm{X})^l \mid \mathrm{X} \sim \Gamma(m,\gamma)\big]$	$\sum_{p=0}^{l} \binom{l}{p} r^{l-p} \frac{(m+p-1)_p}{\gamma^p}$	
$(x)_p$	$\binom{x}{p} p!$	
$a(B_j, \boldsymbol{\theta})$	$\begin{cases} \frac{\gamma}{\delta} & \text{if } r_j - i_j \sim \mathrm{Exp}(\gamma) \\ \left(\frac{\gamma}{\delta}\right)^m & \text{if } r_j - i_j \sim \Gamma(m,\gamma), \end{cases}$	MGF of the infectious period of j, evaluated at B_j
τ_{kj}	$r_k \wedge i_j - i_k \wedge i_j$	Time when there is infectious pressure from k to j
ω_{jk}	$\tau_{jk} + \tau_{kj}$	Time when there is infectious pressure between j and k, for $r_k < r_j$

Recalling the definitions of these terms from Equation (3.2.1),

$$\mathbb{E}[\chi_j \phi_j] = \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} \mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} e^{-\sum_{l=n+1}^{N} \beta_{jl}(r_j - i_j)}\right] \tag{3.4.1}$$

$$\begin{aligned} \mathbb{E}[\psi_j] &= \mathbb{E}\left[\exp\left(-\sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)\right)\right] \\ &\approx \prod_{\substack{k=1 \\ k \neq j}}^{n} \mathbb{E}\left[e^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right], \end{aligned} \tag{3.4.2}$$

where the final approximation describes our second key assumption.

These assumptions seem most reasonable for infectious periods with low variance, since if the infection times were known the expectations would reduce to known values. As discussed in Section 3.2, we would also expect the approximations to perform better when the appearance of new infectives is relatively slow compared the expected length of the infectious periods, since we will have less uncertainty about which individuals placed infectious pressure upon which others. Compared to the ED method, our motivation has been to make additional approximations to the likelihood here, in order to find an expression which is not just tractable but also does not require numerical integration. In the following sections, we will calculate these likelihood expressions for various models and infectious period distributions.

SIR model with homogeneous mixing

If we consider instead the simple case of an SIR model in a homogeneously mixing population, we define contact rate $\beta_{ij} = \frac{\beta}{N}$ for all $i, j \in \{1, \ldots, N\}$.

This simplifies the expressions for the components $\mathbb{E}[\chi_j \phi_j]$ (Equation (3.4.1))

and $\mathbb{E}[\psi_j]$ (Equation (3.4.2)) to

$$\begin{aligned}\mathbb{E}[\chi_j\phi_j] &= \sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}}\mathrm{e}^{-\sum_{l=n+1}^{N}\beta_{jl}(r_j-i_j)}\right]\\ &= \frac{\beta}{N}\sum_{\substack{k=1\\k\neq j}}^{n}\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}}\mathrm{e}^{-(N-n)\frac{\beta}{N}(r_j-i_j)}\right]\\ \mathbb{E}[\psi_j] &\approx \prod_{\substack{k=1\\k\neq j}}^{n}\mathbb{E}\left[\mathrm{e}^{-\frac{\beta}{N}(r_k\wedge i_j-i_k\wedge i_j)}\right].\end{aligned}$$

Overall, with either homogeneous or heterogeneous mixing, we obtain the PBLA I likelihood

$$\pi_{\mathrm{I}}(\mathbf{r}\mid\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{\kappa=1}^{n}\pi(\kappa)\left(\prod_{\substack{j=1\\j\neq\kappa}}^{n}\mathbb{E}[\chi_j\phi_j]\mathbb{E}[\psi_j]\right)\mathbb{E}[\phi_\kappa\pi(i_\kappa\mid\kappa)], \tag{3.4.3}$$

which, for computational speed, we may calculate as

$$\pi_{\mathrm{I}}(\mathbf{r}\mid\boldsymbol{\beta},\boldsymbol{\theta}) = \left(\prod_{j=1}^{n}\mathbb{E}[\chi_j\phi_j]\mathbb{E}[\psi_j]\right)\sum_{\kappa=1}^{n}\frac{\pi(\kappa)\mathbb{E}[\phi_\kappa\pi(i_\kappa\mid\kappa)]}{\mathbb{E}[\chi_\kappa\phi_\kappa]\mathbb{E}[\psi_\kappa]}, \tag{3.4.4}$$

since this expression no longer requires the calculation of the product term for each possible κ, which is the most computationally demanding part. We see that the likelihood is written as an independent product over infectives.

As was mentioned in Section 3.1, the PBLA approach is somewhat similar to the method of composite likelihoods, as often used in geostatistics and genetics (see e.g. Fronterrè et al., 2017 and Larribe and Fearnhead, 2011). This technique involves multiplying together a collection of component or marginal likelihoods to act as an estimator of the true likelihood. Although similar in spirit to PBLA, the methods are in practice quite different. To demonstrate this, we take a simple example with $n = 3$ infectives. We assume knowledge of the initial infective as $\kappa = 1$ (i.e. $\pi(1) = 1$, $\pi(2) = \pi(3) = \cdots = 0$) for simplicity, and set $\pi(i_1) = \mathbb{1}_{\{i_1<r_1\}}$; an improper uniform distribution over the

region $(-\infty, r_1)$. Then, the PBLA likelihood will be of the form:

$$\begin{aligned}
\pi_{\mathrm{I}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) &= \prod_{j=2}^{3} \mathbb{E}[\chi_j \phi_j] \mathbb{E}[\psi_j] \mathbb{E}_{i_1}[\phi_1 \pi(i_1)] \\
&= \prod_{j=2}^{3} \left(\sum_{\substack{k=1 \\ k \neq j}}^{3} \beta_{kj} \mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-\sum_{l=n+1}^{N} \beta_{jl}(r_j - i_j)} \right] \right) \mathbb{E}[\psi_j] \\
&\quad \times \mathbb{E}_{i_1}\left[\mathrm{e}^{-\sum_{l=n+1}^{N} \beta_{1l}(r_1 - i_1)} \mathbb{1}_{\{i_1 < r_1\}} \right] \\
&= \Bigg(\mathbb{E}[\psi_2] \Bigg(\beta_{21} \mathbb{E}\left[\mathbb{1}_{\{1 \text{ infective at } i_2\}} \mathrm{e}^{-\sum_{l=n+1}^{N} \beta_{2l}(r_2 - i_2)} \right] \\
&\qquad + \beta_{23} \mathbb{E}\left[\mathbb{1}_{\{3 \text{ infective at } i_2\}} \mathrm{e}^{-\sum_{l=n+1}^{N} \beta_{2l}(r_2 - i_2)} \right] \Bigg) \\
&\quad + \mathbb{E}[\psi_3] \Bigg(\beta_{31} \mathbb{E}\left[\mathbb{1}_{\{1 \text{ infective at } i_3\}} \mathrm{e}^{-\sum_{l=n+1}^{N} \beta_{3l}(r_3 - i_3)} \right] \\
&\qquad + \beta_{32} \mathbb{E}\left[\mathbb{1}_{\{2 \text{ infective at } i_3\}} \mathrm{e}^{-\sum_{l=n+1}^{N} \beta_{3l}(r_3 - i_3)} \right] \Bigg) \Bigg) \\
&\quad \times \mathbb{E}_{i_1}\left[\mathrm{e}^{-\sum_{l=n+1}^{N} \beta_{1l}(r_1 - i_1)} \mathbb{1}_{\{i_1 < r_1\}} \right]. \qquad (3.4.5)
\end{aligned}$$

Due to the pair-based structure of our approximation, the pairwise likelihood of Cox and Read (2004) bears the most resemblance to it of the different composite methods. The Cox and Read likelihood is written as a double product over all pairs of observations. For example, for observations y_j where $j = 1, 2, \ldots, n$ with PDFs $f(y; \theta)$, the pairwise likelihood is given by

$$\pi_{\mathrm{pair}}(\theta; y) = \prod_{j=1}^{n-1} \prod_{k=j+1}^{n} f(y_j, y_k; \theta).$$

Hence, for the PBLA likelihood with $n = 3$ to be equivalent to a pairwise composite likelihood, we would need to be able to write it in the form

$$\pi_{\mathrm{pair}}(\theta; y) = \prod_{j=1}^{2} \prod_{k=2}^{3} f(y_j, y_k; \theta),$$

where each f represents the contribution to the likelihood for a given pair j, k. Returning to Equation (3.4.5), we see that, despite the outer product over j, the PBLA likelihood involves a sum over different pairs of infectives, and cannot be written as a double product as required. Although the idea of PBLA is

similar to that of composite likelihoods in breaking down the likelihood into contributions from individual pairs, in practice the two methods are certainly distinct.

Returning to the PBLA likelihood calculations, for a given infectious period distribution, we may now explicitly calculate the likelihood, via calculation of $\mathbb{E}[\chi_j\phi_j]$ and $\mathbb{E}[\psi_j]$. We will explore the use of exponential and gamma distributions for the infectious periods, for which the likelihood expression will be analytically tractable. In terms of the initial infective and their infection time i_κ, we may select any prior distribution, including the improper distribution used in the previous example.

3.4.2 PBLA I: Likelihood Calculations for Exponential Infectious Periods

We first consider the case of exponentially distributed infectious periods, so that $f_I(r_j - i_j \mid \gamma) = \gamma e^{-\gamma(r_j - i_j)}$. The likelihood as given in Equation (3.4.3) requires the calculation of two expressions; $\mathbb{E}[\chi_j\phi_j]$ and $\mathbb{E}[\psi_j]$, details of which we will provide here.

3.4.2.1 Expression one: $\mathbb{E}[\chi_j\phi_j]$

We recall that

$$\mathbb{E}[\chi_j\phi_j] = \sum_{\substack{k=1\\k\neq j}}^{n} \beta_{kj}\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} e^{-\sum_{l=n+1}^{N}\beta_{jl}(r_j - i_j)}\right],$$

so that for any given j and k we must calculate $\mathbb{E}[\mathbb{1}_{\{k \text{ infective at } i_j\}} e^{-B_j(r_j - i_j)}]$, where

$$B_j = \sum_{l=n+1}^{N} \beta_{jl}. \tag{3.4.6}$$

Now,

$$\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} e^{-B_j(r_j - i_j)}\right] = \int_{-\infty}^{r_k}\int_{-\infty}^{r_j} \mathbb{1}_{\{i_k < i_j < r_k\}} e^{-B_j(r_j - i_j)} f_I(r_j - i_j) f_I(r_k - i_k)\ di_j\ di_k,$$

assuming that i_j and i_k are independent. Then for a given j and k, this expression will take one of two forms, determined by the values of r_k and r_j.

Case (i): $r_k \geq r_j$

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j-i_j)}\right] \\
&\quad= \int_{-\infty}^{r_j} \int_{i_k}^{r_j} \mathrm{e}^{-B_j(r_j-i_j)} \gamma \mathrm{e}^{-\gamma(r_j-i_j)} \gamma \mathrm{e}^{-\gamma(r_k-i_k)} \; di_j \; di_k \\
&\quad= \int_{-\infty}^{r_j} \gamma^2 \mathrm{e}^{-B_j r_j} \mathrm{e}^{-\gamma(r_k-i_k)} \mathrm{e}^{-\gamma r_j} \left(\int_{i_k}^{r_j} \mathrm{e}^{(\gamma+B_j)i_j} \; di_j \right) \; di_k \\
&\quad= \int_{-\infty}^{r_j} \gamma^2 \mathrm{e}^{-B_j r_j} \mathrm{e}^{-\gamma(r_k-i_k)} \mathrm{e}^{-\gamma r_j} \left(\frac{\mathrm{e}^{(\gamma+B_j)r_j}}{\gamma+B_j} - \frac{\mathrm{e}^{(\gamma+B_j)i_k}}{\gamma+B_j} \right) \; di_k \\
&\quad= \frac{\gamma}{\gamma+B_j} \int_{-\infty}^{r_j} \mathrm{e}^{-B_j r_j} \mathrm{e}^{-\gamma r_k} \mathrm{e}^{-\gamma r_j} \gamma \mathrm{e}^{\gamma i_k} \left(\mathrm{e}^{(\gamma+B_j)r_j} - \mathrm{e}^{(\gamma+B_j)i_k} \right) \; di_k \\
&\quad= \frac{\gamma}{\gamma+B_j} \mathrm{e}^{-B_j r_j} \mathrm{e}^{-\gamma(r_k+r_j)} \left(\mathrm{e}^{(\gamma+B_j)r_j} \mathrm{e}^{\gamma r_j} - \frac{\gamma}{2\gamma+B_j} \mathrm{e}^{(2\gamma+B_j)r_j} \right) \\
&\quad= \frac{\gamma}{\gamma+B_j} \mathrm{e}^{-B_j r_j} \mathrm{e}^{-\gamma(r_k+r_j)} \mathrm{e}^{(2\gamma+B_j)r_j} \left(1 - \frac{\gamma}{2\gamma+B_j}\right) \\
&\quad= \frac{\gamma}{2\gamma+B_j} \mathrm{e}^{-\gamma(r_k-r_j)}. \qquad (3.4.7)
\end{aligned}$$

Case (ii): $r_k < r_j$

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j-i_j)}\right] \\
&\quad= \int_{-\infty}^{r_k} \int_{i_k}^{r_k} \mathrm{e}^{-B_j(r_j-i_j)} \gamma \mathrm{e}^{-\gamma(r_j-i_j)} \gamma \mathrm{e}^{-\gamma(r_k-i_k)} \; di_j \; di_k
\end{aligned}$$

which proceeds as in the $r_k \geq r_j$ case to

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j-i_j)}\right] \\
&\quad= \frac{\gamma^2}{\gamma+B_j} \int_{-\infty}^{r_k} \mathrm{e}^{-B_j r_j} \mathrm{e}^{-\gamma(r_k-i_k)} \mathrm{e}^{-\gamma r_j} \left(\mathrm{e}^{(B_j+\gamma)r_k} - \mathrm{e}^{(B_j+\gamma)i_k} \right) \; di_k \\
&\quad= \frac{\gamma}{B_j+\gamma} \mathrm{e}^{-B_j r_j} \mathrm{e}^{-\gamma(r_k+r_j)} \int_{-\infty}^{r_k} \gamma \mathrm{e}^{\gamma i_k} \left(\mathrm{e}^{(B_j+\gamma)r_k} - \mathrm{e}^{(B_j+\gamma)i_k} \right) \; di_k \\
&\quad= \frac{\gamma}{B_j+\gamma} \mathrm{e}^{-B_j r_j} \mathrm{e}^{-\gamma(r_k+r_j)} \left(\mathrm{e}^{(B_j+\gamma)r_k} \mathrm{e}^{\gamma r_k} - \frac{\gamma}{2\gamma+B_j} \mathrm{e}^{(2\gamma+B)r_k} \right) \\
&\quad= \frac{\gamma}{2\gamma+B_j} \mathrm{e}^{-B_j(r_j-r_k)} \mathrm{e}^{-\gamma(r_j-r_k)}. \qquad (3.4.8)
\end{aligned}$$

Then, combining the two cases $r_k \geq r_j$ and $r_k < r_j$,

$$\mathbb{E}[\chi_j \phi_j] = \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} \frac{\gamma}{2\gamma + B_j} \mathrm{e}^{-\gamma |r_k - r_j| - B_j((r_j - r_k) \vee 0)},$$

where $a \vee b$ represents the maximum of a and b.

3.4.2.2 Expression two: $\mathbb{E}[\psi_j]$

For this expression, recall that

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1 \\ k \neq j}}^{n} \mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right].$$

Any given term in this product will take the form

$$\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right] = \int_{-\infty}^{r_k} \int_{-\infty}^{r_j} \mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)} f_I(r_k - i_k) f_I(r_j - i_j) \; di_j \; di_k.$$

The integral must then again be split into sections, dependent upon the ordering of r_j and r_k. In the calculations that follow, we will write $\beta_{kj} = \beta$ for simplicity.

Case (i): $r_k \geq r_j$

In this case,

$$r_k \wedge i_j - i_k \wedge i_j = \begin{cases} i_j - i_k & \text{if } i_k < i_j < r_k, \\ 0 & \text{otherwise,} \end{cases}$$

and so

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\beta(r_k \wedge i_j - i_k \wedge i_j)} \mid r_k \geq r_j\right] &= \int_{-\infty}^{r_j} \int_{i_k}^{r_j} \mathrm{e}^{-\beta(i_j - i_k)} f_I(r_j - i_j) f_I(r_k - i_k) \; di_j \; di_k \\ &\quad + \int_{r_j}^{r_k} \int_{-\infty}^{r_j} 1 f_I(r_j - i_j) f_I(r_k - i_k) \; di_j \; di_k \\ &\quad + \int_{-\infty}^{r_j} \int_{-\infty}^{i_k} 1 f_I(r_j - i_j) f_I(r_k - i_k) \; di_j \; di_k, \end{aligned}$$

where we have assumed i_j and i_k are independent, and will calculate each integral separately.

(i) To begin,

$$\begin{aligned}
\int_{-\infty}^{r_j}\int_{i_k}^{r_j} \mathrm{e}^{-\beta(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_j\ di_k & \\
&= \int_{-\infty}^{r_j}\int_{i_k}^{r_j} \mathrm{e}^{-\beta(i_j-i_k)}\gamma\mathrm{e}^{-\gamma(r_j-i_j)}\gamma\mathrm{e}^{-\gamma(r_k-i_k)}\ di_j\ di_k \\
&= \int_{-\infty}^{r_j} \gamma^2\mathrm{e}^{-\gamma(r_k-i_k)}\left[\frac{\mathrm{e}^{-\beta(i_j-i_k)-\gamma(r_j-i_j)}}{\gamma-\beta}\right]_{i_k}^{r_j} di_k \\
&= \int_{-\infty}^{r_j} \frac{\gamma^2}{\gamma-\beta}\mathrm{e}^{-\gamma(r_k-i_k)}\left(\mathrm{e}^{-\beta(r_j-i_k)}-\mathrm{e}^{-\gamma(r_j-i_k)}\right) di_k \\
&= \frac{\gamma^2}{\gamma-\beta}\left[\frac{\mathrm{e}^{-\gamma(r_k-i_k)-\beta(r_j-i_k)}}{\gamma+\beta}-\frac{\mathrm{e}^{-\gamma(r_k+r_j-2i_k)}}{2\gamma}\right]_{-\infty}^{r_j} \\
&= \frac{\gamma^2}{\gamma-\beta}\left(\frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{\gamma+\beta}-\frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{2\gamma}\right) \\
&= \frac{\gamma}{\gamma+\beta}\frac{1}{2}\mathrm{e}^{-\gamma(r_k-r_j)}.
\end{aligned}$$

(ii) The second integral is equal to

$$\begin{aligned}
\int_{r_j}^{r_k}\int_{-\infty}^{r_j} 1 f_I(r_j-i_j) f_I(r_k-i_k)\ di_j\ di_k &= \int_{r_j}^{r_k}\int_{-\infty}^{r_j} \gamma\mathrm{e}^{-\gamma(r_j-i_j)}\gamma\mathrm{e}^{-\gamma(r_k-i_k)}\ di_j\ di_k \\
&= \int_{r_j}^{r_k} \gamma^2\mathrm{e}^{-\gamma(r_k-i_k)}\left[\frac{1}{\gamma}\mathrm{e}^{-\gamma(r_j-i_j)}\right]_{-\infty}^{r_j} di_k \\
&= \int_{r_j}^{r_k} \gamma\mathrm{e}^{-\gamma(r_k-i_k)}\ di_k \\
&= \gamma\left[\frac{1}{\gamma}\mathrm{e}^{-\gamma(r_k-i_k)}\right]_{r_j}^{r_k} \\
&= 1-\mathrm{e}^{-\gamma(r_k-r_j)}.
\end{aligned}$$

(iii) Finally,

$$\begin{aligned}
\int_{-\infty}^{r_j}\int_{-\infty}^{i_k} 1 f_I(r_j-i_j) f_I(r_k-i_k)\ di_j\ di_k &= \int_{-\infty}^{r_j}\int_{-\infty}^{i_k} \gamma\mathrm{e}^{-\gamma(r_j-i_j)}\gamma\mathrm{e}^{-\gamma(r_k-i_k)}\ di_j\ di_k \\
&= \int_{-\infty}^{r_j} \gamma^2\mathrm{e}^{-\gamma(r_k-i_k)}\left[\frac{1}{\gamma}\mathrm{e}^{-\gamma(r_j-i_j)}\right]_{-\infty}^{i_k} di_k \\
&= \int_{-\infty}^{r_j} \gamma\mathrm{e}^{-\gamma(r_k-i_k)}\mathrm{e}^{-\gamma(r_j-i_k)}\ di_k \\
&= \gamma\left[\frac{1}{2\gamma}\mathrm{e}^{-\gamma(r_k+r_j-2i_k)}\right]_{-\infty}^{r_j} \\
&= \frac{1}{2}\mathrm{e}^{-\gamma(r_k-r_j)}.
\end{aligned}$$

Thus, for $r_k \geq r_j$,

$$\mathbb{E}\left[e^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right] = 1 - \frac{\beta_{kj}}{2(\beta_{kj}+\gamma)} e^{-\gamma(r_k - r_j)}. \tag{3.4.9}$$

Case (ii): $r_k < r_j$

If $r_k < r_j$,

$$r_k \wedge i_j - i_k \wedge i_j = \begin{cases} r_k - i_k & \text{if } r_k < i_j, \\ i_j - i_k & \text{if } i_k < i_j < r_k, \\ 0 & \text{otherwise,} \end{cases}$$

and so

$$\begin{aligned} \mathbb{E}[e^{-\beta(r_k \wedge i_j - i_k \wedge i_j)} \mid r_k \geq r_j] \;=\; & \int_{-\infty}^{r_k}\int_{r_k}^{r_j} e^{-\beta(r_k - i_k)} f_I(r_j - i_j) f_I(r_k - i_k)\; di_j\; di_k \\ & + \int_{-\infty}^{r_k}\int_{i_k}^{r_k} e^{-\beta(i_j - i_k)} f_I(r_j - i_j) f_I(r_k - i_k)\; di_j\; di_k \\ & + \int_{-\infty}^{r_k}\int_{-\infty}^{i_k} 1 f_I(r_j - i_j) f_I(r_k - i_k)\; di_j\; di_k, \end{aligned}$$

where we have again assumed i_j and i_k are independent. We will again calculate these three integrals individually.

(i) The first integral takes the form

$$\begin{aligned} &\int_{-\infty}^{r_k}\int_{r_k}^{r_j} e^{-\beta(r_k - i_k)} f_I(r_j - i_j) f_I(r_k - i_k)\; di_j\; di_k \\ &\qquad = \int_{-\infty}^{r_k}\int_{r_k}^{r_j} e^{-\beta(r_k - i_k)} \gamma e^{-\gamma(r_j - i_j)} \gamma e^{-\gamma(r_k - i_k)}\; di_j\; di_k \\ &\qquad = \int_{-\infty}^{r_k} \gamma^2 e^{-\beta(r_k - i_k)} e^{-\gamma(r_k - i_k)} \left[\frac{1}{\gamma} e^{-\gamma(r_j - i_j)}\right]_{r_k}^{r_j} di_k \\ &\qquad = (1 - e^{-\gamma(r_j - r_k)}) \int_{-\infty}^{r_k} \gamma e^{-(\beta+\gamma)(r_k - i_k)}\; di_k \\ &\qquad = \gamma(1 - e^{-\gamma(r_j - r_k)}) \left[\frac{1}{\beta+\gamma} e^{-(\beta+\gamma)(r_k - i_k)}\right]_{-\infty}^{r_k} \\ &\qquad = \frac{\gamma}{\beta+\gamma}(1 - e^{-\gamma(r_j - r_k)}). \end{aligned}$$

(ii) Similarly,

$$\begin{aligned}
&\int_{-\infty}^{r_k}\int_{i_k}^{r_k} \mathrm{e}^{-\beta(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_j\ di_k \\
&\qquad = \int_{-\infty}^{r_k}\int_{i_k}^{r_k} \mathrm{e}^{-\beta(i_j-i_k)} \gamma \mathrm{e}^{-\gamma(r_j-i_j)} \gamma \mathrm{e}^{-\gamma(r_k-i_k)}\ di_j\ di_k \\
&\qquad = \int_{-\infty}^{r_k} \gamma^2 \gamma \mathrm{e}^{-\gamma(r_k-i_k)} \left[\frac{1}{\gamma-\beta}\mathrm{e}^{-\beta(i_j-i_k)-\gamma(r_j-i_j)}\right]_{i_k}^{r_k} di_k \\
&\qquad = \int_{-\infty}^{r_k} \frac{\gamma^2}{\gamma-\beta}\left(\mathrm{e}^{-(\beta+\gamma)(r_k-i_k)-\gamma(r_j-r_k)} - \mathrm{e}^{-\gamma(r_k+r_j-2i_k)}\right) di_k \\
&\qquad = \frac{\gamma^2}{\gamma-\beta}\left[\frac{1}{\beta+\gamma}\mathrm{e}^{-(\beta+\gamma)(r_k-i_k)-\gamma(r_j-r_k)} - \frac{1}{2\gamma}\mathrm{e}^{-\gamma(r_k+r_j-2i_k)}\right]_{-\infty}^{r_k} \\
&\qquad = \frac{\gamma}{\beta+\gamma}\frac{1}{2}\mathrm{e}^{-\gamma(r_j-r_k)}.
\end{aligned}$$

(iii) Finally,

$$\begin{aligned}
&\int_{-\infty}^{r_k}\int_{-\infty}^{i_k} 1 f_I(r_j-i_j) f_I(r_k-i_k)\ di_j\ di_k \\
&\qquad = \int_{-\infty}^{r_k}\int_{-\infty}^{i_k} \gamma \mathrm{e}^{-\gamma(r_j-i_j)} \gamma \mathrm{e}^{-\gamma(r_k-i_k)}\ di_j\ di_k \\
&\qquad = \int_{-\infty}^{r_k} \gamma^2 \mathrm{e}^{-\gamma(r_k-i_k)} \left[\frac{1}{\gamma}\mathrm{e}^{-\gamma(r_j-i_j)}\right]_{-\infty}^{i_k} di_k \\
&\qquad = \int_{-\infty}^{r_k} \gamma \mathrm{e}^{-\gamma(r_k+r_j-2i_k)}\ di_k \\
&\qquad = \gamma\left[\frac{1}{2\gamma}\mathrm{e}^{-\gamma(r_k+r_j-2i_k)}\right]_{-\infty}^{r_k} \\
&\qquad = \frac{1}{2}\mathrm{e}^{\gamma(r_j-r_k)}.
\end{aligned}$$

Thus, for $r_k < r_j$,

$$\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k\wedge i_j - i_k\wedge i_j)}\right] = \frac{\gamma}{\beta_{kj}+\gamma} + \frac{\beta_{kj}}{2(\beta_{kj}+\gamma)}\mathrm{e}^{-\gamma(r_j-r_k)}. \tag{3.4.10}$$

Combining Equations (3.4.9) and (3.4.10), we obtain overall

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1\\k\neq j}}^{n} \begin{cases} 1 - \frac{\beta_{kj}}{2(\beta_{kj}+\gamma)}\mathrm{e}^{-\gamma(r_k-r_j)} & \text{if } r_k \geq r_j, \\ \frac{\gamma}{\beta_{kj}+\gamma} + \frac{\beta_{kj}}{2(\beta_{kj}+\gamma)}\mathrm{e}^{-\gamma(r_j-r_k)} & \text{if } r_k < r_j. \end{cases}$$

With the expressions for $\mathbb{E}[\chi_j\phi_j]$ and $\mathbb{E}[\psi_j]$ obtained, for any given choice of prior probability mass function $\pi(\kappa)$ and prior probability density function $\pi(i_\kappa \mid \kappa)$ it is possible to explicitly calculate the approximate likelihood

$\pi(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta})$, avoiding numerical integration. In summary,

$$\pi_{\mathrm{I}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \Big(\prod_{j=1}^{n} \mathbb{E}[\chi_j \phi_j]\mathbb{E}[\psi_j]\Big) \sum_{\kappa=1}^{n} \frac{\pi(\kappa)\mathbb{E}[\phi_\kappa \pi(i_\kappa \mid \kappa)]}{\mathbb{E}[\chi_\kappa \phi_\kappa]\mathbb{E}[\psi_\kappa]},$$

where

$$\begin{aligned}
\mathbb{E}[\chi_j \phi_j] &= \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} \frac{\gamma}{2\gamma + B_j} \mathrm{e}^{-\gamma|r_k - r_j| - B_j((r_j - r_k) \vee 0)} \\
\mathbb{E}[\psi_j] &= \prod_{\substack{k=1 \\ k \neq j}}^{n} \begin{cases} 1 - \frac{\beta_{kj}}{2(\beta_{kj}+\gamma)} \mathrm{e}^{-\gamma(r_k - r_j)} & \text{if } r_k \geq r_j, \\ \frac{\gamma}{\beta_{kj}+\gamma} + \frac{\beta_{kj}}{2(\beta_{kj}+\gamma)} \mathrm{e}^{-\gamma(r_j - r_k)} & \text{if } r_k < r_j. \end{cases}
\end{aligned} \tag{3.4.11}$$

3.4.3 PBLA I: Likelihood Calculations for Gamma Infectious Periods

In this section we provide the likelihood expression for the Pair-Based Likelihood Approximation (version I) with gamma distributed infectious periods. Here, $f_I(r_j - i_j \mid m, \gamma) = \frac{\gamma^m}{\Gamma(m)}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)}$. We will restrict shape parameter m to integer values for all of the PBLA versions, so the distribution is in fact Erlang. As in the exponential case, we require expressions for $\mathbb{E}[\chi_j \phi_j]$ and $\mathbb{E}[\psi_j]$. Integration arguments may be made in much the same manner as the exponential case, and so we do not provide the full calculations here. These may instead be found in Appendix B.

The resulting likelihood is given by:

$$\pi_{\mathrm{I}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \Big(\prod_{j=1}^{n} \mathbb{E}[\chi_j \phi_j]\mathbb{E}[\psi_j]\Big) \sum_{\kappa=1}^{n} \frac{\pi(\kappa)\mathbb{E}[\phi_\kappa \pi(i_\kappa \mid \kappa)]}{\mathbb{E}[\chi_\kappa \phi_\kappa]\mathbb{E}[\psi_\kappa]},$$

where

$$\mathbb{E}[\chi_j\phi_j] = \sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\begin{cases}\left(\frac{\gamma}{\gamma+B_j}\right)^m(1-F_{m,\gamma}(r_k-r_j)) & \\ \quad -\sum_{l=0}^{m-1}\frac{\gamma^{2m}}{(2\gamma+B_j)^{l+1}}\frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{(\gamma+B_j)^{m-l}\Gamma(m)} & \\ \quad \times\mathbb{E}\left[(r_k-r_j+Y)^{m-1}\mid Y\sim\Gamma(l+1,2\gamma+B_j)\right] & \text{if } r_k\geq r_j,\\ \left(\frac{\gamma}{\gamma+B_j}\right)^m\left(1-F_{m,\gamma+B_j}(r_j-r_k)\right) & \\ \quad -\left(\frac{\gamma}{\gamma+B_j}\right)^m\left(\frac{\gamma}{2\gamma+B_j}\right)^m\mathrm{e}^{-(\gamma+B_j)(r_j-r_k)}\sum_{l=0}^{m-1}\frac{(\gamma+B_j)^l}{l!} & \\ \quad \times\mathbb{E}\left[(r_j-r_k+Y)^{l}\mid Y\sim\Gamma(m,2\gamma+B_j)\right] & \text{if } r_k< r_j,\end{cases}$$

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1\\k\neq j}}^{n}\begin{cases}1+\sum_{l=0}^{m-1}\frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{l!2^m}\mathbb{E}\left[(r_k-r_j+Y)^{l}\mid Y\sim\Gamma(m,2\gamma)\right] & \\ \quad\times\left(\left(\frac{\gamma}{\gamma+\beta_{kj}}\right)^m(\gamma+\beta_{kj})^l-\gamma^l\right) & \text{if } r_k\geq r_j,\\ 1-F_{m,\gamma}(r_j-r_k)\left(1-\left(\frac{\gamma}{\gamma+\beta_{kj}}\right)^m\right)+\sum_{l=0}^{m-1}\frac{\gamma^{m-1}\mathrm{e}^{-\gamma(r_j-r_k)}}{2^{l+1}\Gamma(m)} & \\ \quad\times\mathbb{E}\left[(r_j-r_k+Y)^{m-1}\mid Y\sim\Gamma(l+1,2\gamma)\right] & \\ \quad\times\left(\left(\frac{\gamma}{\gamma+\beta_{kj}}\right)^m\left(\frac{\gamma+\beta_{kj}}{\gamma}\right)^l-1\right) & \text{if } r_k< r_j,\end{cases} \tag{3.4.12}$$

where $F_{k,\theta}$ is the CDF of a gamma distribution with shape k and rate θ, i.e.

$$F_{k,\theta}(x) = 1-\sum_{l=0}^{k-1}\frac{1}{l!}(\theta x)^l\mathrm{e}^{-\theta x}.$$

The expectation terms, which define the expectation of a function of a gamma distributed variable, are given by

$$\mathbb{E}\left[(r+X)^l\mid X\sim\Gamma(m,\gamma)\right] = \sum_{p=0}^{l}\binom{l}{p}r^{l-p}\frac{(m+p-1)_p}{\gamma^p},$$

with $(x)_p = \binom{x}{p}p!$.

3.4.4 PBLA I: Probabilistic Arguments

The likelihood expressions obtained in the previous sections can also be explained via probabilistic arguments, which perhaps provide more of an intuition as to how the results arise. We will illustrate this here for the case

of exponentially distributed infectious periods, and similar arguments for the gamma distribution may be found in Appendix B. First, we provide a proposition which will be necessary for these arguments.

Proposition 3.4.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, for any sample space Ω, set of events $\mathcal{F}$ and probability function P. For a random variable X on $(\Omega, \mathcal{F}, P)$ and event $A \in \mathcal{F}$,*

$$\mathbb{E}[\mathbb{1}_{\{A\}} X] = \mathbb{E}[X \mid A] P(A).$$

Proof. By definition,

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{\{A\}} X] &= \int_{\Omega} \mathbb{1}_{\{A\}} \ X \ dP \\ &= \int_{A} X \ dP. \end{aligned}$$

We define a σ-field $\mathcal{D} = \{A, A^c, \emptyset, \Omega\}$. Then $\mathcal{D}$ is a sub-σ-field of $\mathcal{F}$ and

$$\int_D \mathbb{E}[X \mid \mathcal{D}] \ dP = \int_D X \ dP \quad \forall \ D \in \mathcal{D}.$$

Then,

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{\{A\}} X] &= \int_{A} X \ dP \\ &= \int_{A} \mathbb{E}[X \mid \mathcal{D}] \ dP. \end{aligned}$$

Now, $\mathbb{E}[X \mid \mathcal{D}]$ is $\mathcal{D}$-measurable, and so by definition is constant on the atoms of $\mathcal{D}$, meaning

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{\{A\}} X] &= \mathbb{E}[X \mid A] \int_{A} \ dP \\ &= \mathbb{E}[X \mid A] P(A), \end{aligned}$$

as required. □

Now we may proceed to the probability arguments for the PBLA I method. These arguments essentially work by moving backwards in time from individual j's removal time r_j, and considering all possible combinations of events.

3.4.4.1 Exponential Infectious Periods

Expression one: $\mathbb{E}[\chi_j \phi_j]$

We begin with the exponential case and consider the calculation of $\mathbb{E}[\chi_j \phi_j]$. We have found that if we set $B_j = \sum_{l=n+1}^{N} \beta_{jl}$, as in Equation (3.4.6), then we must calculate $\mathbb{E}\left[\mathbb{1}_{\{i_k<i_j<r_k\}} e^{-B_j(r_j-i_j)}\right]$ for all pairs of infectives j and k,

We apply Proposition 3.4.1 so that

$$\mathbb{E}\left[\mathbb{1}_{\{i_k<i_j<r_k\}} e^{-B_j(r_j-i_j)}\right] = \mathbb{E}\left[e^{-B_j(r_j-i_j)} \mid i_k < i_j < r_k\right] \mathbb{P}(i_k < i_j < r_k),$$

and begin with the case $r_k \geq r_j$, which is shown in Figure 3.1.

Case (i): $r_k \geq r_j$

Firstly, consider the term $\mathbb{E}\left[e^{-B_j(r_j-i_j)} \mid i_k < i_j < r_k\right]$. The expression $e^{-B_j(r_j-i_j)}$ is equal to the probability that there are no points in a Poisson process of rate B_j which runs backwards from time r_j to i_j. Since $i_j > i_k$, i_j is the minimum of two independent exponentially distributed periods of rate γ running backwards from r_j, and hence is exponentially distributed with rate 2γ. (Recall that we are working backwards in time, so the minimum corresponds to the closest event to r_j). The probability that this event takes place before a point in the process of rate B_j is therefore simply $\frac{2\gamma}{2\gamma+B_j}$.

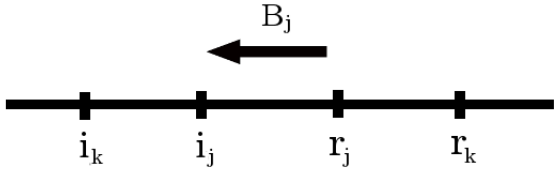


Figure 3.1: Order of events if $r_k \geq r_j$.

Secondly, the probability that $i_k < i_j < r_k$, with r_k known, may be considered in two parts. Firstly, moving backwards in time from r_k, we require that r_j is the first event to occur or equivalently that r_j occurs before i_k. This is given

by a Poisson process of rate γ running backwards in time from r_k to r_j with no points, and hence is of probability $\mathrm{e}^{-\gamma(r_j - r_k)}$. This is followed (still moving backwards in time) by the event i_j occurring before i_k. Backwards from r_j, the processes governing events i_j and i_k occurring are given by two independent Poisson processes of rate γ, and hence the probability of either occurring first is simply $\frac{1}{2}$. Therefore $\mathbb{P}(i_k < i_j < r_k) = \frac{1}{2}\mathrm{e}^{-\gamma(r_k - r_j)}$.

Combining these expressions,

$$\mathbb{E}\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \mathrm{e}^{-B_j(r_j - i_j)} \mid r_k \geq r_j\right] = \frac{\gamma}{2\gamma + B_j}\mathrm{e}^{-\gamma(r_k - r_j)}, \tag{3.4.13}$$

as was obtained in Equation (3.4.7).

Case (ii): $r_k < r_j$

The case $r_k < r_j$ is similar, and shown in Figure 3.2. The calculation of $\mathbb{E}\left[\mathrm{e}^{-B_j(r_j - i_j)} \mid i_k < i_j < r_k\right]$ may be explained as in the case $r_k \geq r_j$ except that the Poisson process runs backwards from r_k rather than r_j. The result is the same as in the case $r_k \geq r_j$: $\mathbb{E}\left[\mathrm{e}^{-B_j(r_j - i_j)} \mid i_k < i_j < r_k\right] = \frac{2\gamma}{2\gamma + B_j}$.

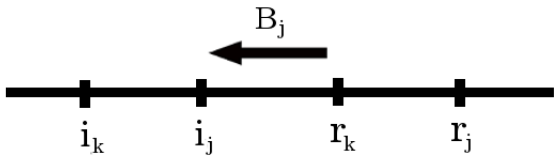


Figure 3.2: Order of events if $r_k < r_j$.

The term $\mathbb{P}(i_k < i_j < r_k)$ follows much the same reasoning as before, but now also requires no points in the Poisson process of rate B_j running backwards between r_j and r_k. This will have probability $\mathrm{e}^{-B_j(r_j - r_k)}$. Then backwards from r_k follows the same argument as before but with r_j and r_k reversed so that $\mathbb{P}(i_k < i_j < r_k) = \frac{1}{2}\mathrm{e}^{-\gamma(r_j - r_k)}\mathrm{e}^{-B_j(r_j - r_k)}$, and

$$\mathbb{E}\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \mathrm{e}^{-B_j(r_j - i_j)} \mid r_k < r_j\right] = \frac{2\gamma}{2\gamma + B_j}\frac{1}{2}\mathrm{e}^{-\gamma(r_j - r_k)}\mathrm{e}^{-B_j(r_j - r_k)}, \tag{3.4.14}$$

equal to Equation (3.4.8).

Then overall, combining Equations (3.4.13) and (3.4.14),

$$\mathbb{E}[\chi_j \phi_j] = \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} \frac{\gamma}{2\gamma + B_j} \mathrm{e}^{-\gamma |r_k - r_j| - B_j((r_j - r_k) \vee 0)}$$

as was found in the previous calculations in Equation (3.4.11).

Expression two: $\mathbb{E}[\psi_j]$

Next we consider the calculation of $\mathbb{E}[\psi_j]$. Recall that under PBLA,

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1 \\ k \neq j}}^{n} \mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right],$$

and so for a given j and k we need to find $\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right]$. For simplicity, we define $\tau_{kj} = r_k \wedge i_j - i_k \wedge i_j$ as the length of time infectious pressure is applied from k to j. For ease of exposition, we write $\beta_{kj} = \beta$ in the calculations that follow.

We begin by conditioning on the values of r_j and r_k as before.

Case (i): $r_k \geq r_j$

In this case,

$$\tau_{kj} = \begin{cases} i_j - i_k & \text{if } i_k < i_j < r_j, \\ 0 & \text{otherwise.} \end{cases}$$

Then, conditioning on the possible values of i_j and i_k,

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}} \mid r_k \geq r_j\right] = \mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}} \mathbb{1}_{\{i_k < i_j < r_j\}}\right] + \mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}} \mathbb{1}_{\{i_j < i_k < r_j\}}\right] \\ + \mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}} \mathbb{1}_{\{i_j < r_j < i_k\}}\right], \end{aligned}$$

where we may calculate each term in the sum individually, applying Proposition 3.4.1 in each case.

(i) First,

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}} \mathbb{1}_{\{i_k < i_j < r_j\}}\right] &= \mathbb{E}\left[\mathrm{e}^{-\beta(i_j - i_k)} \mid i_k < i_j < r_j\right] \mathbb{P}(i_k < i_j < r_j) \\ &= \frac{\gamma}{\gamma + \beta} \times \frac{1}{2} \mathrm{e}^{-\gamma(r_k - r_j)}, \end{aligned}$$

where this situation is as displayed in Figure 3.1, except without the B_j process. Here, $e^{-\beta(i_j - i_k)}$ is equal to the probability of there being no points in a Poisson process of rate β between i_j and i_k. Working backwards from i_j, this concerns the event that the Poisson process of rate γ leading to i_k has a point before this one of rate β. This is hence of probability $\frac{\gamma}{\gamma+\beta}$. The probability that $i_k < i_j < r_j$ is given in part by the probability that there are no points in a Poisson process of rate γ from r_k back to r_j, which is $e^{-\gamma(r_k - r_j)}$. This is multiplied by $\frac{1}{2}$, since backwards from r_j the infectious periods for j and k are both governed by Poisson processes of rate γ, and so occur with equal probability.

Figure 3.3: Order of events if $r_k \geq r_j$.

(ii) Similarly,

$$\begin{aligned}\mathbb{E}\left[e^{-\beta\tau_{kj}}\mathbb{1}_{\{i_j<i_k<r_j\}}\right] &= \mathbb{E}\left[e^{-\beta(0)} \mid i_j < i_k < r_j\right]\mathbb{P}(i_j < i_k) \\ &= \mathbb{P}(i_j < i_k) \\ &= \frac{1}{2}e^{-\gamma(r_k - r_j)}.\end{aligned}$$

Shown in Figure 3.3, the probability here follows exactly the same reasoning as in (i) and the expectation term collapses to 1.

Figure 3.4: Order of events if $r_k \geq r_j$.

(iii) Lastly,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}}\mathbb{1}_{\{i_j<r_j<i_k\}}\right] &= \mathbb{E}\left[\mathrm{e}^{-\beta(0)} \mid i_j < r_j < i_k\right]\mathbb{P}(i_j < r_j < i_k)\\ &= \mathbb{P}(i_j < r_j < i_k)\\ &= 1 - \mathrm{e}^{-\gamma(r_k - r_j)}.\end{aligned}$$

Here again the expectation simplifies to 1. The probability that $i_j < r_j < i_k$ (this ordering of events is shown in Figure 3.4) is equal to one minus the probability of there being no points in a Poisson process of rate γ between r_k and r_j, since i_k can occur either before or after r_j.

Combining the expressions from (i), (ii) and (iii), we obtain an expression equal to that found in Equation (3.4.9),

$$\mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}} \mid r_k \geq r_j\right] = 1 - \frac{\beta}{2(\beta+\gamma)}\mathrm{e}^{-\gamma(r_k - r_j)}. \tag{3.4.15}$$

Case (ii): $r_k < r_j$

If $r_k < r_j$

$$r_k \wedge i_j - i_k \wedge i_j = \begin{cases} r_k - i_k & \text{if } r_k < i_j, \\ i_j - i_k & \text{if } i_k < i_j < r_k, \\ 0 & \text{otherwise.} \end{cases}$$

Conditioning on the possible values of i_j and i_k, and setting $\beta_{kj} = \beta$,

$$\mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}} \mid r_k < r_j\right] = \mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}}\mathbb{1}_{\{i_j<i_k\}}\right] + \mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}}\mathbb{1}_{\{i_k<i_j<r_k\}}\right] + \mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}}\mathbb{1}_{\{r_k<i_j\}}\right].$$

Applying Proposition 3.4.1, we may calculate each term in the sum individually.

Figure 3.5: Order of events if $r_k \geq r_j$.

(i) This case is shown in Figure 3.5.

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}}\mathbb{1}_{\{i_j<i_k\}}\right] &= \mathbb{E}\left[\mathrm{e}^{-\beta(0)} \mid i_j < i_k\right]\mathbb{P}(i_j < i_k)\\ &= \mathbb{P}(i_j < i_k)\\ &= \frac{1}{2}\mathrm{e}^{-\gamma(r_j-r_k)},\end{aligned}$$

using similar logic to the case $r_k \geq r_j$. The expectation term reduces to 1, and the probability is equal to the probability of no points in a Poisson process of rate γ backwards from r_j to r_k, followed by the equal chance of i_j and i_k occurring first before r_k.

(ii) In this case, which is the same as shown in Figure 3.2 except without the B_j process,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}}\mathbb{1}_{\{i_k<i_j<r_k\}}\right] &= \mathbb{E}\left[\mathrm{e}^{-\beta(i_j-i_k)} \mid i_k < i_j < r_k\right]\mathbb{P}(i_k < i_j < r_k)\\ &= \frac{\gamma}{\gamma+\beta} \times \frac{1}{2}\mathrm{e}^{-\gamma(r_j-r_k)}.\end{aligned}$$

This follows the same logic we have seen in previous cases; the expectation equal to the probability that the Poisson process governing i_k has a point before the process of rate β, and the probability the same as in case (i).

Figure 3.6: Order of events if $r_k \geq r_j$.

(iii) The final expectation, with timeline shown in Figure 3.6, is equal to

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta\tau_{kj}}\mathbb{1}_{\{r_k<i_j\}}\right] &= \mathbb{E}\left[\mathrm{e}^{-\beta(r_k-i_k)} \mid r_k < i_j\right]\mathbb{P}(r_k < i_j)\\ &= \frac{\gamma}{\gamma+\beta} \times (1 - \mathrm{e}^{-\gamma(r_j-r_k)}),\end{aligned}$$

where the expectation term is as before, and the probability is equal to 1 minus the probability of there being no points in the Poisson process of rate γ between r_j and r_k.

Combining these expressions,

$$\mathbb{E}\left[e^{-\beta\tau_{kj}} \mid r_k < r_j\right] = \frac{\gamma}{\beta+\gamma} + \frac{\beta}{2(\beta+\gamma)}e^{-\gamma(r_j-r_k)}, \tag{3.4.16}$$

equal to Equation(3.4.10).

Overall, combining Equations (3.4.15) and (3.4.16),

$$\mathbb{E}\left[e^{-\beta_{kj}\tau_{kj}}\right] = \begin{cases} 1 - \frac{\beta_{kj}}{2(\beta_{kj}+\gamma)}e^{-\gamma(r_k-r_j)} & \text{if } r_k \geq r_j, \\ \frac{\gamma}{\beta_{kj}+\gamma} + \frac{\beta_{kj}}{2(\beta_{kj}+\gamma)}e^{-\gamma(r_j-r_k)} & \text{if } r_k < r_j, \end{cases}$$

as was found in the previous integral calculations, in Equation (3.4.11).

3.4.4.2 Gamma Infectious Periods

This same general method can be extended to the case of Gamma distributed infectious periods. Since the arguments are very similar we do not include them here, but they can be found in Appendix B. In the case of exponential infectious periods, our arguments were based upon the fact that the probability of an event occurring is independent of time. In the gamma case, we use the method of stages to split a $\Gamma(m, \gamma)$ time period into m exponentially distributed sections, so that similar arguments may be used. The likelihood expression obtained is equal to that found using integration arguments in Equation (3.4.12).

3.4.5 PBLA II: Improvements to the Approximation

Having derived the first approximation PBLA I in detail, it is possible to further improve this method both in terms of computational speed and accuracy of estimation. In this section, we provide an alternative format for the likeli-

hood approximation to improve accuracy. Recall Equation (3.2.2) for the likelihood, where the infectious periods have distribution $f_I(\cdot \mid \boldsymbol{\theta})$,

$$\begin{aligned}\pi(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) &= \int \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \phi_j f_I(r_j - i_j) \right) \phi_\kappa f_I(r_\kappa - i_\kappa) \pi(i_\kappa, \kappa) \, d\mathbf{i} \, di_\kappa \, d\kappa \\ &= \sum_{\kappa=1}^{n} \pi(\kappa) \int \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \phi_j \right) \phi_\kappa \pi(i_\kappa \mid \kappa) \prod_{j=1}^{n} f_I(r_j - i_j) \, d\mathbf{i} \, di_\kappa,\end{aligned}$$

since κ takes discrete values. Rather than separating out $\chi_j \phi_j$ and ψ_j as previously, we rearrange for an approximate likelihood as follows:

$$\pi_{\mathrm{II}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{\kappa=1}^{n} \pi(\kappa) \int \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \right) \pi(i_\kappa \mid \kappa) \prod_{j=1}^{n} \phi_j f_I(r_j - i_j) \, d\mathbf{i} \, di_\kappa.$$

Then, recalling the definition of ϕ_j in Equation (3.2.1), note that

$$\begin{aligned}\phi_j f_I(r_j - i_j) &= \mathrm{e}^{-(r_j - i_j)B_j} f_I(r_j - i_j \mid \boldsymbol{\theta}) \\ &= a(B_j, \boldsymbol{\theta}) g_j(r_j - i_j \mid \boldsymbol{\theta}, B_j)\end{aligned}$$

where

$$\begin{aligned}g_j(r_j - i_j \mid \boldsymbol{\theta}, B_j) &= \frac{\mathrm{e}^{-(r_j - i_j)B_j} f_I(r_j - i_j \mid \boldsymbol{\theta})}{\int \mathrm{e}^{-(r_j - i_j)B_j} f_I(r_j - i_j \mid \boldsymbol{\theta}) \, di_j} \\ &= \frac{\mathrm{e}^{-(r_j - i_j)B_j} f_I(r_j - i_j \mid \boldsymbol{\theta})}{a(B_j, \boldsymbol{\theta})}.\end{aligned}$$

Thus $a(B_j, \boldsymbol{\theta})$ is the moment generating function of the infectious period of individual j evaluated at B_j, and g_j is a probability density function in the sense that it integrates to 1. This amounts to a change of variable in the overall likelihood, and means that it is not necessary to specifically calculate ϕ_j.

The likelihood reduces to

$$\begin{aligned}\pi_{\text{II}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) &= \left(\sum_{\kappa=1}^{n} \pi(\kappa) \int \left(\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \right) \pi(i_\kappa \mid \kappa) \prod_{j=1}^{n} g_j(r_j - i_j) \, d\mathbf{i} \, di_\kappa \right) \\ &\quad \times \prod_{j=1}^{n} a(B_j, \boldsymbol{\theta}) \\ &= \left(\sum_{\kappa=1}^{n} \pi(\kappa) \pi(i_\kappa \mid \kappa) \mathbb{E}^{g}_{\mathbf{i}, i_\kappa} \left[\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \right] \right) \left(\prod_{j=1}^{n} a(B_j, \boldsymbol{\theta}) \right),\end{aligned}$$

where $\mathbb{E}^{g}[\cdot]$ refers to expectation with respect to $\mathbf{i}$, i_κ with probability density function $\prod_{j=1}^{n} g_j(r_j - i_j)$. Alongside defining the prior probability mass/density functions, we need to evaluate $\mathbb{E}^{g}\left[\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \chi_j \psi_j \right]$. This may be approximated as $\prod_{\substack{j=1 \\ j \neq \kappa}}^{n} \mathbb{E}^{g}[\chi_j] \mathbb{E}^{g}[\psi_j]$, as before.

We would expect PBLA II to outperform PBLA I since the absorption of the ϕ_j terms into the expectation means we do not need to evaluate or approximate them. As we will see in sections 3.4.5.1 and 3.4.5.2, the resulting forms of the $a(B_j, \boldsymbol{\theta})$ terms are simple and inexpensive to calculate for both exponential and gamma infectious periods, so this introduces only little additional computational burden by comparison.

We may also again rearrange the likelihood for computational speed, yielding

$$\pi_{\text{II}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \left(\prod_{j=1}^{n} \mathbb{E}^{g}[\chi_j] \mathbb{E}^{g}[\psi_j] a(B_j, \boldsymbol{\theta}) \right) \sum_{\kappa=1}^{n} \frac{\pi(\kappa) \mathbb{E}^{g}[\pi(i_\kappa \mid \kappa)]}{\mathbb{E}^{g}[\chi_\kappa] \mathbb{E}^{g}[\psi_\kappa]}. \tag{3.4.17}$$

For the cases of exponential and gamma distributed infectious periods, we now describe the functions $g_j(r_j - i_j)$ and $a(B_j, \boldsymbol{\theta})$.

3.4.5.1 Exponential Infectious Periods

For infectious periods exponentially distributed with rate γ and infection rate β_{kj} from individual k to individual j,

$$\begin{aligned} g_j(r_j - i_j) &\sim \text{Exp}(\gamma + B_j), \\ a(B_j, \boldsymbol{\theta}) &= \frac{\gamma}{\gamma + B_j}. \end{aligned}$$

A special case of this: for infectious periods exponentially distributed with rate γ and infection rate $\frac{\beta}{N}$,

$$g_j(r_j - i_j) \sim \text{Exp}\Big(\gamma + \frac{\beta}{N}(N-n)\Big),$$
$$a(B_j, \boldsymbol{\theta}) = \frac{\gamma}{\gamma + \frac{\beta}{N}(N-n)}.$$

3.4.5.2 Gamma Infectious Periods

For infectious periods that are gamma distributed with shape m, rate γ, and infection rate β_{kj} from individual k to individual j,

$$g_j(r_j - i_j) \sim \Gamma(m, \gamma + B_j),$$
$$a(B_j, \boldsymbol{\theta}) = \left(\frac{\gamma}{\gamma + B_j}\right)^m.$$

Again considering the special case of homogeneous mixing, for gamma distributed infectious periods with shape m and rate γ, and infection rate $\frac{\beta}{N}$, we have

$$g_j(r_j - i_j) \sim \Gamma(m, \gamma + \frac{\beta}{N}(N-n)),$$
$$a(B_j, \boldsymbol{\theta}) = \left(\frac{\gamma}{\gamma + \frac{\beta}{N}(N-n)}\right)^m.$$

3.4.6 PBLA III: Further Approximation

With regards to the approximation of $\mathbb{E}^g[\prod_{j\neq\kappa}\chi_j\psi_j]$ in PBLA II, an alternative and improved method to assuming independence between the two terms is to observe that

$$\begin{aligned}\mathbb{E}^g[\chi_j\psi_j] &= \sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}^g\left[\mathbb{1}_{\{i_k<i_j<r_k\}}\psi_j\right]\\ &= \sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}^g\left[\mathbb{1}_{\{i_k<i_j<r_k\}}\psi_{jk}\prod_{\substack{l=1\\l\neq k\\l\neq j}}^{n}\psi_{jl}\right],\end{aligned}$$

where $\psi_j = \prod_{\substack{l=1 \\ l\neq j}}^{n} \mathrm{e}^{-\beta_{lj}(r_l \wedge i_j - i_l \wedge i_j)} = \prod_{\substack{l=1 \\ l\neq j}}^{n} \psi_{jl}$, say.

If we define $\chi_{jk}(t) = \beta_{kj} \mathbb{1}_{\{k \text{ infective at } t\}}$ this method combines χ_{jk} with ψ_{jk}, essentially minimising the number of 'cases' which must be considered in the calculation. Recall that, in Section 3.4.2.2 for the calculation of $\mathbb{E}[\psi_j]$ with exponential infectious periods, we needed to calculate double integrals for all possible values of $r_k \wedge i_j - i_k \wedge i_j$, dependent on the orderings of r_k, r_j, i_k and i_j. In combining χ_{jk} with ψ_{jk} here, the indicator function reduces the number of these orderings which are possible.

Then, for computational efficiency,

$$\begin{aligned}
\mathbb{E}^g[\chi_j \psi_j] &= \sum_{\substack{k=1 \\ k\neq j}}^{n} \beta_{kj} \mathbb{E}^g\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}\right] \prod_{\substack{l=1 \\ l\neq k}}^{n} \mathbb{E}^g[\psi_{jl}] \\
&\approx \sum_{\substack{k=1 \\ k\neq j}}^{n} \beta_{kj} \mathbb{E}^g\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}\right] \frac{\prod_{l=1}^{n} \mathbb{E}^g[\psi_{jl}]}{\mathbb{E}^g[\psi_{jk}]} \\
&\approx \mathbb{E}^g[\psi_j] \sum_{\substack{k=1 \\ k\neq j}}^{n} \beta_{kj} \frac{\mathbb{E}^g\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}\right]}{\mathbb{E}^g[\psi_{jk}]}.
\end{aligned}$$

PBLA III offers marginally less approximation than PBLA II, since the χ_j and ψ_j terms do not need to be assumed independent. We expect the real advantage of this approach will be the increase in computational speed, due to the reduction of 'cases' to be considered as discussed. We will compare this method to the previous PBLA versions in Section 4.1.3, which will show that PBLA III offers considerable increased accuracy of estimation over PBLA I, but only marginal gains of accuracy over PBLA II.

3.4.7 PBLA III: Full Likelihood Expressions

Before defining the next the PBLA method, we summarise this section by providing the full likelihood expressions in the cases of exponential and gamma distributed infectious periods for PBLA III. We introduce a new variable to simplify the expressions: $\delta_j = \gamma + \sum_{l=n+1}^{N} \beta_{jl}$, for $j = 1, \ldots, n$. This reduces to

$\delta = \gamma + \frac{\beta}{N}(N-n)$ for homogeneous mixing.

3.4.7.1 Homogeneous mixing

The likelihood expression for the PBLA III likelihood with a homogeneously mixing population may be obtained simply from the PBLA I expressions by replacing γ with δ in the relevant likelihood components. The resulting likelihood is as follows:

$$\pi_{\mathrm{III}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \left(\prod_{j=1}^{n} \mathbb{E}^{\mathcal{G}}[\psi_j] \sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{\beta}{N} \frac{\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}]}{\mathbb{E}^{\mathcal{G}}[\psi_{jk}]} a(B_j, \boldsymbol{\theta}) \right) \times \sum_{\kappa=1}^{n} \frac{\pi(\kappa) \mathbb{E}^{\mathcal{G}}[\pi(i_\kappa \mid \kappa)]}{\mathbb{E}^{\mathcal{G}}[\psi_\kappa] \sum_{\substack{l=1 \\ l \neq \kappa}}^{n} \frac{\beta}{N} \frac{\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_l < i_\kappa < r_l\}} \psi_{\kappa l}]}{\mathbb{E}^{\mathcal{G}}[\psi_{\kappa l}]}}.$$

The expressions that form this likelihood depend upon the distribution of the infectious periods as usual. These are given by:

Exponential Infectious Periods

$$\mathbb{E}^{\mathcal{G}}[\psi_{jk}] = \begin{cases} 1 - \frac{\frac{\beta}{N}}{2\left(\frac{\beta}{N}+\delta\right)} \mathrm{e}^{-\delta(r_k - r_j)} & \text{if } r_k \geq r_j, \\ \frac{\delta}{\frac{\beta}{N}+\delta} + \frac{\frac{\beta}{N}}{2\left(\frac{\beta}{N}+\delta\right)} \mathrm{e}^{-\delta(r_j - r_k)}) & \text{if } r_k < r_j, \end{cases}$$

$$\mathbb{E}^{\mathcal{G}}[\psi_j] = \prod_{\substack{k=1 \\ k \neq j}}^{n} \mathbb{E}^{\mathcal{G}}[\psi_{jk}],$$

$$\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}] = \frac{\delta}{\delta + \frac{\beta}{N}} \frac{1}{2} \mathrm{e}^{-\delta |r_j - r_k|},$$

$$a(B_j, \boldsymbol{\theta}) = \frac{\gamma}{\delta}.$$

Gamma Infectious Periods

$$\mathbb{E}^{\mathcal{G}}[\psi_{jk}] = \begin{cases} 1 + \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\delta(r_k - r_j)}}{l! 2^m} \left(\left(\frac{\delta}{\delta + \frac{\beta}{N}} \right)^m (\delta + \frac{\beta}{N})^l - \delta^l \right) & \\ \quad \times \mathbb{E}\left[(r_k - r_j + \Upsilon)^l \mid \Upsilon \sim \Gamma(m, 2\delta) \right] & \text{if } r_k \geq r_j, \\ 1 - F_{m,\delta}(r_j - r_k) \left(1 - \left(\frac{\delta}{\delta + \frac{\beta}{N}} \right)^m \right) + \sum_{l=0}^{m-1} \frac{\delta^{m-1} \mathrm{e}^{-\delta(r_j - r_k)}}{2^{l+1} \Gamma(m)} & \\ \quad \times \left(\left(\frac{\delta}{\delta + \frac{\beta}{N}} \right)^m \left(\frac{\delta + \frac{\beta}{N}}{\delta} \right)^l - 1 \right) & \\ \quad \times \mathbb{E}\left[(r_j - r_k + \Upsilon)^{m-1} \mid \Upsilon \sim \Gamma(l+1, 2\delta) \right] & \text{if } r_k < r_j, \end{cases}$$

$$\mathbb{E}^{\mathcal{G}}[\psi_j] = \prod_{\substack{k=1 \\ k \neq j}}^{n} \mathbb{E}^{\mathcal{G}}[\psi_{jk}],$$

$$\mathbb{E}^{\mathcal{G}}\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk} \right] = \begin{cases} \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\delta(r_k - r_j)}}{l! 2^m} \left(\frac{\delta}{\delta + \frac{\beta}{N}} \right)^m (\delta + \frac{\beta}{N})^l & \\ \quad \times \mathbb{E}\left[(r_k - r_j + \Upsilon)^l \mid \Upsilon \sim \Gamma(m, 2\delta) \right] & \text{if } r_k \geq r_j, \\ \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\delta(r_j - r_k)}}{2^{l+1}} \left(\frac{\delta}{\delta + \frac{\beta}{N}} \right)^m \left(\frac{\delta + \frac{\beta}{N}}{\delta} \right)^l \frac{\delta^{m-1}}{\Gamma(m)} & \\ \quad \times \mathbb{E}\left[(r_j - r_k + \Upsilon)^{m-1} \mid \Upsilon \sim \Gamma(l+1, 2\delta) \right] & \text{if } r_k < r_j, \end{cases}$$

$$a(B_j, \boldsymbol{\theta}) = \left(\frac{\gamma}{\delta} \right)^m.$$

3.4.7.2 Heterogeneous mixing

For the case of a heterogeneously mixing population, the likelihood expression will be more complex. As defined in Section 3.2, the infection rate between individuals j and k is now given by β_{jk}. Following from this, we note that each expectation is with respect to a pair of infection times i_j, i_k, now with probability density functions g_j and g_k which depend on B_j and B_k, respectively. Each individual in the population is now effectively considered to have an infectious period with a different distribution. In the exponential case for example, for any pair i_j, i_k the density functions will be $\mathrm{Exp}(\delta_j)$ and $\mathrm{Exp}(\delta_k)$. Although the required likelihood expressions may not be directly taken from the PBLA I calculations, very similar integration arguments may be followed, incorporating these different density functions.

These likelihoods may also be extended to individuals whose infectious periods have completely distinct parameters, for example $r_j - i_j \sim \text{Exp}(\gamma_j)$ or $r_j - i_j \sim \Gamma(m_j, \gamma_j)$ for all j in the population. This is of interest since we do not require that all individuals in the population are modelled with the same infectious period distribution. We may, for example, model different groups (by age, gender, occupation etc.) with different infectious periods, according to their behaviour or characteristics.

The overall likelihood expression remains in much the same format,

$$\pi_{\text{III}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \left(\prod_{j=1}^{n} \mathbb{E}^{\mathcal{G}}[\psi_j] \sum_{\substack{k=1 \\ k \neq j}}^{n} \beta_{kj} \frac{\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}]}{\mathbb{E}^{\mathcal{G}}[\psi_{jk}]} a(B_j, \boldsymbol{\theta}) \right) \times \sum_{\kappa=1}^{n} \frac{\pi(\kappa) \mathbb{E}^{\mathcal{G}}[\pi(i_\kappa \mid \kappa)]}{\mathbb{E}^{\mathcal{G}}[\psi_\kappa] \sum_{\substack{l=1 \\ l \neq \kappa}}^{n} \beta_{\kappa j} \frac{\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_l < i_\kappa < r_l\}} \psi_{\kappa l}]}{\mathbb{E}^{\mathcal{G}}[\psi_{\kappa l}]}},$$

where the component expressions will now be given by:

Exponential Infectious Periods

$$\mathbb{E}^{\mathcal{G}}[\psi_{jk}] = \begin{cases} 1 - \frac{\delta_j \beta_{kj}}{(\delta_j + \delta_k)(\delta_k + \beta_{kj})} \mathrm{e}^{-\delta_k (r_k - r_j)} & \text{if } r_k \geq r_j, \\ \frac{\delta_k}{\delta_k + \beta_{kj}} + \frac{\delta_k \beta_{kj}}{(\delta_j + \delta_k)(\delta_k + \beta_{kj})} \mathrm{e}^{-\delta_j (r_j - r_k)}) & \text{if } r_k < r_j, \end{cases}$$

$$\mathbb{E}^{\mathcal{G}}[\psi_j] = \prod_{\substack{k=1 \\ k \neq j}}^{n} \mathbb{E}^{\mathcal{G}}[\psi_{jk}],$$

$$\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}] = \begin{cases} \frac{\delta_j \delta_k}{(\delta_j + \delta_k)(\delta_k + \beta_{kj})} \mathrm{e}^{-\delta_k (r_k - r_j)} & \text{if } r_k \geq r_j, \\ \frac{\delta_j \delta_k}{(\delta_j + \delta_k)(\delta_k + \beta_{kj})} \mathrm{e}^{-\delta_j (r_j - r_k)} & \text{if } r_k < r_j, \end{cases}$$

$$a(B_j, \boldsymbol{\theta}) = \frac{\gamma_j}{\delta_j}.$$

Gamma Infectious Periods

$$\mathbb{E}^{\mathcal{G}}[\psi_{jk}] = \begin{cases} 1 + \sum_{l=0}^{m_k-1} \frac{\mathrm{e}^{-\delta_k(r_k-r_j)}}{l!} \left(\frac{\delta_j}{\delta_j+\delta_k}\right)^{m_j} \\ \quad \times \left(\left(\frac{\delta_k}{\delta_k+\beta_{kj}}\right)^{m_k} (\delta_k+\beta_{kj})^l - \delta_k^l \right) \\ \quad \times \mathbb{E}\left[(r_k - r_j + \Upsilon)^l \mid \Upsilon \sim \Gamma(m_j, \delta_j+\delta_k)\right] & \text{if } r_k \geq r_j, \\ 1 - F_{m_j,\delta_j}(r_j - r_k)\left(1 - \left(\frac{\delta_k}{\delta_k+\beta_{kj}}\right)^{m_k}\right) \\ \quad \times + \sum_{l=0}^{m_k-1} \frac{\delta_j^{m_j}\mathrm{e}^{-\delta_j(r_j-r_k)}}{(\delta_j+\delta_k)^{l+1}\Gamma(m_j)} \\ \quad \times \left(\left(\frac{\delta+k}{\delta_k+\beta_{kj}}\right)^{m_k} \left(\delta_k+\beta_{kj}\right)^l - \delta_k^l \right) \\ \quad \times \mathbb{E}\left[(r_j - r_k + \Upsilon)^{m_j-1} \mid \Upsilon \sim \Gamma(l+1, \delta_j+\delta_k)\right] & \text{if } r_k < r_j, \end{cases}$$

$$\mathbb{E}^{\mathcal{G}}[\psi_j] = \prod_{\substack{k=1 \\ k\neq j}}^{n} \mathbb{E}^{\mathcal{G}}[\psi_{jk}],$$

$$\mathbb{E}^{\mathcal{G}}\left[\mathbb{1}_{\{i_k<i_j<r_k\}}\psi_{jk}\right] = \begin{cases} \sum_{l=0}^{m_k-1} \frac{\mathrm{e}^{-\delta_k(r_k-r_j)}}{l!} \left(\frac{\delta_j}{\delta_j+\delta_k}\right)^{m_j} \left(\frac{\delta_k}{\delta_k+\beta_{kj}}\right)^{m_k} (\delta_k+\beta_{kj})^l \\ \quad \times \mathbb{E}\left[(r_k - r_j + \Upsilon)^l \mid \Upsilon \sim \Gamma(m_j, \delta_j+\delta_k)\right] & \text{if } r_k \geq r_j, \\ \sum_{l=0}^{m_k-1} \mathrm{e}^{-\delta_j(r_j-r_k)} \left(\frac{\delta_k}{\delta_k+\beta_{kj}}\right)^{m_k} \left(\frac{\delta_k+\beta_{kj}}{\delta_j+\delta_k}\right)^l \frac{\delta_j^{m_j}}{(\delta_j+\delta_k)\Gamma(m_j)} \\ \quad \times \mathbb{E}\left[(r_j - r_k + \Upsilon)^{m_j-1} \mid \Upsilon \sim \Gamma(l+1, \delta_j+\delta_k)\right] & \text{if } r_k < r_j, \end{cases}$$

$$a(B_j, \boldsymbol{\theta}) = \left(\frac{\gamma_j}{\delta_j}\right)^{m_j}.$$

3.4.8 PBLA IV: Central Limit Theorem Approximation

In this section we seek to make an approximation to the ψ_j term in the PBLA II likelihood, using moment generating functions and a central limit theorem. The motivation is to improve the speed of the method, making estimation feasible for even larger population sizes. This approximation will only hold for exponentially distributed infectious periods with homogeneous mixing, so we assume in this section that $r_j - i_j \sim \text{Exp}(\gamma)$ for all individuals j, and set infection rate $\frac{\beta}{N}$. However, since we will apply this central limit theorem approximation within the framework of PBLA II we apply the usual change of

variables as explained in Section 3.4.5, so that all expectations are with respect to $g(r_j - i_j) \sim \text{Exp}(\delta)$. For simplicity, although all expectations are with respect to g in this section, we will not explicitly state this.

To begin the derivation, we recall that in the definition of the true likelihood in Equation (3.2.3), we required the calculation of

$$\mathbb{E}\left[\left(\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\chi_j\psi_j\phi_j\right)\phi_\kappa\pi(i_\kappa \mid \kappa)\right].$$

Rather than assuming independence over j for the ψ_j terms so that $\mathbb{E}\left[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\psi_j\right]$ $= \prod_{\substack{j=1\\ j\neq\kappa}}^{n}\mathbb{E}[\psi_j]$, and removing the dependency upon κ as we will rearrange the likelihood as usual for computational efficiency (see Equation (3.4.4)), we instead note that:

$$\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{n}\psi_j\right] &= \mathbb{E}\left[\prod_{j=1}^{n}\exp\left(-\sum_{\substack{k=1\\ k\neq j}}^{n}\frac{\beta}{N}(r_k \wedge i_j - i_k \wedge i_j)\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\sum_{\substack{j,k=1\\ k\neq j}}^{n}\frac{\beta}{N}(r_k \wedge i_j - i_k \wedge i_j)\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\sum_{\substack{j,k=1\\ k\neq j}}^{n}\frac{\beta}{N}\tau_{kj}\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\sum_{\substack{j,k=1\\ j<k}}^{n}\frac{\beta}{N}(\tau_{kj} + \tau_{jk})\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\sum_{\substack{j,k=1\\ j<k}}^{n}\frac{\beta}{N}\omega_{jk}\right)\right],
\end{aligned}$$

where we have paired τ_{kj} and τ_{jk} by symmetry, setting $j < k$ without loss of generality, and then defined

$$\tau_{kj} + \tau_{jk} = \omega_{jk} = \begin{cases} i_j - i_k & \text{if } i_k < i_j, \\ i_k - i_j & \text{if } i_j < i_k < r_j, \\ r_j - i_j & \text{if } i_k > r_j. \end{cases}$$

Figure 3.7: Timeline of a disease outbreak, showing reverse timescale t for PBLA IV.

Note that $\mathbb{E}\left[\prod_{j=1}^{n}\psi_j\right]$ is then in the form of a moment generating function for $\sum_{\substack{j,k=1\\ j<k}}^{n}\omega_{jk}$, evaluated at $-\frac{\beta}{N}$. Therefore, if we can find the distribution, or an approximation to the distribution, of $\sum_{\substack{j,k=1\\ j<k}}^{n}\omega_{jk}$ we will be able to replace the $\mathbb{E}\left[\prod_{j=1}^{n}\psi_j\right]$ term with the corresponding expression for its moment generating function.

In order to find this approximation to the distribution, we first require the following result:

Theorem 3.4.2.

$$\sum_{\substack{j,k=1\\ j<k}}^{n}\omega_{jk} \sim Y_1 + Y_2 + \cdots + Y_{n-1}$$

where $Y_j \sim Exp\left(\frac{\delta}{j}\right)$*, and* $Y_1, \ldots, Y_{n-1}$ *are independent.*

Proof. To prove this result we work backwards in time, similarly to the probabilistic arguments in Section 3.4.4. We define t as our reverse-timescale, where $t = 0$ at r_n^+ (the moment just before r_n in reverse time, see Figure 3.7). Then t increases as we move backwards in time to i_1 (the 'end' of the outbreak, as after this time no infectious pressure is applied). We consider the reverse infectious process as a continuous time Markov process $\{(S(t), I(t)) : t > 0\}$,

starting at r_n^+ where $S(t=0) = 0$ and $I(t=0) = 0$, such that

$$\begin{aligned}&(S, I) \to (S+1, I-1) \text{ at rate } \delta I(t)\\ &(S, I) \to (S, I+1) \text{ deterministically, when a known removal event occurs.}\end{aligned} \tag{3.4.18}$$

The first transition in Equation 3.4.18 then represents an infection event (since in reverse time this involves an individual moving from infective to susceptible), and the second a removal event (in reverse, corresponding to an individual moving into their infectious period). For example, the first event which will occur in reverse time is r_n, when we move from zero current infectives to one. Notably, we see that the number of susceptibles is not changed at any removal times. We generate infections according to our Markov process at rate $\delta I(t)$, since δ determines the length of the infectious period (under the PBLA II framework). We choose who gets 'infected' (i.e. who moves into their susceptible period) uniformly at random from those currently infective. The key aspect here is that we want to count the total number of infectives and susceptibles present in the population at all times, moving backwards from the end of the outbreak.

The quantity we wish to focus on is $\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk} = \int_0^\infty S(t)I(t)\ dt$, which is the total amount of infectious pressure applied, over all individuals and over all time. To construct this quantity in reverse time we define $T(t)$ as the total infectious pressure observed up to time t, so that $T(t) = \int_0^t S(u)I(u)\ du$, and T increases at a deterministic rate $S(t)I(t)$ at time t. Then $T(E) = \sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk}$, where $E = r_n^+ - i_1$ represents the end of the outbreak under reverse timescale t, after which there are no remaining infectives and hence no infectious pressure. Quantity T is therefore a piecewise, linear, non-decreasing function of t, whose gradient changes are determined by the transitions of the Markov process $\{(S(t), I(t)) : t > 0\}$. The whole process will stop when $S(t) = n$ at i_1, after which there will be no more increase in T.

Between any two events (whether infections or removals), the process T will increase at constant rate $S(t)I(t)$, since $S(t)$ and $I(t)$ remain constant between

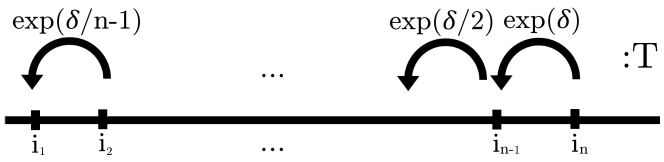


Figure 3.8: Timeline showing how total infectious pressure T is built up, under scaled reverse timescale t^*.

events even as t changes. Noting that an increase of rate x, say, for y time units, is equivalent to an increase of rate 1 for xy time units, we now perform a time scaling for t. We instead consider the construction of T by running time at rate $\frac{1}{I(t)}$ if $I(t) \geq 1$, and at rate 0 if $I(t) = 0$. Under this new timescale, time effectively stops if the number of infectives is 0. This is desired since no new infections can occur, and we simply wait for the next deterministic removal event to restart counting time. This scaling results in new timescale $t^* = \frac{t}{I(t)}$. Now T instead increases at rate $S(t)$. Since we have already shown that the number of susceptibles is unchanged by removal events, removals therefore also have no effect on the rate of increase of T. In the t^* timescale, the Markov process is defined by:

$$(S, I) \to (S + 1, I - 1) \text{ at rate } \delta \text{ if } I(t^*) \geq 1.$$

Quantity T is then formed by a series of stages wherein T is increasing at rate $S(t)$, for $S(t)$ increasing by one at rate δ. This occurs for $n - 1$ infection events at times i_j (when we get to n susceptibles at i_1, no infectious pressure is being placed, as previously discussed). T is then the sum of a series of independent random exponential lengths of time, with rates δ, $\frac{\delta}{2}$ and so on, up to $\frac{\delta}{n-1}$. Figure 3.8 shows this. Therefore, $T(E) = \sum_{\substack{j,k=1 \\ j<k}}^{n} \omega_{jk} \sim Y_1 + Y_2 + \cdots + Y_{n-1}$ where $Y_j \sim \text{Exp}\left(\frac{\delta}{j}\right)$, and $Y_1, \ldots, Y_{n-1}$ are independent, as required. □

Since these exponential distributions are independent but non-identically distributed, Theorem 3.4.2 may then be applied alongside the following central limit theorem, in order to approximate the $\mathbb{E}\left[\prod_{j=1}^{n} \psi_j\right]$ term with a moment generating function.

Theorem 3.4.3. *Define μ_n as the expectation of $\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk}$ and σ_n^2 as its variance. Then, as $n \to \infty$, $\frac{1}{\sigma_n}\left(\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk} - \mu_n \right)$ converges in distribution to a standard normal random variable, i.e.*

$$\frac{\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk} - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1),$$

where

$$\begin{aligned} \mu_n &= \frac{1}{\delta}\binom{n}{2}, \\ \sigma_n^2 &= \frac{1}{\delta^2}\binom{n}{2}\left(\frac{2n-1}{3}\right). \end{aligned}$$

Proof. To prove this we use the following result, Theorem 2.1 in Barbour and Eagleson (1985):

Lemma 3.4.4. *Define random variable X_{jk} with zero mean and finite variance, where (j,k) is a 2-subset of $\{1,2,...,n\}$ and $n \geq 1$. Let $s^2 = \sum_{\substack{i,j,k,l=1\\ i<j,k<l}}^{n} \mathbb{E}[X_{ij}X_{kl}]$. Then $\frac{1}{s}\sum_{\substack{j,k=1\\ j<k}}^{n} X_{jk}$ converges in distribution as $n \to \infty$ to the standard normal distribution if*

1. *$\mathbb{E}\left[|X_{jk}|^3\right] < \infty$ for all n and all pairs (j,k);*
2. $$\epsilon_n' = \frac{1}{s^3} \sum_{\substack{i,j=1\\ i<j}}^{n} \mathbb{E}\left[|X_{ij}|^3\right]^{\frac{1}{3}} \Bigg(\sum_{\substack{k,l=1\\ (k,l)\cap(i,j)\neq\emptyset}}^{n} \mathbb{E}\left[|X_{kl}|^3\right]^{\frac{1}{3}} \Bigg)^2 \to 0.$$

We therefore seek some $X_{jk} = f(\omega_{jk})$ which meets the requirements of this lemma. For infectives j and k, pairs (j,k) are always 2-subsets of $\{1,2,...,n\}$. Hence, if we are able to show that $f(\omega_{jk})$ has zero mean, finite variance s^2 and fulfills the two criteria, then $\frac{1}{s}\sum_{\substack{j,k=1\\ j<k}}^{n} f(\omega_{jk})$ can be approximated by a standard normal distribution, and we may approximate $\mathbb{E}\left[\prod_{j=1}^{n} \psi_j\right]$ with a normal moment generating function.

By applying Theorem 3.4.2 with $n = 2$, we know that $\omega_{jk} \sim \text{Exp}(\delta)$. Therefore, using standard results,

$$\begin{aligned} \mathbb{E}[\omega_{jk}] &= \frac{1}{\delta} \\ \text{Var}(\omega_{jk}) &= \frac{1}{\delta^2}. \end{aligned}$$

Setting $X_{jk} = \omega_{jk} - \mathbb{E}\big[\omega_{jk}\big] = \omega_{jk} - \frac{1}{\delta}$, we then obtain a random variable with zero mean and finite variance. Let us next explore the necessary requirements of lemma (3.4.4).

1. Firstly, $\mathbb{E}\big[|X_{jk}|^3\big] = \mathbb{E}\big[|\omega_{jk} - \frac{1}{\delta}|^3\big] < \infty$ since $\omega_{jk} \sim \text{Exp}(\delta)$.

2. Using the formula for s^2,

$$\begin{aligned} s^2 &= \sum_{\substack{i,j,k,l=1 \\ i<j,k<l}}^{n} \mathbb{E}[X_{ij}X_{kl}] \\ &= \sum_{\substack{i,j,k,l=1 \\ i<j,k<l}}^{n} \mathbb{E}\left[\left(\omega_{ij} - \frac{1}{\delta})(\omega_{kl} - \frac{1}{\delta}\right)\right] \\ &= \sum_{\substack{i,j,k,l=1 \\ i<j,k<l}}^{n} \mathbb{E}\left[\left(\omega_{ij}\omega_{kl} - \frac{1}{\delta}\omega_{kl} - \frac{1}{\delta}\omega_{ij} + \frac{1}{\delta^2}\right)\right] \\ &= \sum_{\substack{i,j,k,l=1 \\ i<j,k<l}}^{n} \mathbb{E}[\omega_{ij}\omega_{kl}] - \frac{2}{\delta}\sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbb{E}[\omega_{ij}] + \frac{1}{\delta^2}\binom{n}{2}^2 \\ &= \sum_{\substack{i,j,k,l=1 \\ i<j,k<l}}^{n} \mathbb{E}[\omega_{ij}\omega_{kl}] - \frac{2}{\delta}\frac{1}{\delta}\binom{n}{2} + \frac{1}{\delta^2}\binom{n}{2}^2, \end{aligned} \tag{3.4.19}$$

since $\sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbb{E}[\omega_{ij}] = \sum_{\substack{i,j=1 \\ i<j}}^{n} \frac{1}{\delta}$. We split the first term in Equation (3.4.19) into three sections based on the number of indices in common, so that

$$\sum_{\substack{i,j,k,l=1 \\ i<j,k<l}}^{n} \mathbb{E}[\omega_{ij}\omega_{kl}] = \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbb{E}[\omega_{ij}^2] + \sum_{\substack{i,j,l=1 \\ i\neq j,i\neq l}}^{n} \mathbb{E}[\omega_{ij}\omega_{il}] + \sum_{\substack{i,j,k,l=1 \\ i\neq j,k\neq l \\ i<j,k<l}}^{n} \mathbb{E}[\omega_{ij}\omega_{kl}]. \tag{3.4.20}$$

Here, $\sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbb{E}[\omega_{ij}^2]$ provides all terms in Equation (3.4.20) for which there are two indices in common, i.e. $i = k$ and $j = l$. There are $\binom{n}{2}$ possibilities for

this, where each term takes the form of a second moment of an exponentially distributed variable with parameter δ. Hence, $\sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbb{E}[\omega_{ij}^2] = \frac{2}{\delta^2}\binom{n}{2}$.

Similarly, $\sum_{\substack{i,j,k,l=1 \\ i\neq j,k\neq l \\ i<j,k<l}}^{n} \mathbb{E}[\omega_{ij}\omega_{kl}]$ includes those terms in Equation (3.4.20) which have no indices in common. There are $3\binom{n}{4}$ terms in this sum, where each represents the expectation of two independent exponentially distributed variables with parameter δ. Hence, $\sum_{\substack{i,j,k,l=1 \\ i\neq j,k\neq l \\ i<j,k<l}}^{n} \mathbb{E}[\omega_{ij}\omega_{kl}] = \frac{3}{\delta^2}\binom{n}{4}$.

Lastly, $\sum_{\substack{i,j,l=1 \\ i\neq j,i\neq l}}^{n} \mathbb{E}[\omega_{ij}\omega_{il}]$ provides the terms in Equation (3.4.20) which have 1 index in common. We find these slightly differently than the previous two cases, by grouping together all triplets of expectations with the same indices (i.e. $\mathbb{E}[\omega_{12}\omega_{13}]$, $\mathbb{E}[\omega_{12}\omega_{23}]$ and $\mathbb{E}[\omega_{13}\omega_{23}]$). There will be $\binom{n}{3}$ sets of such triplets. For each, with $i < j < k$ without loss of generality, we seek $\mathbb{E}[\omega_{ij}\omega_{ik}] + \mathbb{E}[\omega_{ij}\omega_{jk}] + \mathbb{E}[\omega_{ik}\omega_{jk}] = \mathbb{E}[\Omega_{ijk}]$, say. Then, the standard result

$$\begin{aligned}
\mathrm{Var}(\omega_{jk} + \omega_{jl} + \omega_{kl}) &= \mathrm{Var}(\omega_{jk}) + \mathrm{Var}(\omega_{jl}) + \mathrm{Var}(\omega_{kl}) + \\
&\quad 2\mathrm{Cov}(\omega_{ij}\omega_{ik} + \omega_{ij}\omega_{jk} + \omega_{ik}\omega_{jk}) \\
&= \mathrm{Var}(\omega_{jk}) + \mathrm{Var}(\omega_{jl}) + \mathrm{Var}(\omega_{kl}) + \\
&\quad 2(\mathbb{E}[\omega_{ij}\omega_{ik}] - \mathbb{E}[\omega_{ij}]\mathbb{E}[\omega_{ik}] + \mathbb{E}[\omega_{ij}\omega_{jk}] \\
&\quad -\mathbb{E}[\omega_{ij}]\mathbb{E}[\omega_{jk}] + \mathbb{E}[\omega_{ik}\omega_{jk}] - \mathbb{E}[\omega_{ik}]\mathbb{E}[\omega_{jk}]),
\end{aligned}$$

implies that

$$\begin{aligned}
\mathbb{E}[\Omega_{ijk}] &= \mathbb{E}[\omega_{ij}\omega_{ik}] + \mathbb{E}[\omega_{ij}\omega_{jk}] + \mathbb{E}[\omega_{ik}\omega_{jk}] \\
&= \frac{1}{2}\Big(\mathrm{Var}(\omega_{jk} + \omega_{jl} + \omega_{kl}) - \mathrm{Var}(\omega_{jk}) - \mathrm{Var}(\omega_{jl}) - \mathrm{Var}(\omega_{kl}) + \\
&\quad 2\mathbb{E}[\omega_{ij}]\mathbb{E}[\omega_{ik}] + 2\mathbb{E}[\omega_{ij}]\mathbb{E}[\omega_{jk}] + 2\mathbb{E}[\omega_{ik}]\mathbb{E}[\omega_{jk}]\Big) \\
&= \frac{1}{2}\left(\mathrm{Var}(\omega_{jk} + \omega_{jl} + \omega_{kl}) - \frac{3}{\delta^2} + 2\left(\frac{3}{\delta^2}\right)\right),
\end{aligned}$$

since $\omega_{ij} \sim \mathrm{Exp}(\delta)$ has mean $\frac{1}{\delta}$ and variance $\frac{1}{\delta^2}$ for all i, j, as stated previously using Theorem 3.4.2 with $n = 2$. Applying Theorem 3.4.2 with $n = 3$, we also find that

$$\omega_{jk} + \omega_{jl} + \omega_{kl} \sim \mathrm{Exp}(\delta) + \mathrm{Exp}\left(\frac{\delta}{2}\right),$$

and so

$$\text{Var}(\omega_{jk} + \omega_{jl} + \omega_{kl}) = \frac{1}{\delta^2} + \frac{4}{\delta^2} = \frac{5}{\delta^2}.$$

Therefore,

$$\begin{aligned} \mathbb{E}[\Omega_{ijk}] &= \frac{1}{2}\big(\text{Var}(\omega_{jk} + \omega_{jl} + \omega_{kl}) + \frac{3}{\delta^2}\big) \\ &= \frac{4}{\delta^2}, \end{aligned}$$

and hence we obtain an expression for the second term of Equation (3.4.20), $\sum_{\substack{i,j,l=1 \\ i\neq j, i\neq l}}^{n} \mathbb{E}[\omega_{ij}\omega_{il}] = \frac{4}{\delta^2}\binom{n}{3}$.

Combining these results, Equation (3.4.20) becomes,

$$\sum_{\substack{i,j,k,l=1 \\ i<j,k<l}}^{n} \mathbb{E}[\omega_{ij}\omega_{kl}] = \frac{2}{\delta^2}\binom{n}{2} + \frac{3}{\delta^2}\binom{n}{4} + \frac{4}{\delta^2}\binom{n}{3}. \tag{3.4.21}$$

Recombining Equation (3.4.21) with Equation (3.4.19), the expression for s^2 is then given by

$$\begin{aligned} s^2 &= \frac{2}{\delta^2}\binom{n}{2} + \frac{3}{\delta^2}\binom{n}{4} + \frac{4}{\delta^2}\binom{n}{3} - \frac{2}{\delta^2}\binom{n}{2} + \frac{1}{\delta^2}\binom{n}{2}^2 \\ &= O(n^4). \end{aligned} \tag{3.4.22}$$

Therefore,

$$\begin{aligned} \epsilon'_n &= \frac{1}{s^3}\sum_{i<j}\mathbb{E}[|X_{ij}|^3]^{\frac{1}{3}}\Big(\sum_{\substack{k,l \\ (k,l)\cap(i,j)\neq\emptyset}} \mathbb{E}[|X_{kl}|^3]^{\frac{1}{3}}\Big)^2 \\ &= \frac{1}{s^3}\sum_{i<j} A \times \big(A(2n-3)\big)^2, \end{aligned}$$

where $A = \mathbb{E}\big[|X_{ij}|^3\big]^{\frac{1}{3}}$, and since $2n-3$ is the number of pairs (k,l) where at least one of k and l is the same as i or j. Then, the terms inside the i,j sum are independent of i and j and

$$\epsilon'_n = \frac{1}{s^3}\binom{n}{2} A^3 (2n-3)^2,$$

since there are $\binom{n}{2}$ choices for this pair (i,j). Now, $s^3 = O(n^6)$ from Equation (3.4.22), and $\binom{n}{2}(2n-3)^2 = O(n^4)$. A is independent of n, so overall,

$$\begin{aligned} \epsilon'_n &= O(n^{-2}) \\ &\xrightarrow[n\to\infty]{} 0. \end{aligned}$$

We have therefore met all of the requirements of lemma 3.4.4, which implies that $\frac{1}{s}\sum_{\substack{j,k=1\\ j<k}}^{n} X_{jk} \xrightarrow{d} \mathrm{N}(1,0)$ as $n \to \infty$, from which we may approximate the distribution of $\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk}$ as $n \to \infty$.

In order to implement this result, we must specifically calculate the mean and variance of $\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk}$. These may be easily obtained from Theorem 3.4.2, since we know that $\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk} \sim \mathrm{Exp}(\delta) + \mathrm{Exp}\left(\frac{\delta}{2}\right) + \cdots + \mathrm{Exp}\left(\frac{\delta}{n-1}\right)$, and so:

$$\begin{aligned} \mathbb{E}\left[\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk}\right] &= \frac{1}{\delta} + \frac{2}{\delta} + \cdots + \frac{n-1}{\delta} \\ &= \frac{1}{\delta}\binom{n}{2}, \\ \mathrm{Var}\left(\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk}\right) &= \frac{1}{\delta^2} + \frac{4}{\delta^2} + \cdots + \frac{(n-1)^2}{\delta^2} \\ &= \frac{1}{\delta^2}\binom{n}{2}\left(\frac{2n-1}{3}\right). \end{aligned}$$

In summary, as $n \to \infty$

$$\frac{\sum_{\substack{j,k=1\\ j<k}}^{n} \omega_{jk} - \mu_n}{\sigma_n} \xrightarrow{d} \mathrm{N}(0,1),$$

where

$$\begin{aligned} \mu_n &= \frac{1}{\delta}\binom{n}{2}, \\ \sigma_n^2 &= \frac{1}{\delta^2}\binom{n}{2}\left(\frac{2n-1}{3}\right), \end{aligned}$$

as required. □

Therefore, to implement the PBLA IV method we must replace the expression for $\mathbb{E}\left[\prod_{j=1}^{n}\psi_j\right]$ with the moment generating function of a normal distribution, so that

$$\begin{aligned}\mathbb{E}\left[\prod_{j=1}^{n}\psi_j\right] &= \mathbb{E}\left[\exp\left(-\frac{\beta}{N}\sum_{\substack{j,k=1\\ j<k}}^{n}\omega_{jk}\right)\right]\\ &\approx \exp\left(-\frac{\beta}{N}\mu+\frac{1}{2}\sigma^2\left(\frac{\beta}{N}\right)^2\right)\\ &= \exp\left(-\frac{\beta}{\delta N}\binom{n}{2}+\frac{1}{2\delta^2}\binom{n}{2}\left(\frac{2n-1}{3}\right)\left(\frac{\beta}{N}\right)^2\right).\end{aligned}$$

Since this method does not involve the calculation of many sums over infectives, which are computationally intensive for large outbreaks, it provides a useful approximation for use under large outbreak sizes. However, it is certainly more specific in its requirements, only having been derived for exponential infectious periods and requiring large n. There is potential for the method to be extended to gamma distributed infectious periods, but the calculations are far more complex and we do not pursue them here. We will compare this central limit theorem approximation to the earlier PBLA versions in Chapter 4, but first we define our fifth and final PBLA method.

3.4.9 PBLA V

We now consider an alternative expression for $\mathbb{E}[\prod_{j\neq k}\chi_j\psi_j]$ within the PBLA II framework. Under the previous PBLA methods, we would consider the infectious pressure from any individual j to any individual k independently to the pressure from k to j, despite one of these two quantities being necessarily zero (as one individual must have been infected before the other). Here we derive an expression which considers these two pairs together, hopefully resulting in improved approximation. Again, all expectations in this section are with respect to density function $g_j(r_j - i_j)$ since we are in the PBLA II framework. This method will be applicable for both homogeneous and heterogeneous mixing, though we will require that $\beta_{kj} = \beta_{jk}$ for all k and j.

To begin, as before we split the expectation into two parts so that

$$\mathbb{E}\left[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\chi_j\psi_j\right] \approx \mathbb{E}\left[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\chi_j\right]\mathbb{E}\left[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\psi_j\right].$$

We then calculate $\mathbb{E}\left[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\chi_j\right]$ as previously. For the ψ term however:

$$\begin{aligned}\mathbb{E}\left[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\psi_j\right] &= \mathbb{E}\left[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\exp\Big(-\sum_{k\neq j}\beta_{kj}\tau_{kj}\Big)\right]\\ &= \mathbb{E}\left[\exp\left(-\sum_{\substack{j=1\\ j\neq\kappa}}^{n}\sum_{\substack{k=1\\ k\neq j}}^{n}\beta_{kj}\tau_{kj}\right)\right]. \qquad (3.4.23)\end{aligned}$$

We may split the inner sum into cases, depending on if $k=\kappa$, so that

$$\begin{aligned}\sum_{\substack{j=1\\ j\neq\kappa}}^{n}\sum_{\substack{k=1\\ k\neq j}}^{n}\beta_{kj}\tau_{kj} &= \sum_{\substack{j=1\\ j\neq\kappa}}^{n}\left(\sum_{\substack{k=1\\ k\neq j,\kappa}}^{n}\beta_{kj}\tau_{kj}+\beta_{\kappa j}\tau_{\kappa j}\right)\\ &= \sum_{\substack{k,j=1\\ k\neq j,k,j\neq\kappa}}^{n}\beta_{kj}\tau_{kj}+\sum_{\substack{j=1\\ j\neq\kappa}}^{n}\beta_{\kappa j}\tau_{\kappa j}\\ &= \sum_{\substack{j,k=1\\ k<j,k,j\neq\kappa}}^{n}\beta_{kj}(\tau_{kj}+\tau_{jk})+\sum_{\substack{j=1\\ j\neq\kappa}}^{n}\beta_{\kappa j}\tau_{\kappa j},\end{aligned}$$

where we have paired up the pressure from j to k and from k to j, without loss of generality setting $k<j$. Note that it is this last step which requires $\beta_{kj}=\beta_{jk}$.

Splitting up the exponential in Equation (3.4.23), our expression becomes

$$\begin{aligned}\mathbb{E}\left[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\psi_j\right] &\approx \mathbb{E}\left[\exp\left(-\sum_{\substack{j,k=1\\ k<j,k,j\neq\kappa}}^{n}\beta_{kj}(\tau_{kj}+\tau_{jk})\right)\right]\\ &\quad\times\mathbb{E}\left[\exp\left(-\sum_{\substack{j=1\\ j\neq\kappa}}^{n}\beta_{\kappa j}\tau_{\kappa j}\right)\right]\\ &\approx \prod_{\substack{j,k=1\\ k<j,k,j\neq\kappa}}^{n}\mathbb{E}\Big[\exp\big(-\beta_{kj}(\tau_{kj}+\tau_{jk})\big)\Big]\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\mathbb{E}\Big[\exp\big(-\beta_{\kappa j}\tau_{\kappa j}\big)\Big],\end{aligned} \qquad (3.4.24)$$

which may be specifically calculated, depending on the distribution of the infectious periods as usual.

3.4.9.1 Homogeneous mixing

Exponential Infectious Periods

If we assume that the infectious periods are exponentially distributed and the population mixes homogeneously, the expectations in Equation (3.4.24) are simple to calculate. Using Theorem 3.4.2 and recalling that $\tau_{kj} + \tau_{jk} = \omega_{jk}$ for exponential infectious periods, it is known that $\tau_{kj} + \tau_{jk} \sim \text{Exp}(\delta)$ (since we are working within the framework of PBLA II including the change of variables). It is also simple to show that $\tau_{\kappa j} \sim \text{Exp}(\delta)$, similarly. Therefore, both of the expectations in the expressions for $\mathbb{E}\big[\prod_{j\neq\kappa}\psi_j\big]$ take the form of moment generating functions for an exponentially distributed variable with mean δ.

$$\begin{aligned}\mathbb{E}\Big[\exp\big(-\frac{\beta}{N}(\tau_{kj}+\tau_{jk})\big)\Big] &= \frac{\delta}{\delta+\frac{\beta}{N}}\\ \mathbb{E}\Big[\exp\big(-\frac{\beta}{N}\tau_{\kappa j}\big)\Big] &= \frac{\delta}{\delta+\frac{\beta}{N}}.\end{aligned}$$

Then, under PBLA,

$$\begin{aligned}\mathbb{E}\big[\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\psi_j\big] &= \prod_{\substack{j,k=1\\ k<j,k,j\neq\kappa}}^{n}\frac{\delta}{\delta+\frac{\beta}{N}}\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\frac{\delta}{\delta+\frac{\beta}{N}}\\ &= \left(\frac{\delta}{\delta+\frac{\beta}{N}}\right)^{\frac{(n-1)(n-2)}{2}+n-1}\\ &= \left(\frac{\delta}{\delta+\frac{\beta}{N}}\right)^{\frac{n(n-1)}{2}},\end{aligned}$$

since there are $\binom{n-1}{2}$ terms in the product over k and j, and $n-1$ terms in the product over $j \neq \kappa$.

Gamma Infectious Periods

In the case of gamma distributed infectious periods and a homogeneously mixing population, the previous arguments using moment generating functions cannot be applied since we have no equivalent expression for the distribution of ω_{jk}. However, we can still calculate the expectations in Equation

(3.4.24) using the same method as in PBLA I. Recall that, since $r_k < r_j$,

$$\tau_{kj} + \tau_{jk} = \begin{cases} i_k - i_j & \text{if } i_j < i_k, \\ i_j - i_k & \text{if } i_k < i_j < r_k, \\ r_k - i_k & \text{if } r_k < i_j, \end{cases}$$

then

$$\mathbb{E}[e^{-\frac{\beta}{N}(\tau_{kj}+\tau_{jk})}] = \mathbb{E}[e^{-\frac{\beta}{N}(i_k-i_j)}\mathbb{1}_{i_j<i_k}] + \mathbb{E}[e^{-\frac{\beta}{N}(i_j-i_k)}\mathbb{1}_{i_k<i_j<r_k}] + \mathbb{E}[e^{-\frac{\beta}{N}(r_k-i_k)}\mathbb{1}_{r_k<i_j}].$$

Each of these expectations may be explicitly calculated, either through direct integration or probability type arguments as before. Recombining these, we are able to obtain the expression

$$\begin{aligned} \mathbb{E}[e^{-\frac{\beta}{N}(\tau_{kj}+\tau_{jk})}] \;=\; & \left(\frac{\delta}{\delta+\frac{\beta}{N}}\right)^m + \left(\frac{\delta}{\delta+\frac{\beta}{N}}\right)^m \sum_{l=0}^{m-1} e^{-\delta(r_j-r_k)} \\ & \times \left(\frac{1}{2^m l!}\mathbb{E}[(r_j - r_k + Y)^l \mid Y \sim \Gamma(m, 2\delta)]\left(\delta+\frac{\beta}{N}\right)^l \right. \\ & \quad + \frac{\delta^{m-1}}{2^{l+1}\Gamma(m)}\mathbb{E}[(r_j - r_k + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\delta)] \\ & \quad \left. \times \left(\frac{\delta}{\delta+\frac{\beta}{N}}\right)^{-l} - \frac{\delta^l}{l!}(r_j - r_k)^l \right). \end{aligned}$$

The expression for the κ terms may be calculated similarly, and both substituted into Equation (3.4.24). Although this form is not as simple as in the exponential case, it is still independent of the infection times and may be computed with relative speed.

3.4.9.2 Heterogeneous mixing

As in the PBLA III calculations, these expressions may be extended to heterogeneously mixing populations. We will not include these calculations as they follow much the same format as in previous sections, but the resulting expressions are as follows:

Exponential Infectious Periods

For exponentially distributed infectious periods such that $r_j - i_j \sim \text{Exp}(\delta_j)$ for $j = 1, \ldots, n$:

$$\mathbb{E}\Big[\exp\big(-\beta_{kj}(\tau_{kj}+\tau_{jk})\big)\Big] = \frac{\delta_k}{\delta_k+\beta_{kj}} + \mathrm{e}^{-\delta_j(r_j-r_k)} \times\Big(\frac{\delta_k\beta_{kj}(\delta_j-\delta_k)}{(\delta_k+\beta_{kj})(\delta_j+\delta_k)(\delta_j+\beta_{kj})}\Big).$$

Gamma Infectious Periods

For gamma distributed infectious periods such that $r_j - i_j \sim \Gamma(m_j, \delta_j)$ for $j = 1, \ldots, n$:

$$\begin{aligned}\mathbb{E}\Big[\exp\big(-\beta_{kj}(\tau_{kj}+\tau_{jk})\big)\Big] &= \Big(\frac{\delta_k}{\delta_k+\beta_{kj}}\Big)^{m_k} + \sum_{l=0}^{m_k-1}\mathrm{e}^{-\delta_j(r_j-r_k)}\Big(\frac{\delta_k}{\delta_k+\beta_{kj}}\Big)^{m_k} \\ &\times\mathbb{E}[(r_j-r_k+Y)^{m_j-1} \mid Y\sim\Gamma(l+1,\delta_j+\delta_k)]\Big(\frac{\delta_k+\beta_{kj}}{\delta_j+\delta_k}\Big)^{l}\frac{\delta_j^{m_j}}{\Gamma(m_j)(\delta_j+\delta_k)} \\ &+ \sum_{l=0}^{m_j-1}\mathrm{e}^{-\delta_j(r_j-r_k)}\Big(\frac{\delta_j}{\delta_j+\beta_{kj}}\Big)^{m_j}\mathbb{E}[(r_j-r_k+Y)^{l} \mid Y\sim\Gamma(m_k,\delta_j+\delta_k)] \\ &\times\Big(\frac{\delta_k}{\delta_j+\delta_k}\Big)^{m_k}(\delta_j+\beta_{kj})^{l} - \sum_{l=0}^{m_j-1}\mathrm{e}^{-\delta_j(r_j-r_k)}\Big(\frac{\delta_k}{\delta_k+\beta_{kj}}\Big)^{m_k}\frac{\delta_j^{l}}{l!}(r_j-r_k)^{l}.\end{aligned}$$

Again, equivalent expressions may be obtained for the terms involving the initial infective. However, it is worthwhile to note that, in reality, when we calculate these quantities we make a small additional approximation in not considering the contribution of the initial infective to be different than that of any other infective.

3.4.9.3 Comparison with PBLA III

Since PBLA V calculates the pressure from individual j to individual k combined with the pressure from individual k to individual j, rather than assuming these are independent quantities, we would expect it to outperform the

previous PBLA versions. In practice, however, numerical investigation in Section 4.1.3 will reveal that the PBLA V method does not offer much improvement over PBLA III. Considering the homogenous mixing case, the only difference is in the $\mathbb{E}[\psi_{jk}]$ term. We recall under PBLA III that this is given by $\mathbb{E}[e^{-\beta\tau_{jk}}]\mathbb{E}[e^{-\beta\tau_{kj}}]$ and under PBLA V is given by $\mathbb{E}[e^{-\beta(\tau_{jk}+\tau_{kj})}]$. In the exponential infectious periods case, for a given pair of individuals j, k, we set $r_j < r_k$ without loss of generality. Then,

$$\mathbb{E}[e^{-\frac{\beta}{N}\tau_{jk}}]\mathbb{E}[e^{-\frac{\beta}{N}\tau_{kj}}] = \left(1 - \frac{\frac{\beta}{N}}{2(\frac{\beta}{N}+\delta)}e^{-\delta(r_k-r_j)}\right)\left(\frac{\delta}{\frac{\beta}{N}+\delta} + \frac{\frac{\beta}{N}}{2(\frac{\beta}{N}+\delta)}e^{-\delta(r_k-r_j)}\right)$$

and

$$\mathbb{E}[e^{-\frac{\beta}{N}(\tau_{jk}+\tau_{kj})}] = \frac{\delta}{\frac{\beta}{N}+\delta},$$

so the difference between these two quantities is given by

$$\frac{\frac{\beta}{N}}{2(\frac{\beta}{N}+\delta)}e^{-\delta(r_k-r_j)} - \frac{\frac{\beta}{N}\delta}{2(\frac{\beta}{N}+\delta)^2}e^{-\delta(r_k-r_j)} - \frac{\frac{\beta}{N}^2}{4(\frac{\beta}{N}+\delta)^2}e^{-2\delta(r_k-r_j)}.$$

This is a monotone decreasing function of $r_k - r_j$, and hence the maximum occurs when $r_k = r_j$. The maximum difference in the expectation for any given pair j, k is therefore $\frac{1}{4}\left(\frac{\frac{\beta}{N}}{\frac{\beta}{N}+\delta}\right)^2$. Under the kind of parameter values that we will explore in this thesis and commonly see in practice, this equates to a maximum difference of around 0.7% between the expectations under each method for pair j, k, a difference so small that it is very unlikely to have any significant impact on estimation.

3.4.10 Equal Removal Times

Recall that in our initial definition of the model in Section 3.2, we required that all removal times are ordered such that $r_1 < r_2 < \cdots < r_n$; each much be strictly greater than the previous. The exact likelihood will be 0 if any removal times are equal, since in continuous time no two removal times will be the same with probability 1 . In terms of PBLA, it can be found that the method fails to work well when there are equal removal times within the data, i.e.

there exist $j, k \in (1, \ldots, n) : r_j = r_k$. In this section we demonstrate how this issue arises.

Take a simple example with three infectives. We define $\mathbf{r} = (1, 2, x)$ and assume a contact rate β between all pairs of individuals, with exponentially distributed infectious periods with mean γ. We will explore the effect on the PBLA III likelihood as x varies from below, to equal to, to greater than $r_2 = 2$. First, Figure 3.9 displays the (β, γ) likelihood surface over a range of values of x, for a population size of 10. We see that as x approaches $r_2 = 2$ from both above and below, the maximum point on the contour increases in both the β and γ directions. If we plot the $x = 2$ surface for increasingly large (β, γ) values, we see that the maximum tends off towards (∞, ∞).

To demonstrate the specifics of the likelihood calculation, we continue to focus on PBLA III with exponentially distributed infectious periods (the arguments being similar for other PBLA versions and for gamma infectious periods), as well as homogeneous mixing. Recall that the likelihood under this method is given by

$$\pi_{\text{III}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \Big(\prod_{j=1}^{n} \mathbb{E}^{\mathcal{G}}[\psi_j] \sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{\beta}{N} \frac{\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}]}{\mathbb{E}^{\mathcal{G}}[\psi_{jk}]} a(B_j, \boldsymbol{\theta}) \Big) \times \sum_{\kappa=1}^{n} \frac{\pi(\kappa) \mathbb{E}^{\mathcal{G}}[\pi(i_\kappa \mid \kappa)]}{\mathbb{E}^{\mathcal{G}}[\psi_\kappa] \sum_{\substack{l=1 \\ l \neq \kappa}}^{n} \frac{\beta}{N} \frac{\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_l < i_\kappa < r_l\}} \psi_{\kappa l}]}{\mathbb{E}^{\mathcal{G}}[\psi_{\kappa l}]}}.$$

Ignoring the initial infective here for simplicity, we must consider expressions $\mathbb{E}^{\mathcal{G}}[\psi_{jk}]$ and $\mathbb{E}^{\mathcal{G}}[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}]$, when $r_k = r_j$.

For a given pair of infectives j, k, recall that

$$\mathbb{E}[\psi_{jk}] = \begin{cases} 1 - \frac{\frac{\beta}{N}}{2(\frac{\beta}{N} + \delta)} \mathrm{e}^{-\delta(r_k - r_j)} & \text{if } r_k \geq r_j, \\ \frac{\delta}{\frac{\beta}{N} + \delta} + \frac{\frac{\beta}{N}}{2(\frac{\beta}{N} + \delta)} \mathrm{e}^{-\delta(r_j - r_k)}) & \text{if } r_k < r_j, \end{cases}$$

where $\delta = \gamma + \sum_{l=n+1}^{N} \beta_{jl}$ was obtained from the change of variables in PBLA II.

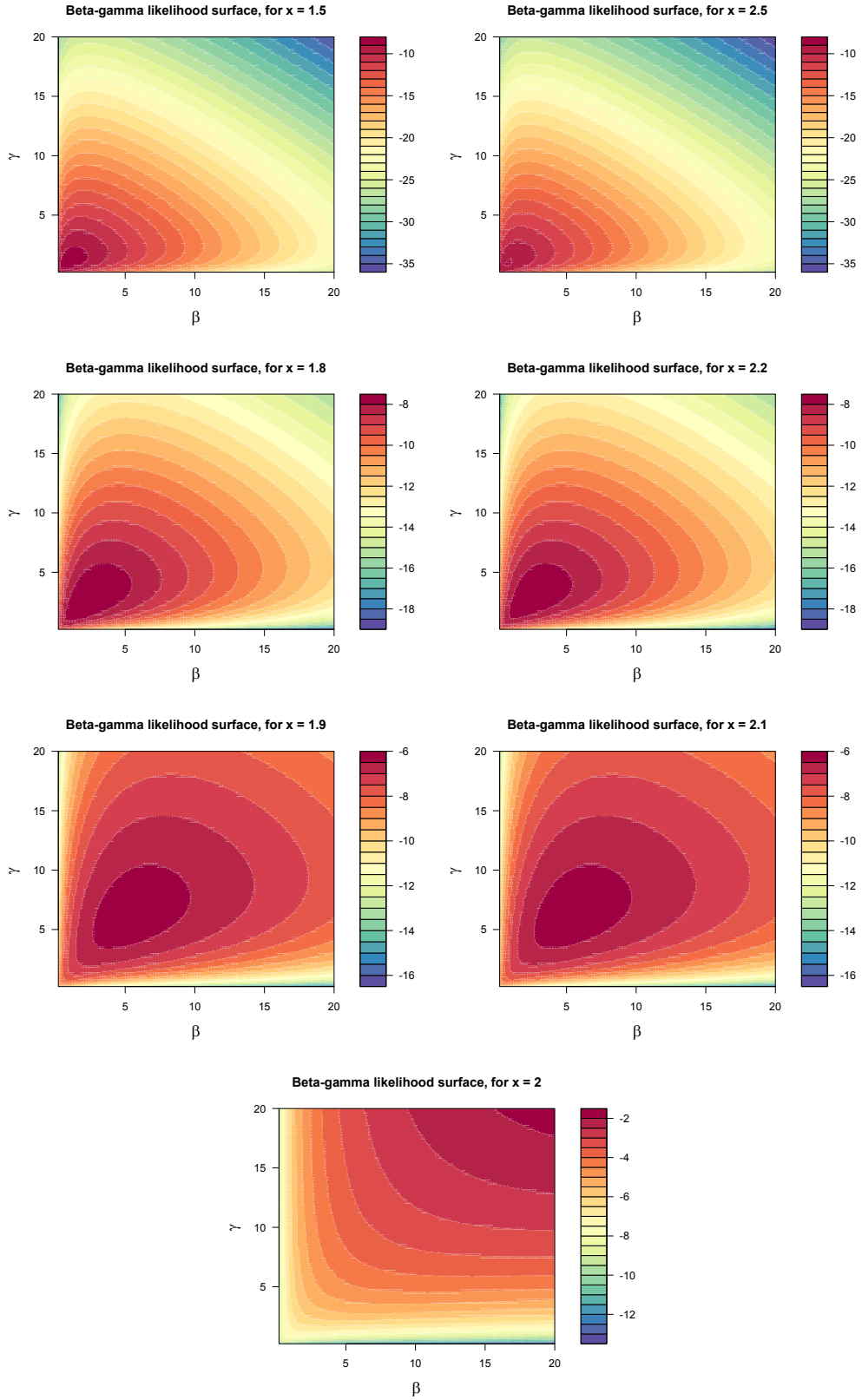


Figure 3.9: Likelihood surfaces for β and γ under the PBLA III approximation, where $\mathbf{r} = (1, 2, x)$ and x varies. This demonstrates the impact on estimation of equal removal times. $N = 10$ in all cases.

If we set equal removal times $r_k = r_j$, our expression reduces to

$$\mathbb{E}[\psi_{jk}] = 1 - \frac{\frac{\beta}{N}}{2(\frac{\beta}{N} + \delta)} = \frac{2\delta + \frac{\beta}{N}}{2(\delta + \frac{\beta}{N})}.$$

This is clearly bounded by 0 and 1, as we would expect for a probability. As γ and β tend to infinity the expression tends to 0, as we would hope for in this likelihood. Therefore, the $\mathbb{E}[\psi_j]$ section of the likelihood behaves appropriately as $r_k - r_j \to 0$.

Consider next the expression

$$\mathbb{E}\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}\right] = \frac{\delta_j}{2(\delta + \frac{\beta}{N})} \exp^{-\delta |r_j - r_k|}.$$

Hence, when $r_j = r_k$,

$$\mathbb{E}\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}\right] = \frac{\delta}{2(\delta + \frac{\beta}{N})}.$$

The combined relevant likelihood term for pair j, k such that $r_k = r_j$, recalling that $a(B_j, \boldsymbol{\theta}) = \frac{\gamma}{\delta}$, is then given by

$$\frac{\beta}{N} \frac{\mathbb{E}\left[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}\right]}{\mathbb{E}[\psi_{jk}]} a(B_j, \boldsymbol{\theta}) = \frac{\beta}{N} \frac{\frac{\delta}{2(\delta + \frac{\beta}{N})}}{\frac{2\delta + \frac{\beta}{N}}{2(\delta + \frac{\beta}{N})}} \frac{\gamma}{\delta} = \frac{\frac{\beta}{N}\gamma}{2\delta + \frac{\beta}{N}}.$$

Recall that $\delta = \gamma + \frac{\beta(N-n)}{N}$. Then as $\beta \to \infty$ and $\gamma \to \infty$, so will $\delta \to \infty$. Hence our entire likelihood expression will tend to infinity as both β and γ do. This is in agreement with what have seen in the three removal example in Figure 3.9, and demonstrates why the PBLA method may not be used with equal removal times. Fortunately, in practice it suffices to slightly jitter equal removal times in order to perform the approximation. For any pair j, k such that $r_j = r_k$, we simply set $r_k = r_k + \epsilon$ for some small ϵ of much lower order of magnitude than the removal times. This avoids the problem of equal removal times without changing the data any impactful amount, and may be extended to any larger number of equal removals, for example for daily or weekly collected data.

3.4.11 Extension to the SEIR model

So far in this chapter, our analysis has focused on application of approximation methods to only the SIR model. However, it is possible to extend the PBLA methods for use with an SEIR model. We conclude this chapter with a brief exploration of this, and the next chapter will also include an application in Section 4.3.

We define the SEIR model similarly to the SIR case: at any given time, every individual in the closed population of size N will be in one of four states: susceptible, exposed, infectious or removed. During an individual's exposed period, they are infected but not yet infectious. Then, for $j = 1, 2, \ldots, n$, e_j denotes the time of exposure of individual j, i_j denotes their infection time, and r_j their removal time. The exposure times $\mathbf{e} = \{e_j : j = 1, 2, \ldots, \kappa - 1, \kappa + 1, \ldots, n\}$ (where κ is the initial infective), and infection times $\mathbf{i} = \{i_j : j = 1, 2, \ldots, n\}$ are unknown, and the data still consist of ordered removal times $\mathbf{r} = \{r_j : j = 1, 2, \ldots, n, \text{ where } r_1 < r_2 < \cdots < r_n\}$.

The outbreak begins with the infection of the initial infective κ, at time i_κ, and continues until no infectious individuals remain. We do not allow for reinfection. During any individual i's infectious stage, they will have contact with any other individual j at times given by the points of a Poisson process of rate β_{ij}, where all such Poisson processes are assumed mutually independent. Any contact with a susceptible individual results in their immediate infection. Then $\boldsymbol{\beta} = \{\beta_{ij} : i, j \in \{1, 2, \ldots N\}\}$ provides a matrix of these contact rates, which may again be defined to incorporate a wide range of population structures. For a given outbreak of any disease, the infectious periods will have probability density (or mass) function $f_I(\cdot \mid \boldsymbol{\theta_I})$, where f_I has parameter vector $\boldsymbol{\theta_I}$. We must also now define the lengths of the exposed periods, which we state as having probability density (or mass) function $f_E(\cdot \mid \boldsymbol{\theta_E})$. First, however, we will restrict our attention to fixed length exposed periods.

3.4.11.1 Fixed length exposed periods

For both exponential and gamma infectious periods, it is possible to revisit all of the PBLA calculations using an SEIR model with fixed length exposed periods. The integrals may all be similarly calculated to achieve expressions for $\mathbb{E}[\chi_j]$, $\mathbb{E}[\psi_j]$ etc. However, we may bypass these by noting some key facts about the way the exposed period will affect our arguments.

Take, for example, $\mathbb{E}[\psi_j]$. Recall that this represents the expected probability that individual j avoids infection until the time they become infected, now e_j. Hence,

$$\mathbb{E}[\psi_j] \approx \prod_{\substack{k=1 \\ k\neq j}}^{n} \mathbb{E}[\exp\big(-\beta_{kj}(r_k \wedge e_j - i_k \wedge e_j)\big)].$$

However, with fixed length infectious periods, say of length c, we may state that $e_j = i_j - c$, for $j \in 1, \ldots, n$. Hence,

$$\mathbb{E}[\psi_j] \approx \prod_{\substack{k=1 \\ k\neq j}}^{n} \mathbb{E}[\exp\big(-\beta_{kj}(r_k \wedge (i_j - c) - i_k \wedge (i_j - c))\big)].$$

All of the integration arguments will therefore be the same as in the SIR case, but with i_j shifted by c time units. This will result in the same final expressions as before, but with r_j replaced with $r_j - c$. Interpretively, the probability arguments relating the time some infective k puts infectious pressure on j, for instance, will still apply, but will be shifted c time units earlier to represent the infection of j at e_j rather than i_j. We use this to write down the likelihood expressions for both exponential and gamma distributed infectious periods directly.

Likelihood expressions for fixed length exposed periods

As with the SIR model, for PBLA III the likelihood will be of the form

$$\begin{aligned}\pi_{\mathrm{III}}(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = &\Big(\prod_{j=1}^{n} \mathbb{E}^{g}[\psi_j] \sum_{\substack{k=1 \\ k\neq j}}^{n} \beta_{kj} \frac{\mathbb{E}^{g}[\mathbb{1}_{\{i_k < i_j < r_k\}}\psi_{jk}]}{\mathbb{E}^{g}[\psi_{jk}]} a(B_j, \boldsymbol{\theta})\Big) \\ &\times \sum_{\kappa=1}^{n} \frac{\pi(\kappa)\mathbb{E}^{g}[\pi(i_\kappa \mid \kappa)]}{\mathbb{E}^{g}[\psi_\kappa] \sum_{\substack{l=1 \\ l\neq\kappa}}^{n} \beta_{l\kappa} \frac{\mathbb{E}^{g}[\mathbb{1}_{\{i_l < i_\kappa < r_l\}}\psi_{\kappa l}]}{\mathbb{E}^{g}[\psi_{\kappa l}]}}.\end{aligned}$$

For exponential infectious periods, the required expressions are

$$\mathbb{E}^{g}[\psi_{jk}] = \begin{cases} 1 - \frac{\beta_{kj}}{2(\beta_{kj}+\delta_j)} \mathrm{e}^{-\delta_j(r_k - r_j + c)} & \text{if } r_k \geq r_j - c, \\ \frac{\delta_j}{\beta_{kj}+\delta_j} + \frac{\beta_{kj}}{2(\beta_{kj}+\delta_j)} \mathrm{e}^{-\delta_j(r_j - r_k - c)} & \text{if } r_k < r_j - c, \end{cases}$$

$$\mathbb{E}^{g}[\psi_j] = \prod_{\substack{k=1 \\ k \neq j}}^{n} \mathbb{E}^{g}[\psi_{jk}],$$

$$\mathbb{E}^{g}[\mathbb{1}_{\{i_k < i_j < r_k\}} \psi_{jk}] = \begin{cases} \frac{\delta_j}{\delta_j+\beta_{kj}} \frac{1}{2} \mathrm{e}^{-\delta_j(r_k - r_j + c)} & \text{if } r_k \geq r_j - c, \\ \frac{\delta_j}{\delta_j+\beta_{kj}} \frac{1}{2} \mathrm{e}^{-\delta_j(r_j - r_k - c)} & \text{if } r_k < r_j - c, \end{cases}$$

$$a(B_j, \boldsymbol{\theta}) = \frac{\gamma}{\delta_j}.$$

Equivalently, for gamma distributed infectious periods we have

$$\mathbb{E}^g[\psi_{jk}] = \begin{cases} 1 + \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\delta_j(r_k - r_j + c)}}{l! 2^m} \left(\left(\frac{\delta_j}{\delta_j + \beta_{kj}} \right)^m (\delta_j + \beta_{kj})^l - \delta_j^l \right) \\ \quad \times \mathbb{E}\left[((r_k - r_j + c + Y)^l) \mid Y \sim \Gamma(m, 2\delta_j) \right] \\ \qquad\qquad \text{if } r_k \geq r_j - c, \\ 1 - F_{m,\delta_j}(r_j - r_k - c) \left(1 - \left(\frac{\delta_j}{\delta_j + \beta_{kj}} \right)^m \right) \\ \quad + \sum_{l=0}^{m-1} \frac{\delta_j^{m-1} \mathrm{e}^{-\delta_j(r_j - r_k - c)}}{2^{l+1}\Gamma(m)} \left(\left(\frac{\delta_j}{\delta_j + \beta_{kj}} \right)^m \left(\frac{\delta_j + \beta_{kj}}{\delta_j} \right)^l - 1 \right) \\ \quad \times \mathbb{E}\left[(r_j - r_k - c + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\delta_j) \right] \\ \qquad\qquad \text{if } r_k < r_j - c, \end{cases}$$

$$\mathbb{E}^g[\psi_j] = \prod_{\substack{k=1 \\ k \neq j}}^{n} \mathbb{E}^g[\psi_{jk}],$$

$$\mathbb{E}^g\left[\mathbb{1}_{\{i_k < e_j < r_k\}} \psi_{jk} \right] = \begin{cases} \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\delta_j(r_k - r_j + c)}}{l! 2^m} \left(\frac{\delta_j}{\delta_j + \beta_{kj}} \right)^m (\delta_j + \beta_{kj})^l \\ \quad \times \mathbb{E}\left[(r_k - r_j + c + Y)^l \mid Y \sim \Gamma(m, 2\delta_j) \right] \\ \qquad\qquad \text{if } r_k \geq r_j - c, \\ \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\delta_j(r_j - r_k - c)}}{2^{l+1}} \left(\frac{\delta_j}{\delta_j + \beta_{kj}} \right)^m \left(\frac{\delta_j + \beta_{kj}}{\delta_j} \right)^l \frac{\delta_j^{m-1}}{\Gamma(m)} \\ \quad \times \mathbb{E}\left[(r_j - r_k - c + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\delta_j) \right] \\ \qquad\qquad \text{if } r_k < r_j - c, \end{cases}$$

$$a(B_j, \boldsymbol{\theta}) = \left(\frac{\gamma}{\delta_j} \right)^m.$$

3.4.11.2 Random length exposed periods

We may wish to extend the SEIR case to random length exposed periods. We now define c_j as the exposed period of individual j, ($c_j = i_j - e_j$). Then each c_j takes a random value from $f_E(\cdot \mid \boldsymbol{\theta_E})$, and for certain distributions f_E it may be possible to obtain a likelihood expression. By taking the expectation over all c_j of our likelihood expressions from Section 3.4.11.1, for example, we make similar arguments by conditioning on c_j's value, for each individual j.

For example, for the ψ_{jk} term with any pair j, k,

$$\mathbb{E}[\psi_{jk}] = \mathop{\mathbb{E}}_{c_j}\Big[\mathbb{E}[\psi_{jk} \mid c_j]\Big] = \mathop{\mathbb{E}}_{c_j}\left[\begin{cases} f(c_j) & \text{if } r_k \geq r_j - c_j \\ g(c_j) & \text{if } r_k < r_j - c_j \end{cases}\right],$$

where $f(\cdot)$ and $g(\cdot)$ are functions defined by the choice of infectious period. Then,

$$\mathbb{E}[\psi_{jk}] = \int_{c_j} \mathbb{E}[\psi_{jk} \mid u] f_E(u)\ du = \int_{c_j} \left(\begin{cases} f(u) & \text{if } u \geq r_j - r_k \\ g(u) & \text{if } u < r_j - r_k \end{cases}\right) f_E(u)\ du.$$

If c_j takes only a small number of fixed values, it may be possible to compute this integral directly as a sum. Similarly, if we assume that $c_j \sim \text{Exp}(\rho)$ (i.e. $f_E(c_j \mid \rho) = \rho e^{-\rho c_j}$), as well as exponential infectious periods, for example, we can also calculate the integral using the expression from Section 3.4.11.1 for fixed c_j:

$$\begin{aligned}
\mathbb{E}[\psi_{jk}] = &\int_0^{r_j - r_k} \left(\frac{\delta_j}{\beta_{kj} + \delta_j} + \frac{\beta_{kj}}{2(\beta_{kj} + \delta_j)} e^{-\delta_j(r_j - r_k - c_j)}\right) \rho e^{-\rho c_j}\ dc_j\ + \\
&\int_{r_j - r_k}^{\infty} \left(1 - \frac{\beta_{kj}}{2(\beta_{kj} + \delta_j)} e^{-\delta_j(r_k - r_j + c_j)}\right) \rho e^{-\rho c_j}\ dc_j \\
= &\frac{\delta_j}{\beta_{kj} + \delta_j} + \frac{\beta_{kj}\rho}{2(\beta_{kj} + \delta_j)(\rho - \delta_j)} \left(e^{-\delta_j(r_j - r_k)} - e^{-(\rho - 2\delta_j)(r_j - r_k)}\right) + \\
&\frac{\beta_{kj}}{2(\beta_{kj} + \delta_j)} e^{-\rho(r_j - r_k)}.
\end{aligned}$$

with similar calculations for the other likelihood terms required, depending on the PBLA version used ($\mathbb{E}[\chi_j \phi_j]$ and so on).

We note that this resulting expression for $\mathbb{E}[\psi_{jk}]$ is not particularly simple however, and it will be similarly so for the other likelihood terms, unlike the SIR or fixed exposed period cases. The expressions with gamma distributed infectious and exposure periods prove even more complex. This approach also certainly brings more approximation into the model, and an important question would be whether this affects the accuracy of the method.

In reality, however, we will usually choose to fix c_j to the same value for all j, as in Section 3.4.11.1. With random length exposure periods, we are trying to estimate two quantities (ρ and γ) from one piece of data (the removal time), which is problematic. Further exploration of the random exposure periods approach is certainly possible, but this will be beyond the scope of this work.

3.5 Conclusions

In this chapter, we have explored two new likelihood approximation methods for use in infectious disease modelling: the Eichner and Dietz method, as introduced in Section 3.3 as a generalisation of that used in Eichner and Dietz (2003), and a new series of PBLA methods, as introduced in Section 3.4. A summary of the different PBLA methods can be found in Table 3.2. A general theme of the approximations has been assuming independence between the interactions of individuals in the population, particularly in the PBLA method which assumes that all pairs of individuals make independent likelihood contributions. The overall aim was to obtain approximate likelihood expressions which do not require data augmentation, to avoid issues of correlation in missing data as well as computational issues which occur when using large data sets with DA-MCMC. The ED and PBLA methods may all be used within standard MCMC without data augmentation, since the likelihood expressions are independent of the infection times, or MLEs for the parameters of interest may also be obtained through any choice of optimisation scheme.

Explicit likelihood expressions have been derived for all of the approximation methods explored, for both exponential and gamma distributed infectious periods (in all cases where possible). We chose to focus on these distributions since they are widely used within the infectious disease modelling community and offer simple, interpretable likelihood expressions, but of course it may also be possible to extend the methods to other infectious period distributions of choice. As well as deriving the PBLA expressions analytically with

full integration arguments, we have also provided probabilistic arguments for these (for PBLA I at least, though other PBLA versions may be obtained similarly). This has hopefully provided further understanding of how the likelihood expressions are obtained, and the contextual meaning of the approximations made.

We have focused on the application of the approximation methods to the standard SIR model until Section 3.4.11, which discussed their use with SEIR methods. This has really been limited so far to exposure periods of fixed length, but there is scope to extend this to random length exposure periods in the future. We also explored another limitation of the PBLA method in Section 3.4.10; that we require all non-equal removal times for the method to approximate well. Although it is possible to avoid this problem by simply jittering removal times slightly, this of course introduces some more approximation to the model in terms of the order in which the removals are jittered. For data sets with large numbers of equal removal times, it is possible that this might have a significant impact, and so the amount of jittering applied must be carefully considered so as not to impact analysis.

We have found that there are a large number of approximations which may be made to the true likelihood to result in a tractable likelihood expression, and a number of these have been explored with the different PBLA versions. We next wish to compare these different versions to explore if any offer accuracy or computational advantages. Although we have theoretically defined our approximation methods, another key issue is of course their practical implementation and performance as compared to existing methods. We will explore these points in Chapter 4, firstly with a series of simulation studies to compare the different methods both with each other and with standard DA-MCMC, and then with application to a number of real-life data sets.

Table 3.2: Table summarising the PBLA methods and associated assumptions explored in Chapter 3. All PBLA versions follow the initial approximation of independence over individuals, as in Equation 3.2.4.

Method	Description
PBLA I	The basic PBLA method. As well as the basic independence assumption (1) $\mathop{\mathbb{E}}_{\mathbf{i},i_\kappa}\Big[\big(\prod_{\substack{j=1\\j\neq\kappa}}^{n}\chi_j\psi_j\phi_j\big)\phi_\kappa\pi(i_\kappa\mid\kappa)\Big]\approx\Big(\prod_{\substack{j=1\\j\neq\kappa}}^{n}\mathop{\mathbb{E}}_{\mathbf{i},i_\kappa}\big[\chi_j\psi_j\phi_j\big]\Big)\mathop{\mathbb{E}}_{i_\kappa}\big[\phi_\kappa\pi(i_\kappa\mid\kappa)\big]$, PBLA I assumes (2) independence between $\chi_j\phi_j$ and ψ_j, and sets (3) $\mathbb{E}[\psi_j]\approx\prod_{\substack{k=1\\k\neq j}}^{n}\mathbb{E}\big[\mathrm{e}^{-\beta_{kj}(r_k\wedge i_j-i_k\wedge i_j)}\big]$.
PBLA II	Applies a change of variable to the expectations, so they are with respect to $\mathbf{i}$, i_κ now with probability density function $\prod_{j=1}^{n}g_j(r_j-i_j)$. Essentially, all instances of γ in the likelihood are replaced with δ_j for individual j. This removes the need to calculate the ϕ_j terms, as well as the assumptions (in (1) and (2)) associated with this calculation.
PBLA III	Builds upon PBLA II by combining the expectations of χ_j and ψ_j, rather than assuming these are independent (2). This simplifies the calculations required, and avoids the introduction of unnecessary approximation.
PBLA IV	Only applicable for exponential infectious periods and large outbreak sizes, this method extends PBLA II by using a central limit theorem to form an approximation to the ψ_j terms. Rather than making assumption (3), we instead assume that $\mathbb{E}\Big[\prod_{j=1}^{n}\psi_j\Big]\approx\exp\Big(-\frac{\beta}{\delta N}\binom{n}{2}+\frac{1}{2\delta^2}\binom{n}{2}\big(\frac{2n-1}{3}\big)\big(\frac{\beta}{N}\big)^2\Big)$, using the fact that the sum over all pairs of individuals of the time infectious pressure is placed between them is approximately normally distributed for large n.
PBLA V	Also extending PBLA II, this method considers the infectious pressure from any individual j to any other individual k in combination with the pressure from k to j, rather than assuming these are independent. Instead of making assumption (3), this provides an alternative expression for the ψ_j term such that $\mathbb{E}\Big[\prod_{\substack{j=1\\j\neq\kappa}}^{n}\psi_j\Big]\approx\prod^{n}_{\substack{j,k=1\\k<j,k,j\neq\kappa}}\mathbb{E}\Big[\exp\big(-\beta_{kj}(\tau_{kj}+\tau_{jk})\big)\Big]\prod_{\substack{j=1\\j\neq\kappa}}^{n}\mathbb{E}\Big[\exp\big(-\beta_{\kappa j}\tau_{\kappa j}\big)\Big]$.

CHAPTER 4

Likelihood Approximation Method Simulation Studies and Applications

In Chapter 3 we introduced the generalized Eichner and Dietz (ED) likelihood approximation, as well as the Pair Based Likelihood Approximation (PBLA). This chapter will involve application of these approximation methods, firstly in a series of simulation studies to evaluate the performance of the methods, and then to a variety of real data sets where we compare analysis using the PBLA method to existing published analyses. For easy reference, Table 3.1 provides a summary of the notation used in the likelihood approximations, and a summary of the different PBLA methods can be found in Table 3.2.

In Section 4.1 we will assess the likelihood approximation methods through a series of simulation studies. We first seek to compare the accuracy of the approximation methods, both with each other and with standard DA-MCMC techniques. This will be explored in Section 4.1.1, for both exponential and gamma infectious periods. Section 4.1.2 will then include a more detailed comparison of the ED and PBLA methods only, using a larger number of simulations to more accurately assess their performance and which situations lead to either outperforming the other. In Section 4.1.3 we then use an additional sim-

ulation study to compare the different PBLA methods as defined in Chapter 3. However, we wish to assess not just the approximation methods' accuracy in parameter estimation, but also their computational speed. This will therefore be explored in Section 4.1.4.

Although we use simulated data to compare various versions of the approximation methods, of course it is also of interest to know how likelihood approximations perform with real data, compared to existing techniques. In sections 4.2, 4.3 and 4.4 we apply the PBLA method to various real data sets, encompassing a range of population structures, sizes and models for the infection process.

In Section 4.2 we analyse data concerning respiratory disease on the island of Tristan Da Cunha. We investigate how the PBLA method performs with a multi-type SIR model, comparing our estimates to a previous analysis by Hayakawa et al. (2003). Section 4.3 describes analysis of the West African Ebola virus outbreak of 2014. Here, we wish to compare the PBLA method to a previous analysis in Althaus (2014), who used a model featuring a time-dependent infection rate. Although this is not possible in the PBLA framework, we explore a proxy-time-dependence in the form of a heterogeneously mixing population where the infection rate is a function of the removal time. We also demonstrate the use of the PBLA method with an SEIR model, and highlight the importance of having data on a completed outbreak for the PBLA approach to approximate the true likelihood well. In Section 4.4 we consider data from the 2001 UK Foot and Mouth livestock epidemic. Here we use a heterogeneous mixing model with a spatial component, where the infection rate between any two farms will depend on the geographical distance between them as well as other covariates.

This range of analyses, both using simulated and real data, will highlight the strengths and the weaknesses of the likelihood approximation methods, which we summarise in Section 4.5.

4.1 Simulation Studies

4.1.1 Comparing ED, PBLA and DA-MCMC

To assess the Eichner and Dietz and PBLA methods, we first perform a simulation study comparing both to DA-MCMC, as this represents a gold standard for Bayesian analyses. We therefore compare the parameter estimates from the ED and PBLA methods to those from DA-MCMC, rather than to the true values the outbreaks were simulated from. For both exponential and gamma distributed infectious periods, we simulate 12 outbreaks under a variety of parameter values and population sizes, and perform parameter estimation for each using DA-MCMC, and MCMC with the ED and PBLA likelihoods. We compare the resulting parameter estimates to evaluate the likelihood approximation methods, found in Section 4.1.1.1 for the exponential case and 4.1.1.2 for the gamma case.

4.1.1.1 Exponentially Distributed Infectious Periods

For the homogeneously mixing SIR model with exponentially distributed infectious periods, we will now apply the ED and PBLA methods to 12 simulated data sets of varying population and final sizes. We compare the ability of the approximations to recover the true parameter values against standard DA-MCMC.

Each data set is a simulated outbreak from the given parameters, where in each simulation we start with one initial infective. For any simulations of final size 1, we discard and re-simulate. For analysis, we use MCMC with both the ED and PBLA likelihoods, as well as performing standard DA-MCMC.

For the ED likelihood, the integral involved in the calculation of the likelihood (as given in Equation (3.3.4)) must be computed numerically. In this study we use a simple trapezium rule method for this, which was found to provide suf-

ficiently accurate results when compared with more complex techniques. We set the lower limit of the integral to significantly less than the expected value of the initial infection time. Since this time is when infectious pressure begins to be applied, we can hence be confident that our numerical integration region is capturing the entire epidemic process. We selected 1000 as the lowest number of trapezia for which increasing this number did not impact the likelihood values at the degree of accuracy used in the analysis. All PBLA analysis is performed with the PBLA III method, since this is applicable for both exponential and gamma infectious periods and we expect it to perform better than PBLA I or II (we will explore this further in Section 4.1.3). For all MCMC analyses we take 10,000 samples after an initial burn-in period of 500. We perform Gaussian random walk updates for ED and PBLA, where the variances of the proposed parameter values have been tuned to result in a well-mixing chain. We use independent low rate (10^{-4}) exponential priors for the parameters β and γ.

The results of the simulation study, ordered by population size, are given in Table 4.1. These consist of posterior medians for infection rate β and removal rate γ, under each method. Table 4.2 contains the corresponding results for the estimation of R_0.

As we can see from Table 4.1, prediction using ED and PBLA with exponential infectious periods is much more similar to DA-MCMC for some data sets than others. Generally, the ED approximation estimates are much less similar to DA-MCMC than the PBLA estimates. This is especially true for smaller outbreaks such as simulation 2, where the ED estimates are almost double those from DA-MCMC. PBLA, however, has obtained relatively similar estimates to DA-MCMC in this case. Both methods seem to consistently overestimate both β and γ, but in almost all cases we see that PBLA offers estimates closer to DA-MCMC than ED.

Both methods do seem to struggle when the proportion of infectives is high. We will discuss this further in Section 4.1.2, but we see here that for simula-

Table 4.1: Estimates of infection rate β and removal rate γ for 12 simulated data sets using standard DA-MCMC, the ED approximation and PBLA, for an SIR model with exponential infectious periods.

	True (β, γ)	n/N	ED MCMC medians	PBLA MCMC medians	DA-MCMC medians
1	(1.3,1.3)	27/50	(6.600,4.867)	(4.199,2.649)	(2.672,1.833)
2	(2.5,3.6)	4/60	(9.196,10.642)	(5.079,5.867)	(4.933,5.940)
3	(3.0,1.0)	56/60	(4.159,2.141)	(3.926,1.815)	(2.837,0.963)
4	(1.5,1.4)	26/100	(4.001,3.554)	(2.205,1.826)	(1.997,1.737)
5	(1.3,1.0)	52/100	(2.280,1.695)	(1.514,0.958)	(1.001,0.702)
6	(2.3,1.4)	73/100	(2.631,1.641)	(2.059,1.047)	(1.954,1.029)
7	(3.0,2.0)	88/100	(0.257,0.140)	(0.243,0.112)	(0.119,0.049)
8	(1.6,1.0)	123/200	(3.940,2.716)	(2.788,1.591)	(2.326,1.489)
9	(1.9,1.2)	216/300	(4.131,2.659)	(3.840,1.960)	(2.062,1.167)
10	(4.0,1.0)	295/300	(7.177,3.527)	(9.276,4.556)	(4.419,1.112)
11	(2.0,1.3)	187/400	(3.629,2.797)	(2.309,1.540)	(2.374,1.759)
12	(0.2,0.1)	496/600	(0.867,0.506)	(0.729,0.347)	(0.207,0.099)

tions 3, 7, 10 and 12 (where the proportion of infectives is highest), the estimates are most dissimilar to DA-MCMC. This is also true for ED with very small outbreaks, though we see PBLA estimates well in these situations, such as simulation 2. In practice, we will less often be concerned with very small outbreaks, but further consideration why the methods struggle for larger outbreaks would be of use and will follow in Section 4.1.2. We lastly note that the population size does not appear to have a significant impact on estimation, just the proportion of infectives within it.

In terms of R_0, Table 4.2 shows that the ED and PBLA methods are both much better able to estimate R_0 than β and γ individually. Both no longer consistently overestimate, with estimates generally very similar to those obtained with standard MCMC. It seems that both approximation methods are unable to estimate β and γ well in all cases for exponential infectious periods, but are able to maintain their ratio to result in a reasonable R_0 estimate. In cases

Table 4.2: For the 12 simulated data sets in Table 4.1, this table includes the estimates of R_0 using standard DA-MCMC, the ED approximation and PBLA with exponential infectious periods.

	True R_0	n/N	ED MCMC R_0 medians	PBLA MCMC R_0 medians	DA-MCMC R_0 medians
1	1.0	27/50	1.386	1.595	1.458
2	0.694	4/60	0.900	0.878	0.840
3	3.0	56/60	1.977	2.173	2.917
4	1.071	26/100	1.134	1.237	1.141
5	1.3	52/100	1.364	1.586	1.424
6	1.643	73/100	1.612	1.975	1.901
7	1.5	88/100	1.831	2.143	2.449
8	1.6	123/200	1.453	1.751	1.555
9	1.583	216/300	1.568	1.954	1.772
10	4.0	295/300	2.030	2.044	3.947
11	1.538	187/400	1.304	1.506	1.354
12	2.0	496/600	1.716	2.100	2.100

where only R_0 is of interest rather than the individual infection and removal rates, this highlights that the approximation methods might be especially of use.

4.1.1.2 Gamma Distributed Infectious Periods

We test the approximation methods with gamma distributed infectious periods in the same way as the exponential case. To ensure the analysis is comparative across the two methods, we simulate 12 data sets with the same R_0 values and population sizes as the exponential data sets. We set shape parameter $m = 5$, keeping β the same as the exponential outbreaks and setting γ to keep R_0 constant. As in the exponential case, we will use the trapezium rule to compute the numerical integral in the ED likelihood (with the number of trapezia set to 1000 and the lower limit of the integral significantly less than the expected value of the initial infection time), and use version III of PBLA.

Table 4.3: Estimates of infection rate β and removal rate γ for 12 simulated data sets using standard DA-MCMC, the ED approximation and PBLA, for an SIR model with exponential infectious periods. Shape parameter $m = 5$ is fixed.

	True (β, γ)	n/N	ED MCMC medians	PBLA MCMC medians	DA-MCMC medians
1	(1.3,6.5)	24/50	(1.493,5.576)	(1.412,5.100)	(1.359,4.914)
2	(2.5,18.0)	2/60	(0.834,4.908)	(0.732,4.561)	(0.675,4.263)
3	(3.0,5.0)	58/60	(2.635,4.925)	(2.417,4.685)	(2.953,3.979)
4	(3.0,10.0)	3/100	(69.390,392.570)	(29.549,175.894)	(29.806,173.416)
5	(1.3,5.0)	18/100	(1.221,5.596)	(1.004,4.458)	(1.023,4.663)
6	(1.5,7.0)	89/100	(1.332,3.044)	(1.238,2.838)	(1.393,2.773)
7	(2.3,7.0)	95/100	(2.359,4.686)	(2.181,4.421)	(2.515,3.850)
8	(1.6,5.0)	120/200	(1.344,4.551)	(1.263,4.119)	(1.249,4.146)
9	(1.9,6.0)	258/300	(1.693,4.106)	(1.561,3.645)	(1.694,3.685)
10	(4.0,5.0)	296/300	(3.558,6.426)	(3.206,6.170)	(4.161,4.744)
11	(2.0,6.5)	352/400	(1.611,3.685)	(1.495,3.390)	(1.786,3.700)
12	(0.2,0.5)	3/600	(10.415,60.049)	(8.315,51.423)	(8.354,50.496)

Again, for all MCMC algorithms, we take 10,000 samples after an initial burn-in period of 500, and for ED and PBLA we perform Gaussian random walk updates with low rate (10^{-4}) exponential priors for β and γ.

The results of this are given in Table 4.3, again ordered by population size. The posterior medians obtained from DA-MCMC may be compared to the posterior medians from MCMC with the ED and PBLA likelihoods. We also give the results of R_0 estimation in Table 4.4.

Table 4.3 shows that, in the case of gamma distributed infectious periods, both ED and PBLA generally estimate the parameters much more closely compared to DA-MCMC than in the exponential case. PBLA, however, continues to offer closer estimates to DA-MCMC than ED in almost all cases. Again, this is especially true for smaller outbreaks such as simulations 4 and 12, where the PBLA estimates are similar to DA-MCMC but the ED estimates are con-

siderably larger. The results in Table 4.3 suggest, as in the exponential case, that the total population size does not impact the accuracy of estimation of the approximation methods, but rather the proportion of infectives.

In terms of cases where the proportion of infectives is high such as simulations 3, 7 and 10, both PBLA and ED estimates are far closer to those from DA-MCMC than in the exponential case. We also no longer see consistent overestimation of the parameters, and again PBLA tends to provide better estimation than the ED method.

Considering Table 4.4 which includes the estimates of R_0 for the gamma infectious period simulations, we see that R_0 estimation is generally similar when using the approximation methods or DA-MCMC. Again, the ED and PBLA methods are better able to estimate R_0 than β and γ individually. Despite the improvement in β and γ estimation as compared to the exponential case, the estimation of R_0 is interestingly very similar with gamma infectious periods and exponential. Even for very small or large outbreaks, the R_0 estimates using the approximation methods are very close to those from DA-MCMC, without any consistent over- or under-estimation.

4.1.1.3 Conclusions

Overall, from this simulation study we have seen that the PBLA and ED methods perform much more similarly to DA-MCMC for gamma infectious periods than exponential. For very small proportions of infectives, the ED method especially struggles, and for very large proportions of infectives both methods struggle, to obtain similar estimates to DA-MCMC. However, generally for gamma infectious periods with roughly $20 - 80\%$ of the population infected, both methods (and particularly PBLA) estimate the parameters well. The basic reproduction number R_0 is also generally estimated well by both methods, under both exponential and gamma distributed infectious periods.

We have highlighted that, especially for large outbreaks where DA-MCMC is

Table 4.4: For the 12 simulated data sets in Table 4.3, this table includes the estimates of R_0 using standard DA-MCMC, the ED approximation and PBLA with exponential infectious periods. Shape parameter $m = 5$ is fixed.

	True R_0	n/N	ED MCMC R_0 medians	PBLA MCMC R_0 medians	DA-MCMC R_0 medians
1	1.0	24/50	0.273	0.280	0.280
2	0.694	2/60	0.181	0.165	0.166
3	3.0	58/60	0.532	0.514	0.733
4	1.5	3/100	0.182	0.175	0.177
5	1.3	18/100	0.221	0.220	0.221
6	1.071	89/100	0.434	0.439	0.495
7	1.643	95/100	0.502	0.492	0.645
8	1.6	120/200	0.295	0.366	0.301
9	1.583	258/300	0.413	0.428	0.461
10	4.0	296/300	0.552	0.519	0.877
11	1.538	352/400	0.434	0.441	0.484
12	2.0	3/600	0.183	0.162	0.171

particularly cumbersome in terms of both mixing and speed (we will further explore this in Section 4.1.4), and even more so where inference about the reproduction number is key, either ED or PBLA may offer a useful alternative. PBLA also seems to offer better estimation than ED overall, though we will explore this further in the following section.

Although this section has sufficed as an introductory exploration of the performance of the approximation methods, it is important to note that in this study we have only simulated one outbreak from each set of parameter values. These outbreaks therefore might not necessarily represent what is typical. In our second study, we will simulate a larger number of outbreaks per set of parameter values, and look at the average performance of the ED and PBLA methods.

4.1.2 A more in-depth study comparing PBLA and the Eichner and Dietz approximation

We continue the simulation studies with a more in-depth comparison of the Eichner and Dietz and PBLA III approximations, in order to more clearly demonstrate the differences in performance between them. Although in Section 4.1.1 we compared PBLA and the ED method (to DA-MCMC) for a single simulation under different sets of parameter values, here we simulate a large number of outbreaks for each set of values. We present this first for exponentially distributed infectious periods, and then for gamma.

4.1.2.1 Exponential Infectious Periods

Here we describe a comparison of the Eichner and Dietz likelihood approximation and Pair-Based Likelihood Approximation (version III) for exponential infectious periods. A selection of parameter values are chosen, varying the infection rate β and the removal rate γ as well as the population size N, which in turn influence the final size n and determine the reproduction number R_0. We

choose a wide range of parameter values, to best explore the accuracy of the methods in different situations. For each set of parameter values to be tested we simulate a large number of outbreaks, ranging from 200 to 1000 depending on the computational demands of the population/outbreak size. For each simulation, we then maximise both the ED and PBLA likelihoods. We provide density plots of the distribution of the MLEs obtained from each simulation for comparison of the two methods, and investigate the impact of varying the parameter values. Note that these are not true densities since the MLEs do not exactly represent samples from a probability distribution, but the plots suffice for visualisation purposes.

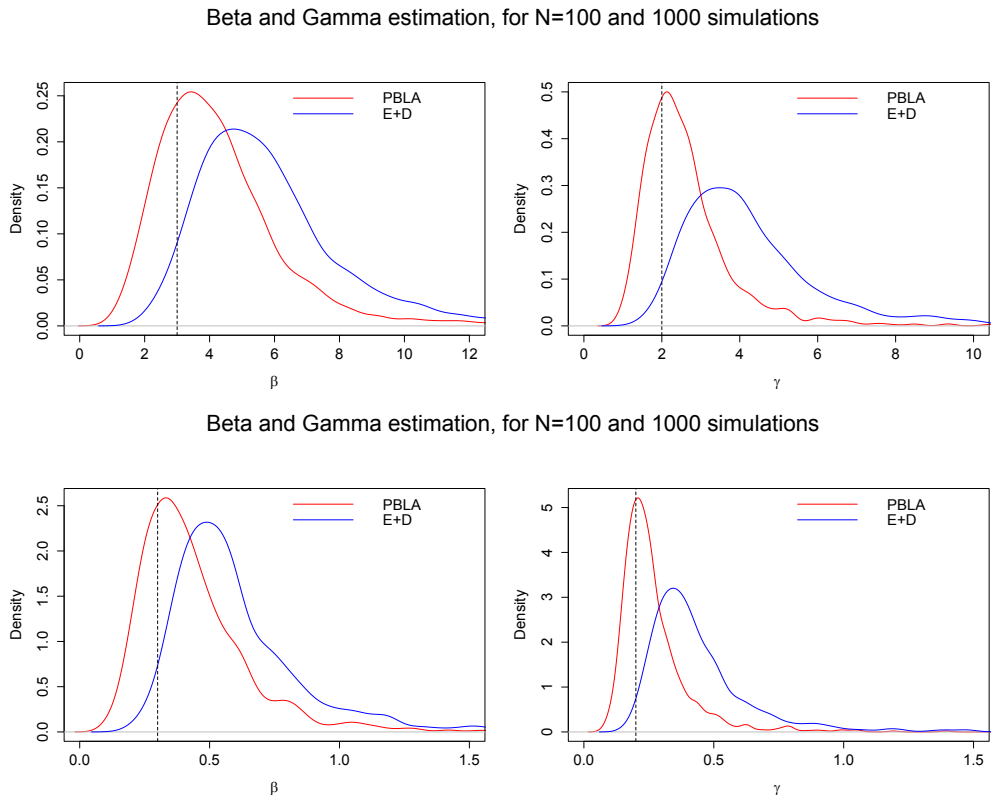


Figure 4.1: To compare the impact of varying β and γ, these figures show densities of MLEs from both the ED and PBLA III approximation methods with exponential infectious periods. Data are from 1000 simulations with $N = 100$ and true values $\beta = 3$, $\gamma = 2$ in the first plot and $\beta = 0.3$, $\gamma = 0.2$ in the second.

Varying parameters β and γ together

To compare varying values of rates β and γ whilst keeping N and R_0 constant, we must vary β and γ whilst keeping their ratio equal. Figure 4.1 contains density plots of the maximum likelihood estimates for 1000 simulations under two sets of values of β and γ. There is no clear difference in estimation performance between the two, despite a 10 times reduction in the true values of both β and γ between the first and second plots. The accuracy of parameter estimation is not considerably changed; at least as long as population size N remains constant. We see that the PBLA method estimates both parameters much more closely, in both cases.

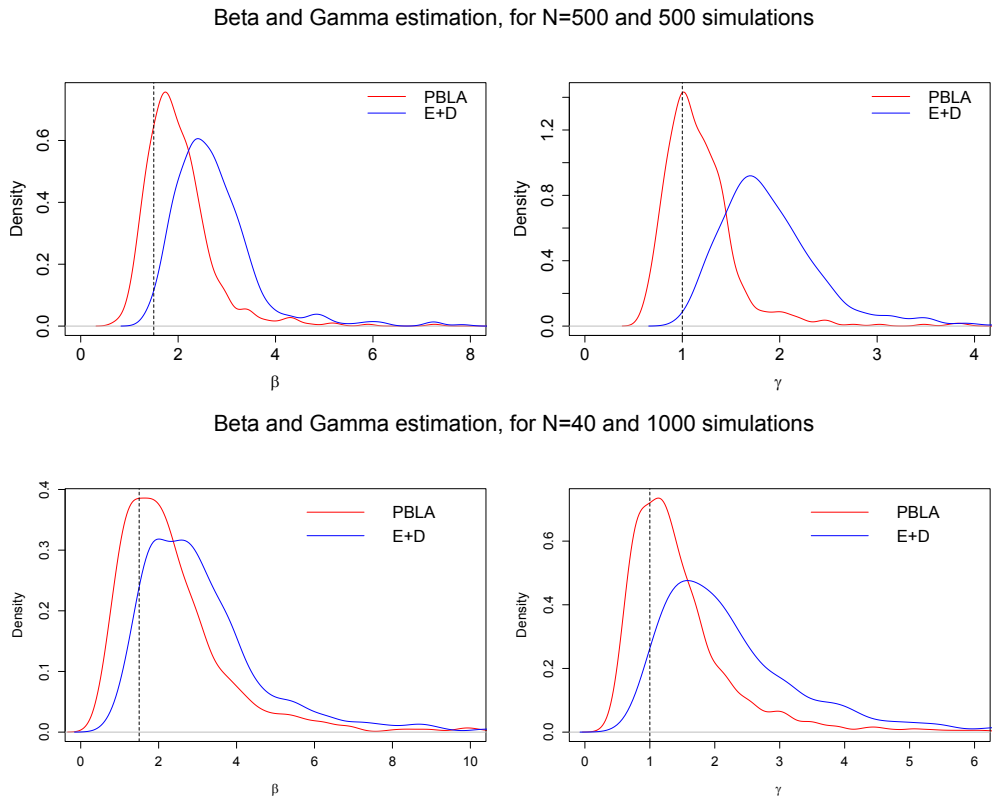


Figure 4.2: To compare the impact of varying N, these figures show densities of MLEs from both the ED and PBLA III approximation methods with exponential infectious periods. Data are from 500 simulations with $N = 500$ and 1000 simulations with $N = 40$, respectively, where in both the true values are $\beta = 1.5$, $\gamma = 1$.

Varying population size N

When we vary the value of N alone, we find that this also does not greatly

affect the estimation of either method. Figure 4.2 shows no significant difference between the two plots, despite the fact that they use values $N = 500$ and $N = 40$, respectively. The only difference seems to be that for larger N there is slightly less variance in the distribution of the MLEs obtained. Again, we see that the PBLA method better estimates both β and γ. A different number of simulations has been used in each case due to computational restraints, but this was not found to significantly impact the analysis.

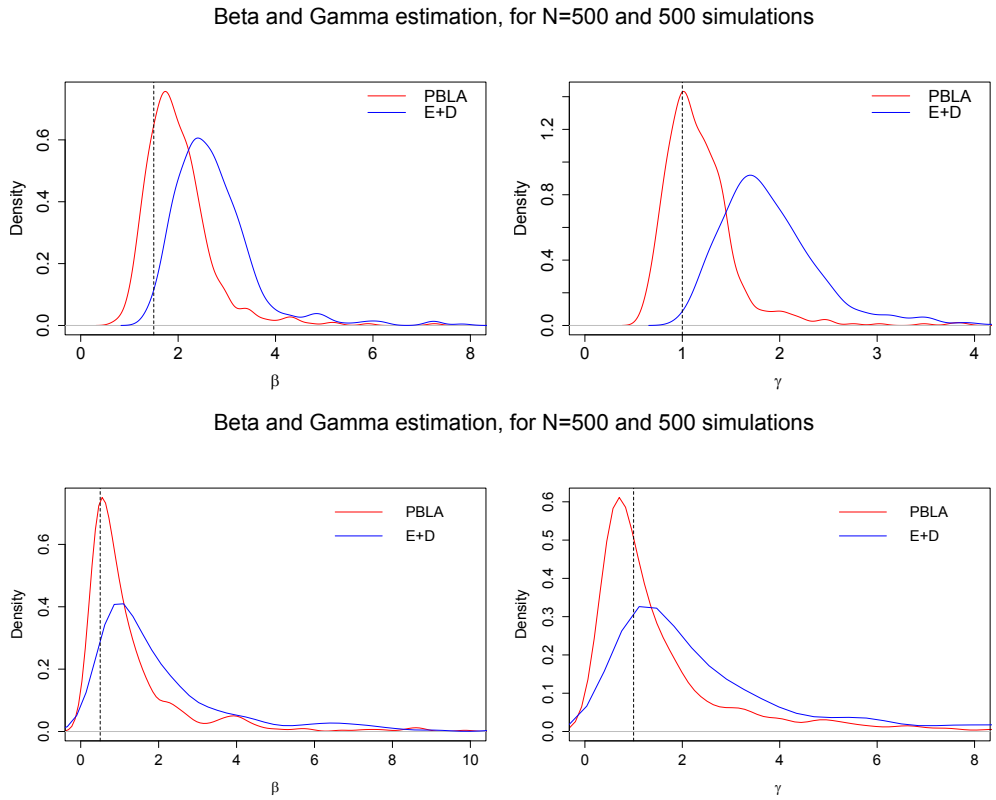


Figure 4.3: To compare the impact of varying R_0, these figures show densities of MLEs from both the ED and PBLA III approximation methods with exponential infectious periods. Data are from 500 simulations with $N = 500$, and true values $\beta = 1.5$, $\gamma = 1$ in the first plot and $\beta = 0.5$, $\gamma = 1$ in the second. This leads to R_0 values of 1.5 and 0.5, respectively.

Varying basic reproduction number R_0

Recall that reproduction number $R_0 = \frac{\beta}{\gamma}$ represents the average number of secondary cases a given infective causes in an entirely susceptible population.

We see that varying this can have a considerable impact on the efficacy of our approximation methods.

In the first plot in Figure 4.3, the true value of R_0 is 1.5, leading to an average proportion of 0.5 of the individuals in the population infected. The parameters β and γ are set as 1.5 and 1, respectively. We see again that estimation of both parameters is fairly good. The PBLA method performs better than ED.

We compare this to the second plot in Figure 4.3, which shows the density of estimates when R_0 is smaller, at 0.5. We see that estimation for β and γ is similarly good to $R_0 = 1.5$, and for the ED method is improved. It appears that for up to at least half of the population infected, both methods are able to estimate the true parameter values fairly well, with PBLA providing slightly more accurate results.

On the other hand, the first plot in Figure 4.4 shows the density of estimates when $R_0 = 3$, resulting in an average proportion of 0.95 of individuals becoming infected. We see much less accuracy in estimation for both methods, with the ED approximation now obtaining the closest results but generally both methods failing to acquire an accurate estimate. This is also seen in the second plot in Figure 4.4 where $R_0 = 2$ (an average of 77% of individuals infected), though here the ED method no longer performs significantly better. Further testing indicates that the drop in accuracy for both methods occurs gradually as the proportion of infectives increases, from around 0.6, and that neither method is consistently better under these conditions for exponential infectious periods. We will discuss this further with our concluding remarks.

Comparing accuracy of R_0 prediction

As well as considering the methods' accuracy in estimating β and γ, we may consider the accuracy of R_0 estimation. This is found to be similar for varying β, γ and N in that there is little impact on accuracy, but again both methods struggle when the true value of R_0 is large. Figure 4.5 shows densities for the estimates of R_0 across a large number of simulations when $R_0 = 1.5$ and $R_0 = 3$. We see that for $R_0 = 1.5$, leading to approximately half of the

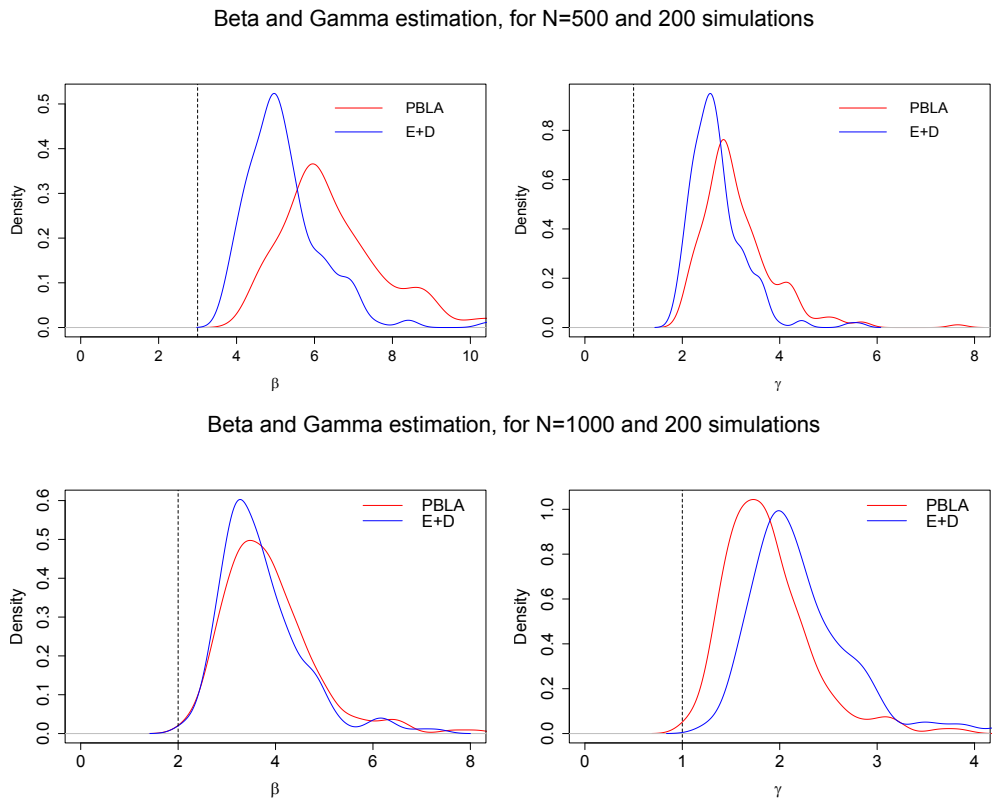


Figure 4.4: To further compare the impact of varying R_0, these figures show densities of MLEs from both the ED and PBLA III approximation methods with exponential infectious periods. Data are from 200 simulations with $N = 500$ and $N = 1000$, respectively, as well as true values $\beta = 3$, $\gamma = 1$ in the first plot and $\beta = 2$, $\gamma = 1$ in the second. This leads to R_0 values of 3 and 2.

population infected, R_0 is estimated fairly well by the mean of the MLEs, and particularly by the ED method, though there is some interesting bimodality observed. However when R_0 is increased to 3, resulting in on average 0.95 of the population infected, the estimation of both methods is considerably worse (perhaps PBLA being slightly more robust to this, though arguably estimation is so poor that this is meaningless). Overall, and despite this, in all cases where estimation is generally good, the ED method is seen to offer slightly improved estimation for R_0.

Concluding remarks

We have found that for both the ED and PBLA methods with exponential in-

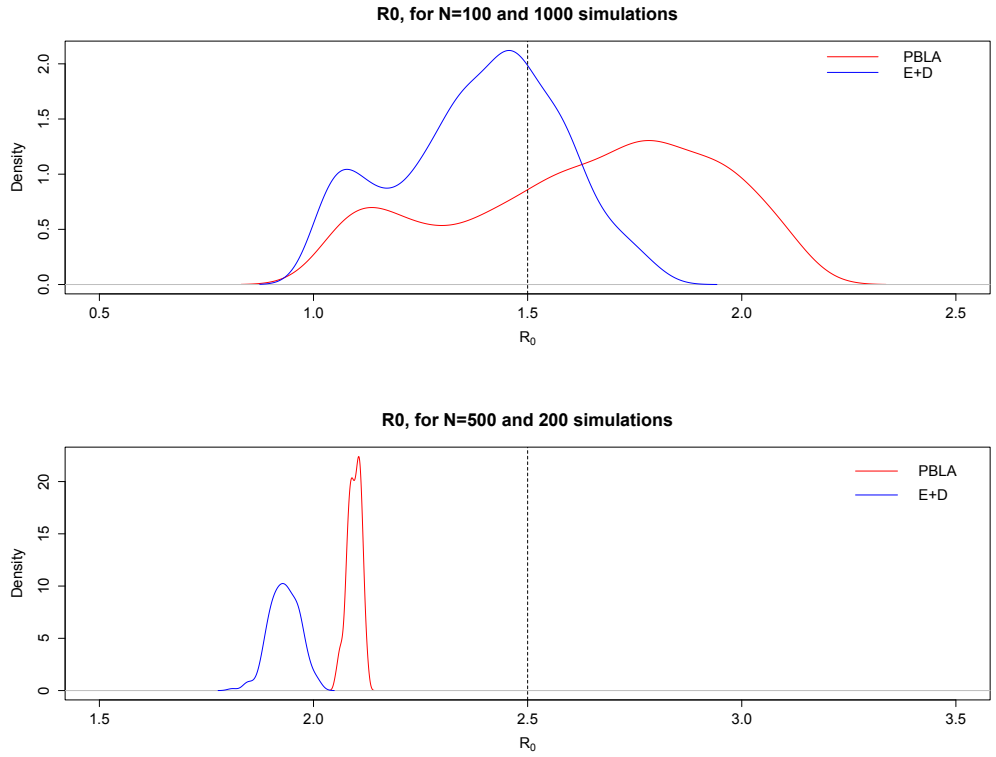


Figure 4.5: To compare R_0 estimation, these figures show densities of MLEs from both the ED and PBLA III approximation methods with exponential infectious periods. Data are from, respectively, 1000 simulations with $N = 100$, $\beta = 1.5$ and $\gamma = 1$, and 200 simulations with $N = 500$, $\beta = 2.5$, $\gamma = 1$. This leads to R_0 values of 1.5 and 2.5.

fectious periods, varying β and γ whilst maintaining their ratio or varying N does not have a significant effect on the accuracy of the methods. However, the size of reproduction number R_0, does have a big impact on their efficacy. Around half of the population infected appears optimum for both methods, with PBLA consistently outperforming ED. However, when the average proportion of infectives increases to above around 0.6, both methods become limited in their efficacy, with estimates far further from their true value. The situation is similar when considering estimation of R_0. Both methods seem to perform well except in cases where the proportion of infectives in the population is very high, in which case the accuracy of estimation rapidly deteriorates.

In general, the ED approximation appears to estimate R_0 more closely, though perhaps less so for larger true R_0 values. PBLA, on the other hand, estimates both β and γ much more closely than the ED method, and so, for exponential infectious periods at least, may be seen as a considerable improvement.

It seems feasible that the limited performance when a large proportion of the population is infected is due to the ψ term in the likelihood (recall e.g. Equation (3.2.1) which defines the different likelihood terms), concerning the avoidance of infection until an infective's exposure. This term involves the most approximation under both ED and PBLA, and further analysis indicates that as the proportion of infectives varies, $\mathbb{E}[\psi_j]$ varies greatly also. Returning to the PBLA I structure for comparison of the different elements of the likelihood, we find that for outbreaks where around half of the population is infected ($R_0 = 1.5$) the ψ term is of size roughly half of the ϕ term. As we increase R_0, leading to a larger proportion of infectives, ψ increases in relation to ϕ. For example, $R_0 = 3$ yields ψ four times larger than ϕ, and $R_0 = 5$ yields ψ twenty times larger than ϕ (for $N = 100$). This implies that as we increase R_0, both methods have a higher contribution from more highly approximated terms, and hence poorer estimation would be expected.

4.1.2.2 Gamma Infectious Periods

Next we consider a comparison of the likelihood approximation methods for gamma infectious periods, with shape parameter m and rate parameter γ. We expect the methods to perform better in this case since as shape parameter m increases, if the mean is fixed then the infectious periods become less variable. There is hence less uncertainty in the expectations over pairs of infection times, and so we would expect the approximations to be closer to the truth. Again a range of parameter values are explored, and we simulate here at least 500 outbreaks in each case. We compare the accuracy of the methods by considering density plots of the MLEs obtained.

Varying parameters β and γ together

First we vary β and γ together, whilst keeping R_0 constant. Figure 4.6 displays this for three decreasing sets of β and γ values, with m fixed at 2. We see that, as in the exponential case, a ten and then a hundred times reduction in both β and γ has little impact on the accuracy of estimation; there is almost no visible difference between the densities. Both methods perform well here, with no considerable difference between the two in terms of accuracy.

Varying population size N

Figures 4.7 and 4.8 explore varying N, with fixed shape $m = 2$. As in the exponential case, we note that as the total population size increases, the mean estimated values of β and γ do not change significantly, but the variance of these estimates decreases. Even for small values of N, both methods are able to estimate β and γ well (with slight overestimation of γ), and PBLA continues to offer very similar estimation to the ED method.

We also investigate the impact of varying N on the bias and mean squared error (MSE) of our β and γ estimates. Figures 4.9 and 4.10 show the bias and mean squared error, respectively, where we test with shape parameter m equal to 2 and 8. We see that an increase in population size leads to a reduction in both bias and mean squared error. Since the mean squared error is equal to the variance plus the squared bias of an estimate, this agrees with what we noted when investigating the estimates of β and γ obtained. The PBLA method generally seems to provide less biased estimates of the parameters, with similar mean squared error to the ED method.

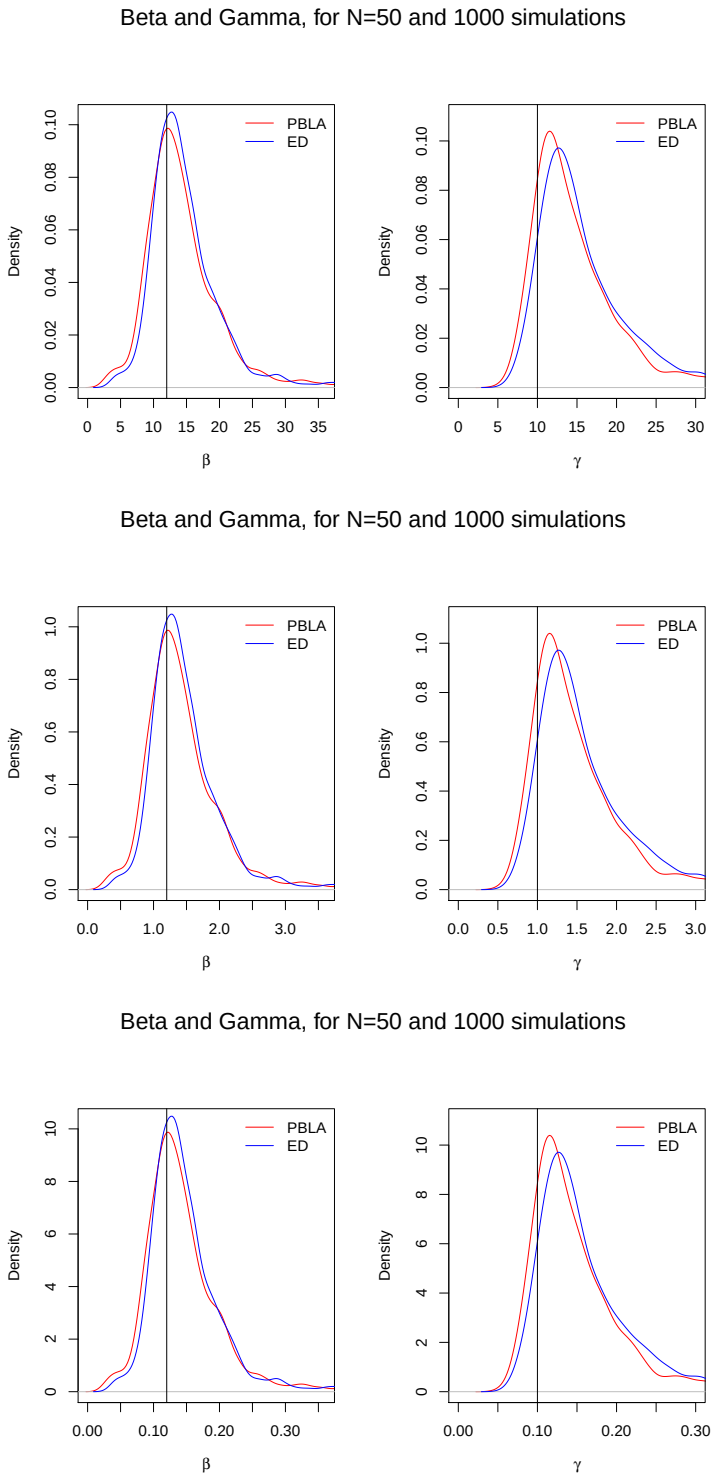


Figure 4.6: To compare the impact of varying β and γ, these figures show densities of MLEs from both the ED and PBLA III methods with gamma infectious periods. Data are from 1000 simulations with $N = 50$ and shape $m = 2$ in all cases, where the true values are $\beta = 12$ and $\gamma = 10$, $\beta = 1.2$ and $\gamma = 1$, and $\beta = 0.12$ and $\gamma = 0.1$, respectively.

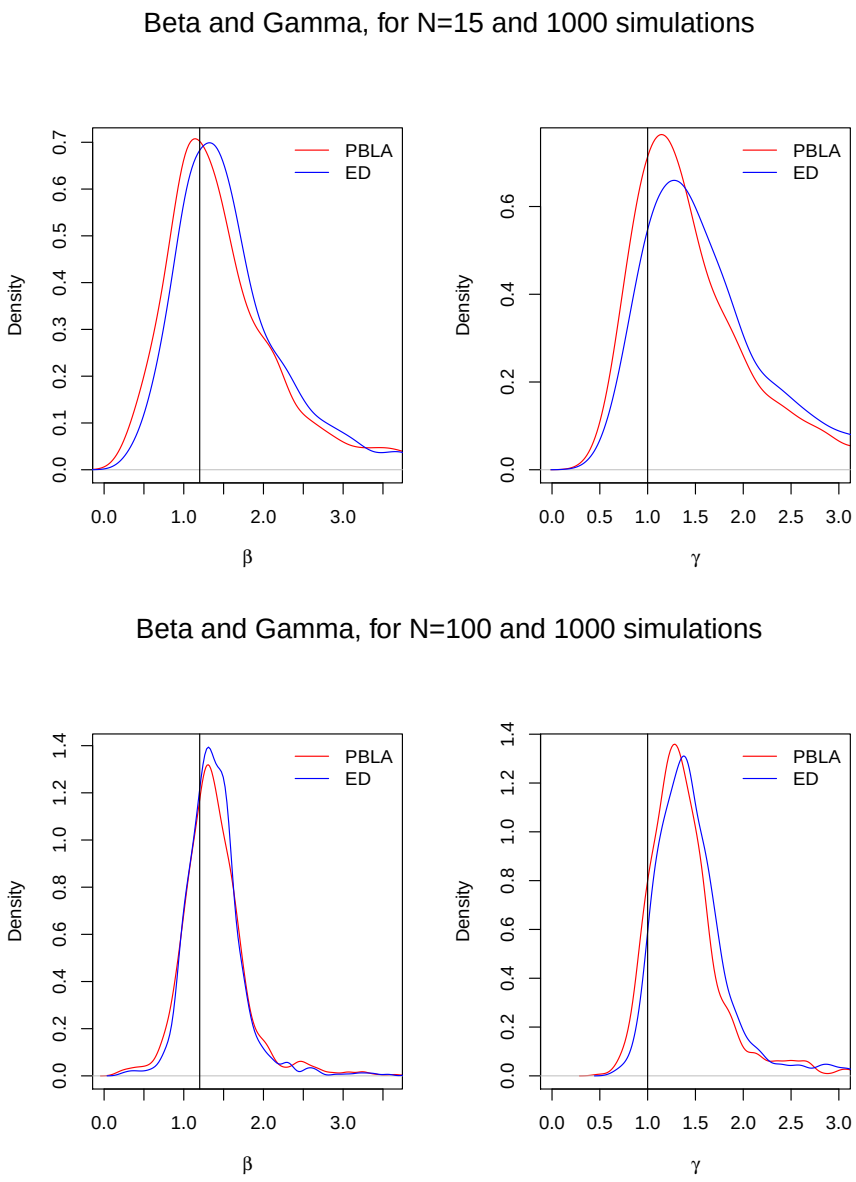


Figure 4.7: To compare the impact of varying N, these figures show densities of MLEs from both the ED and PBLA III methods with gamma infectious periods. Data are from 1000 simulations with shape $m = 2$, $\beta = 1.2$ and $\gamma = 1$. In the upper plots $N = 15$ and in the lower plots $N = 100$.

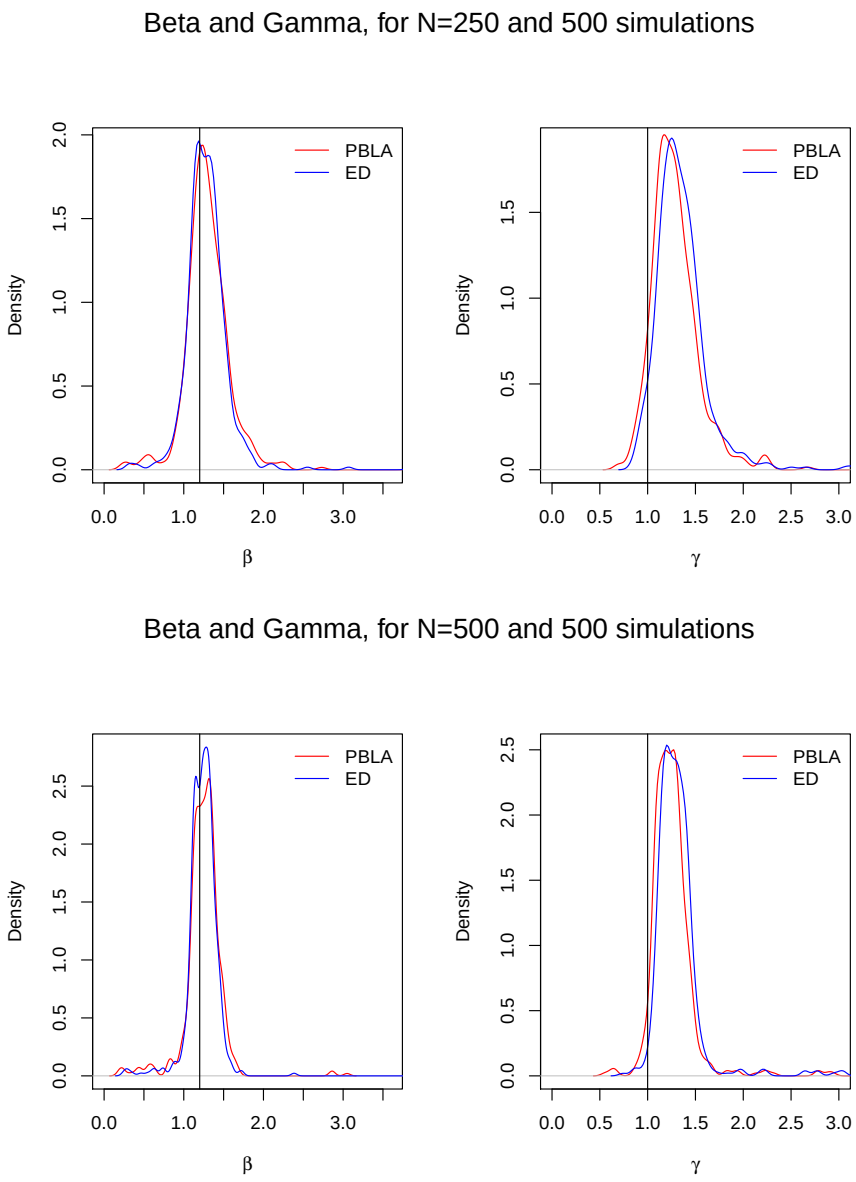


Figure 4.8: To compare the impact of varying N, these figures show densities of MLEs from both the ED and PBLA III methods with gamma infectious periods. Data are from 500 simulations with shape $m = 2$, $\beta = 1.2$ and $\gamma = 1$. In the upper plots $N = 250$ and in the lower plots $N = 500$.

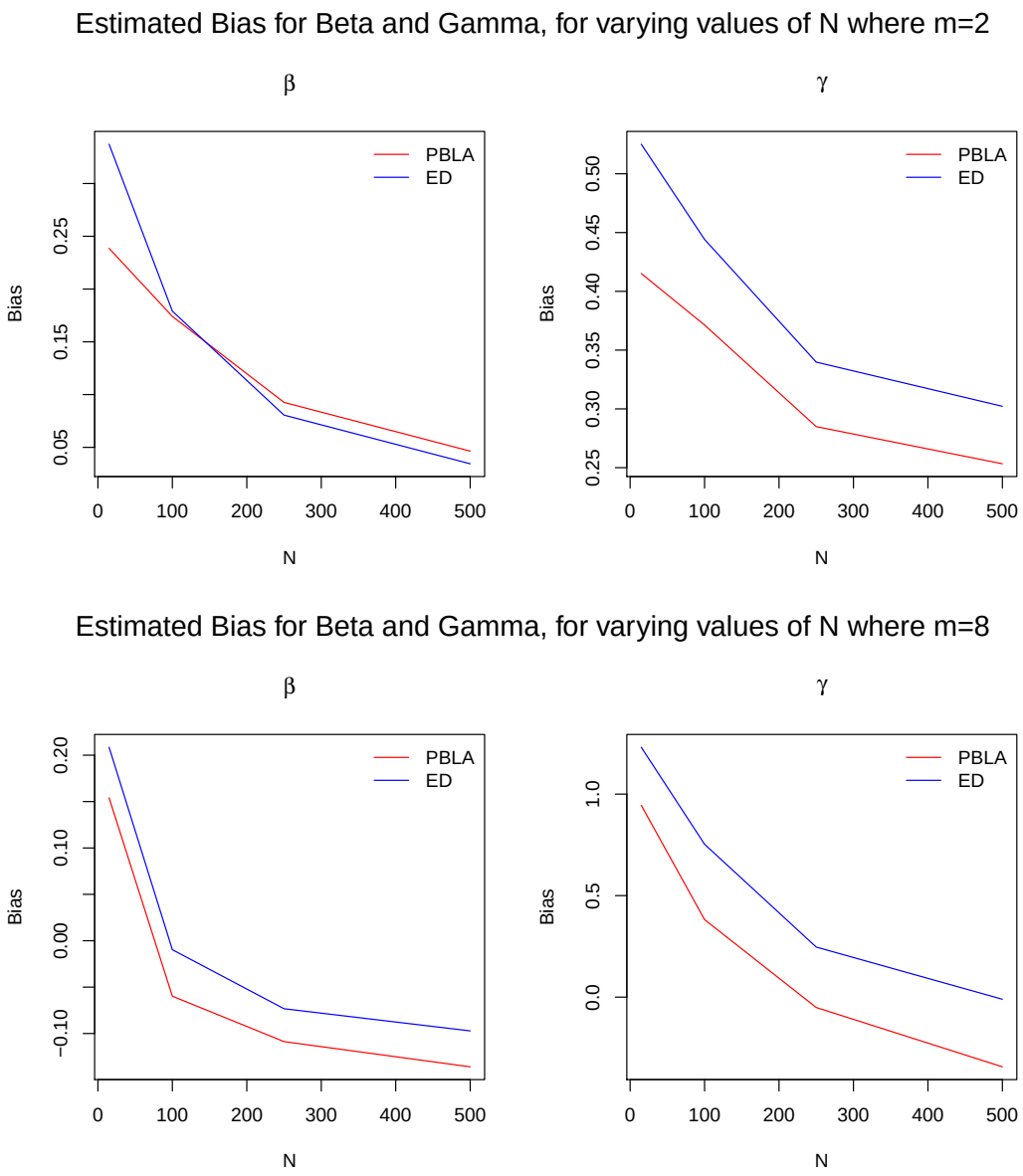


Figure 4.9: These figures show the bias in estimating parameters β and γ as N varies, for both the PBLA and ED methods. Shown are estimated values with shape parameter m equal to 2 and 8, where in all cases 1000 outbreaks were simulated.

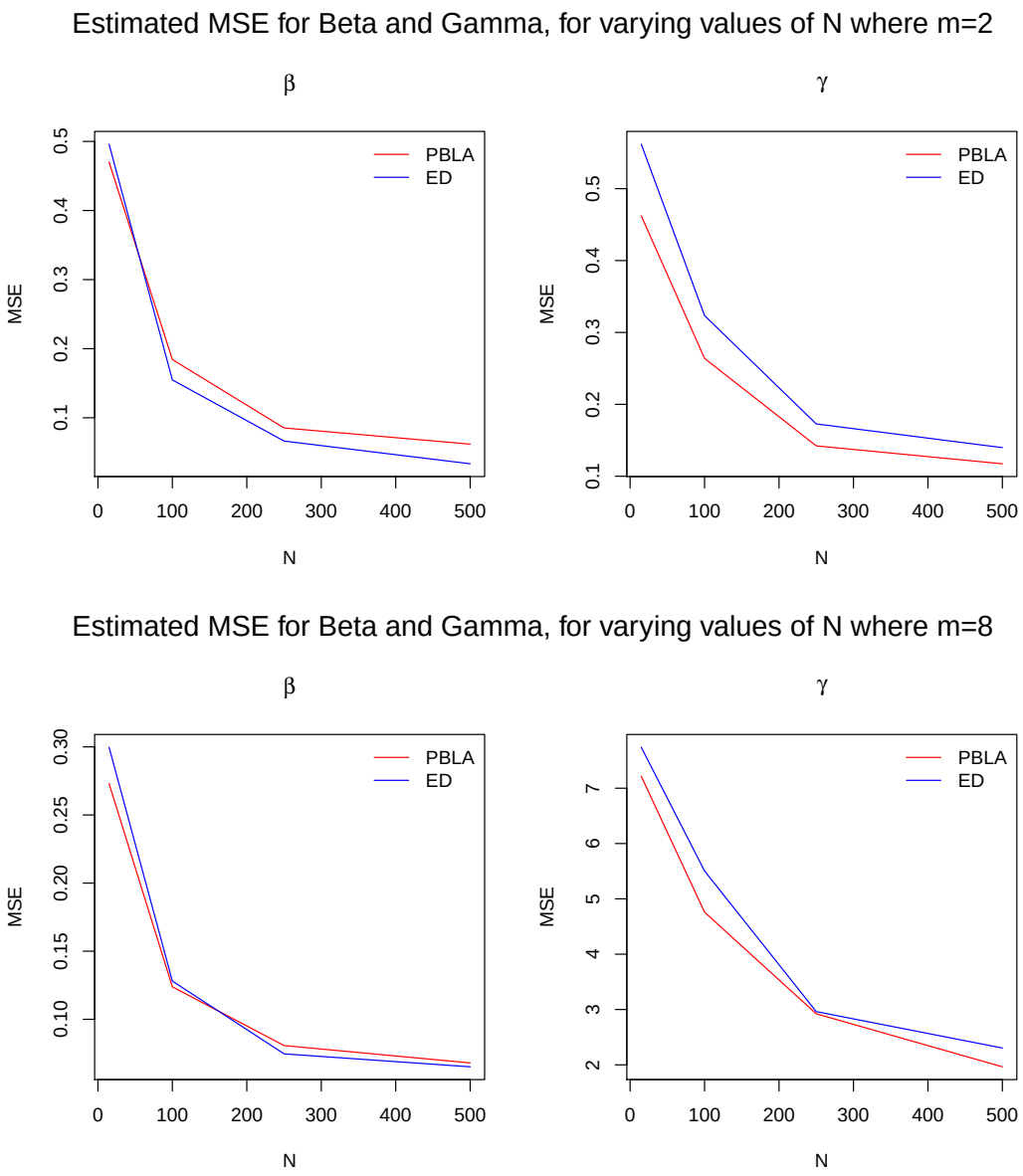


Figure 4.10: These figures show the mean squared error in estimating parameters β and γ as N varies, for both the PBLA and ED methods. Shown are estimated values with shape parameter m equal to 2 and 8, where in all cases 1000 outbreaks were simulated.

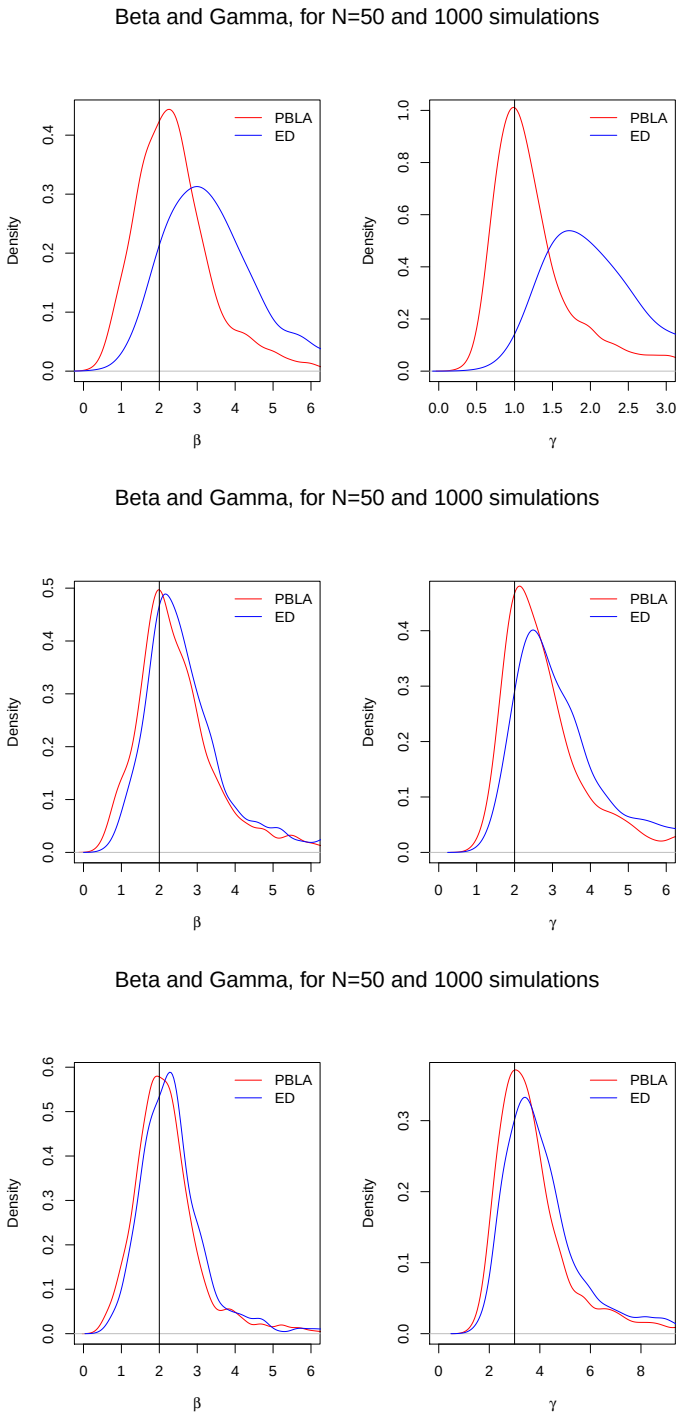


Figure 4.11: To compare the impact of varying m, these figures show densities of MLEs from both the ED and PBLA III methods with gamma infectious periods. Data are from 1000 simulations with $N = 50$ and $\beta = 2$. In the upper plots the true values are $m = \gamma = 1$, in the middle plots $m = \gamma = 2$ and in the lower plots $m = \gamma = 3$.

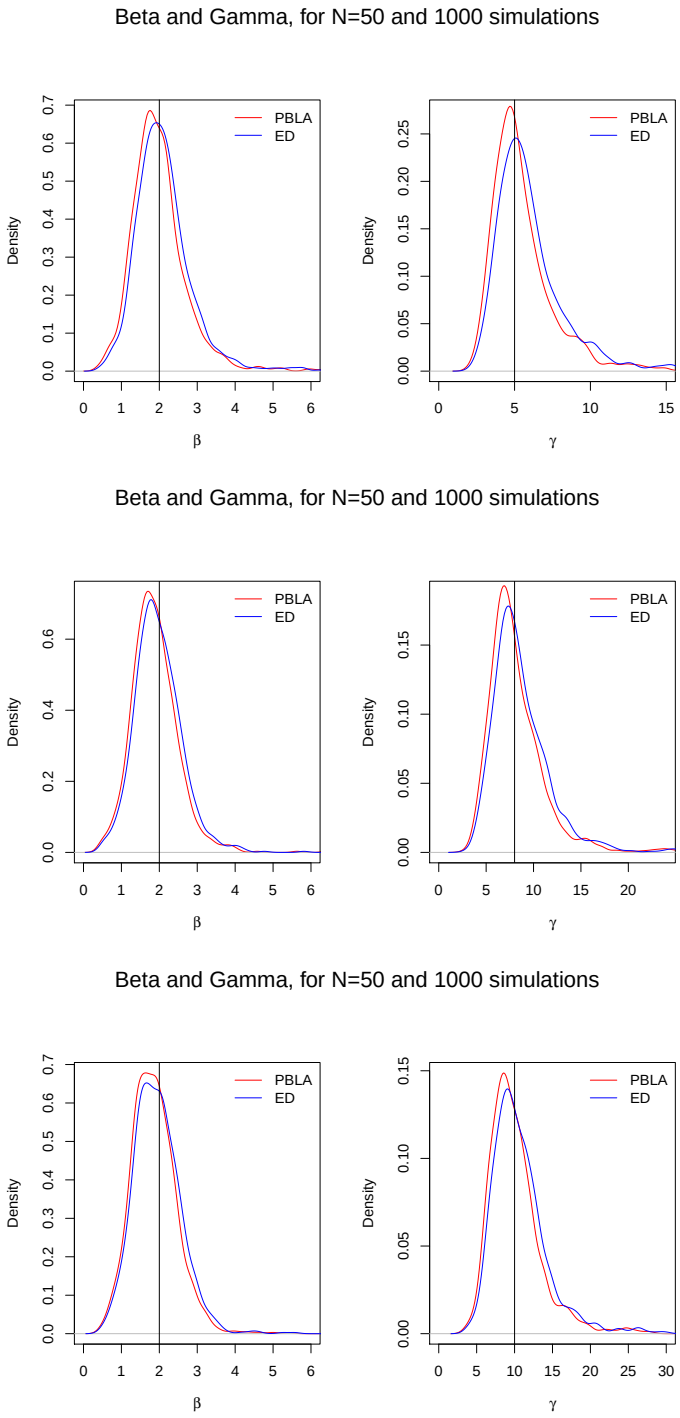


Figure 4.12: To compare the impact of varying m, these figures show densities of MLEs from both the ED and PBLA III methods with gamma infectious periods. Data are from 1000 simulations with $N = 50$ and $\beta = 2$. In the upper plots the true values are $m = \gamma = 5$, in the middle plots $m = \gamma = 8$ and in the lower plots $m = \gamma = 10$.

Varying shape parameter $m = \gamma$

Figures 4.11 and 4.12 display the effect varying shape parameter m, whilst maintaining $\gamma = m$ to keep a constant mean infectious period. We would expect both methods to improve for increased m since the infectious period becomes closer to constant. We see that even for $m = 2$, both methods approximate β and γ well, and almost exactly for $m \geq 3$. Though both methods perform very similarly for larger values of shape m, PBLA offers considerably improved estimation over the ED method for smaller m.

As with varying N, we investigate the impact of varying m, for a fixed R_0, on the bias and mean squared error of our estimates. Figures 4.13 and 4.14 show the estimated bias and mean squared error, respectively, with R_0 fixed to 1.6 and 4. We see that as we increase m (so the infectious periods tend closer to constant), in consequence the bias and mean squared error greatly reduce for both approximation methods. In agreement with what we have seen so far, it seems that moving from $m = 1$ to $m = 2$ provides the largest improvement, especially in the mean squared error, and then larger m values do not alter this so greatly. The PBLA and ED methods have very similar bias and mean squared error for larger m, though both are considerably lower with PBLA for smaller values of the shape parameter.

Varying basic reproduction number R_0

Figures 4.15 and 4.16 show the effect of varying R_0 on estimation. This is achieved by varying β whilst keeping all other parameters fixed, including fixed mean and variance of the infectious period. We see that in the first plot of Figure 4.15 where only on average 8 out of 80 individuals were infected, both methods estimate β and γ well on average, but with a fairly large variance. In the second plot of Figure 4.15 where the proportion of infectives has increased to almost half, the estimation is still good, but with much lower variance. There is no significant difference between the performance of the two methods. Then, as we saw in the exponential case, in both plots in Figure 4.16 (where the average number of infectives has increased to 78 and 80 out of 80,

respectively), the estimation of both β and γ has worsened, by both methods fairly equally.

Concluding remarks

We see a marked improvement in the performance of both approximation methods with gamma distributed infectious periods compared to exponential, even for just shape $m = 2$. In almost all of the examples we have explored, both β and γ are estimated very closely, and the PBLA and ED methods perform very similarly. As for exponential infectious periods however, both methods continue to struggle when the proportion of infectives is close to one.

Considering the error involved in estimation under both methods, we have seen that larger values of shape parameter m and population size N lead to lower bias and mean squared error. Since these methods are intended for large populations, this should be beneficial. We have seen that the PBLA method offers lower bias and mean squared error than the ED method, especially for lower shape parameter values.

In terms of the time taken to run the analysis under both methods, the computational speed is of course affected by the computational parameters chosen. For example, we must choose the number of trapezia to be used in the numerical integration for the ED method. For $k = 1000$ trapezia (any larger number having been found not to significantly increase the accuracy), $m = 8$ and $N = 100$, the PBLA method is roughly four times faster than the ED method. As N increases up to 1000, PBLA is still faster though only now by a factor of two. Indeed, for all $N \geq 1000$ tested, the PBLA method remains approximately twice as fast as the ED approach. This highlights the benefits of the PBLA method: we have avoided numerical integration and provided an approximation method which is more computationally efficient than the ED method, at no apparent cost of accuracy. We will analyse this more fully in Section 4.1.4.

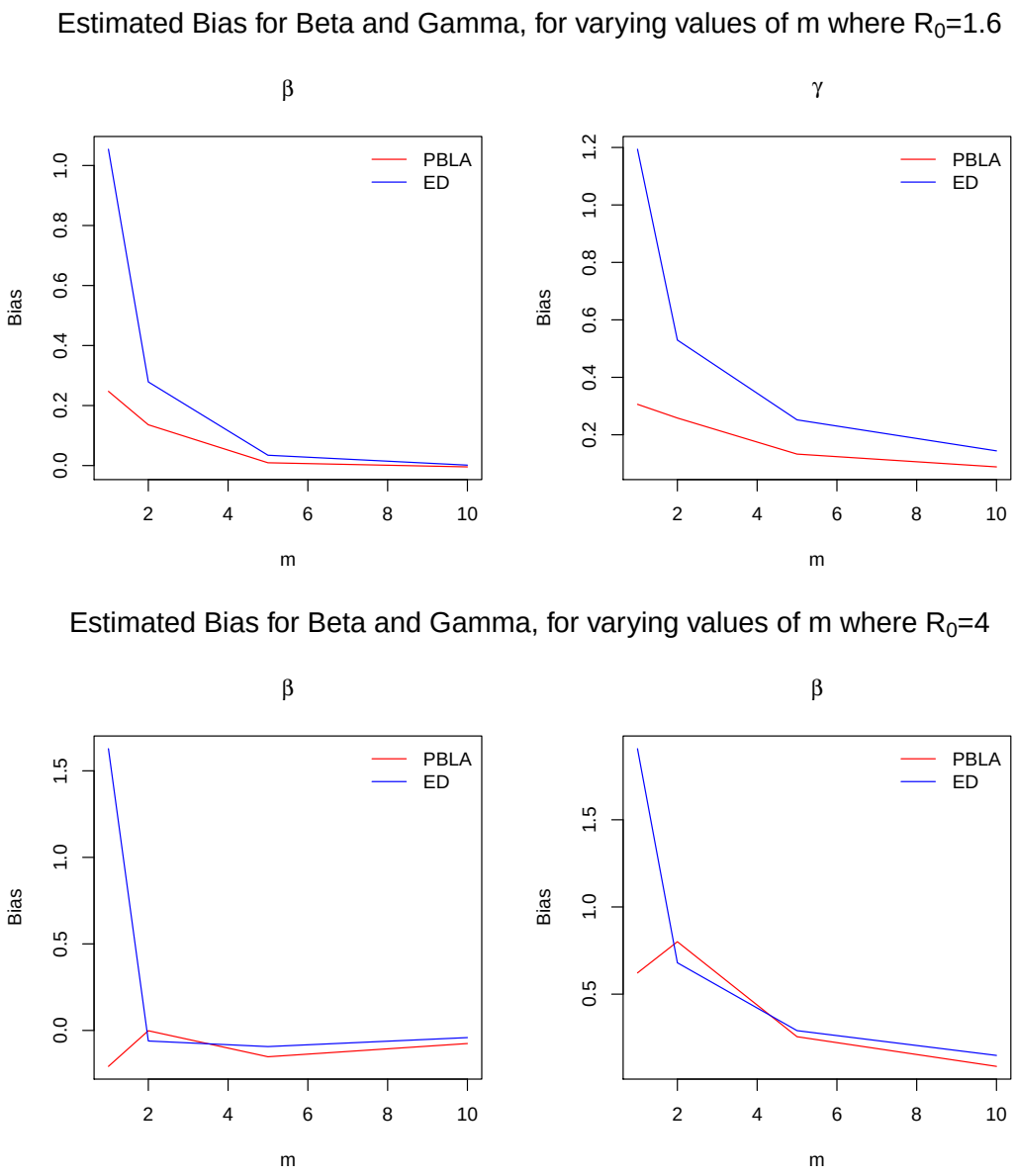


Figure 4.13: These figures show the bias in estimating parameters β and γ as $m = \gamma$ varies, for both the PBLA and ED methods. Shown are estimated values with R_0 fixed to 1.6 and 4, $N = 80$, and where in all cases 1000 outbreaks were simulated.

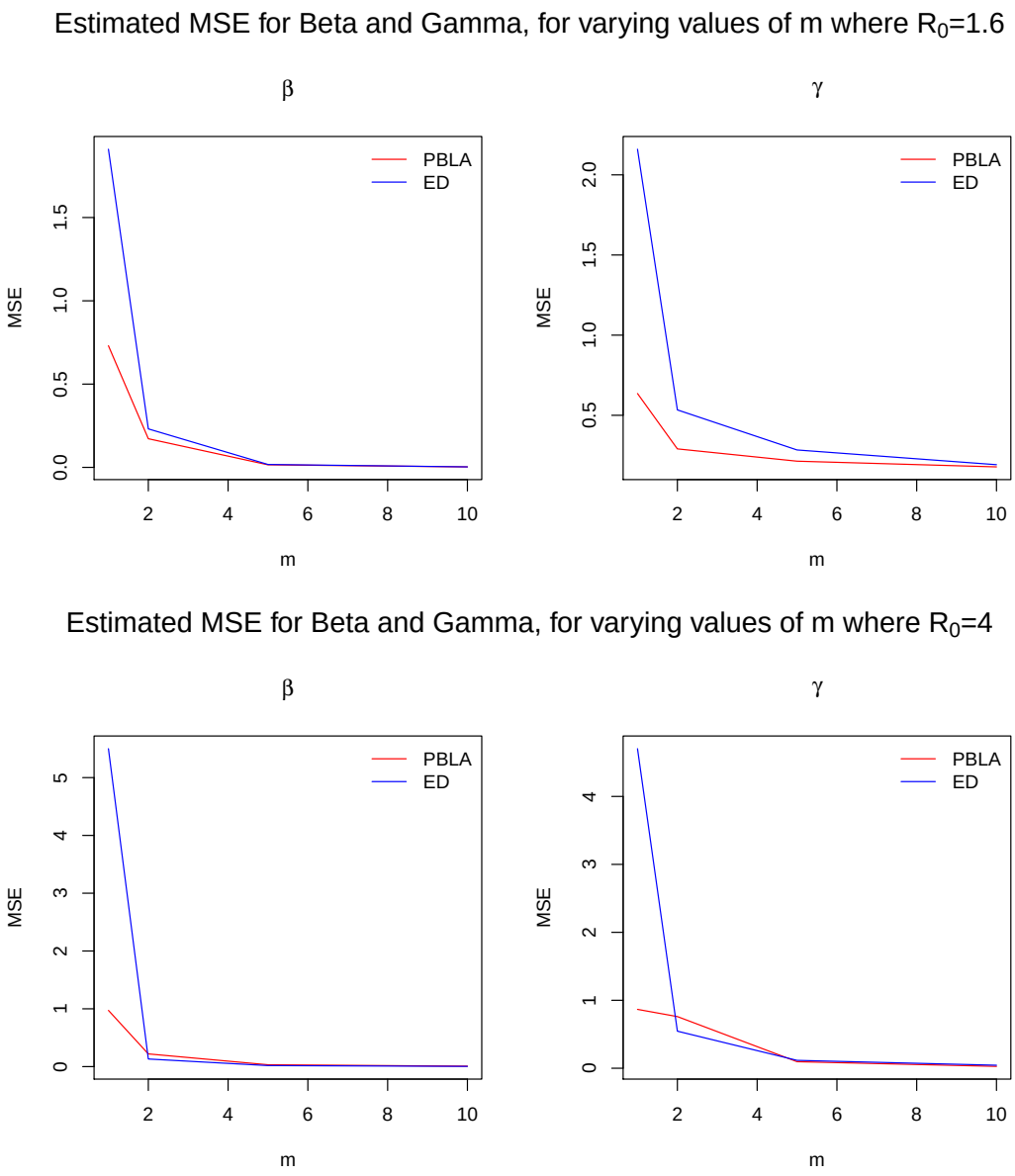


Figure 4.14: These figures show the mean squared error in estimating parameters β and γ as $m = \gamma$ varies, for both the PBLA and ED methods. Shown are estimated values with R_0 fixed to 1.6 and 4, $N = 80$, and where in all cases 1000 outbreaks were simulated.

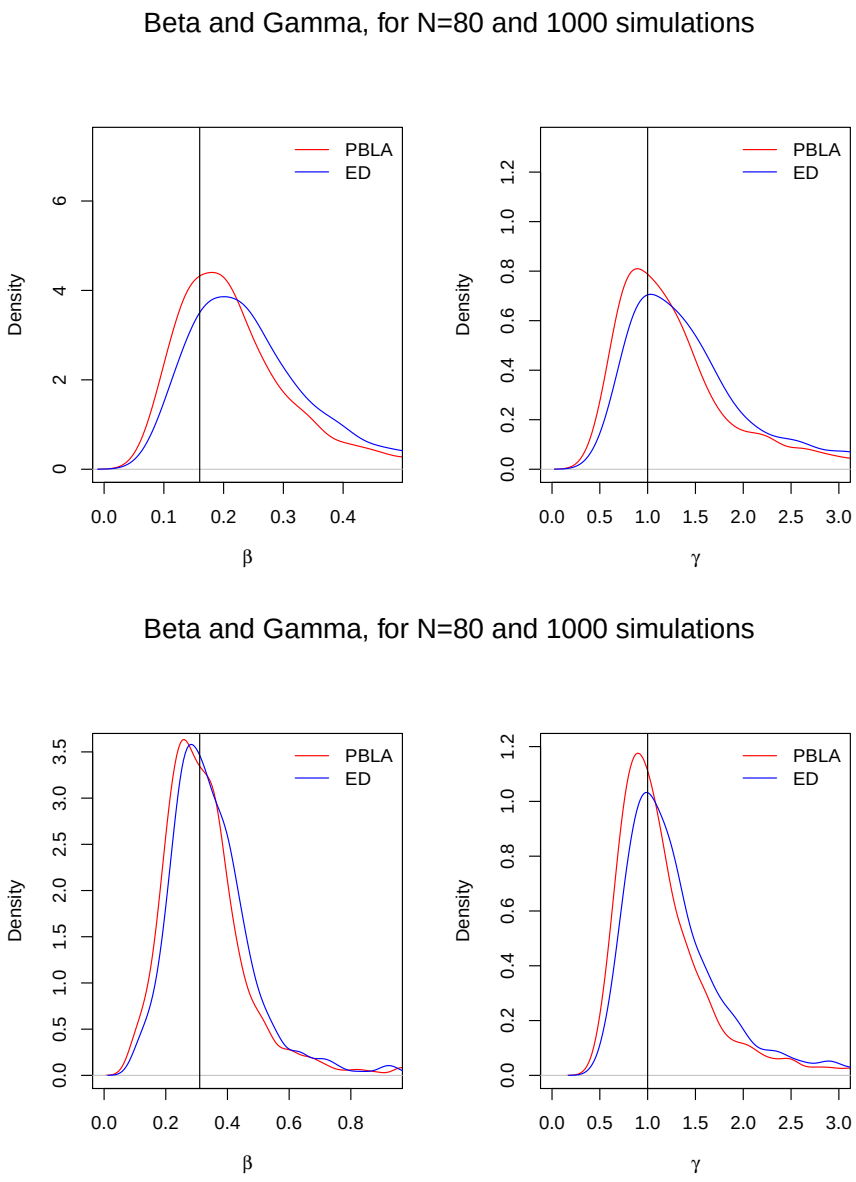


Figure 4.15: To compare the impact of varying R_0, these figures show densities of MLEs from both the ED and PBLA III methods with gamma infectious periods. Data are from 1000 simulations with $N = 80$ and shape $m = 5$. In the upper plots the true values are $\beta = 0.16$ and $\gamma = 1$, and in the lower plots $\beta = 0.31$ and $\gamma = 1$. This leads to R_0 values of 0.8 and 1.55, respectively, with the average number of infectives in these simulations being 8 and 39.

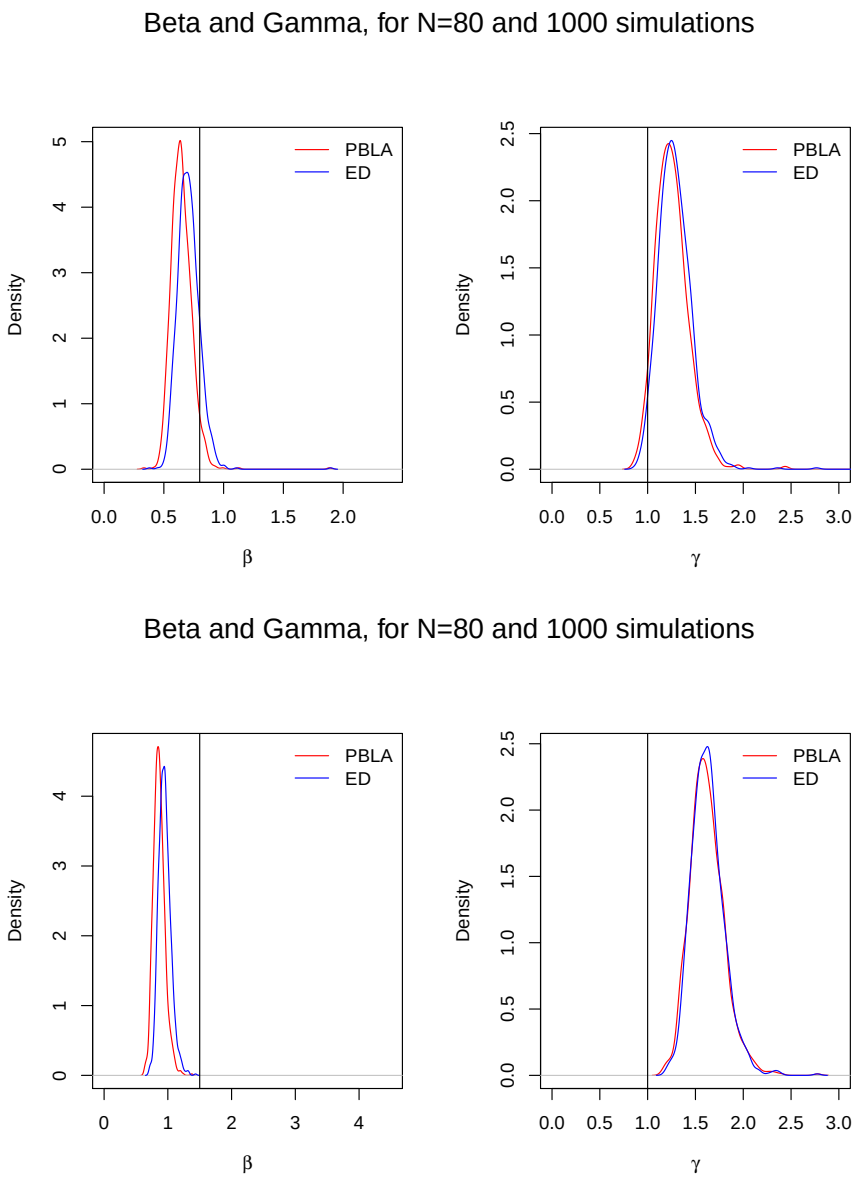


Figure 4.16: To compare the impact of varying R_0, these figures show densities of MLEs from both the ED and PBLA III methods with gamma infectious periods. Data are from 1000 simulations with $N = 80$ and shape $m = 5$. In the upper plots the true values are $\beta = 0.8$ and $\gamma = 1$, and in the lower $\beta = 1.5$ and $\gamma = 1$. This leads to R_0 values of 4 and 7.5, respectively, with the average number of infectives in these simulations being 78 and 80.

4.1.3 A Comparison of PBLA versions

The simulation studies in sections 4.1.1 and 4.1.2 have compared the ED and PBLA methods in terms of parameter estimation, and found that, in general, PBLA offers more accurate estimates. However, so far we have only focused on PBLA version III. In this section, we now perform a brief comparison of the different PBLA versions, with both exponential and gamma infectious periods as usual. Similar to the study in Section 4.1.2, we will simulate a large number of outbreaks for a selection of sets of parameter values, and perform parameter estimation using PBLA methods I through V (where applicable).

4.1.3.1 Exponential Infectious Periods

Considering exponential infectious periods first, we simulate 1000 outbreaks for three sets of parameter values with increasing population size N. In all cases, we use true values $\beta = 1.5, \gamma = 1.0$, with $N = 15$, 100 and 250. In each simulation we begin with one initial infective, and we discard and resimulate any simulations of final size one. Since in this study we are only comparing different versions of PBLA, for which maximum likelihood estimation is applicable, we will perform parameter estimation using the '*optim*' function in R to obtain MLEs.

Figure 4.17 includes densities of the estimates of β and γ using PBLA versions I through V, for these increasing values of the population size N. Most of the methods perform similarly when N is small, but for larger N the later PBLA versions offer considerable improvement over PBLA I. We see that a population size of $N = 15$ is perhaps not large enough for the central limit theorem approximation in PBLA IV to hold, since there is bimodality in the density for γ.

In terms of computational speed, PBLA IV is the fastest version to run, and seems to do so at little cost of accuracy in all but small population sizes. This particularly highlights this method as a strong alternative to DA-MCMC with

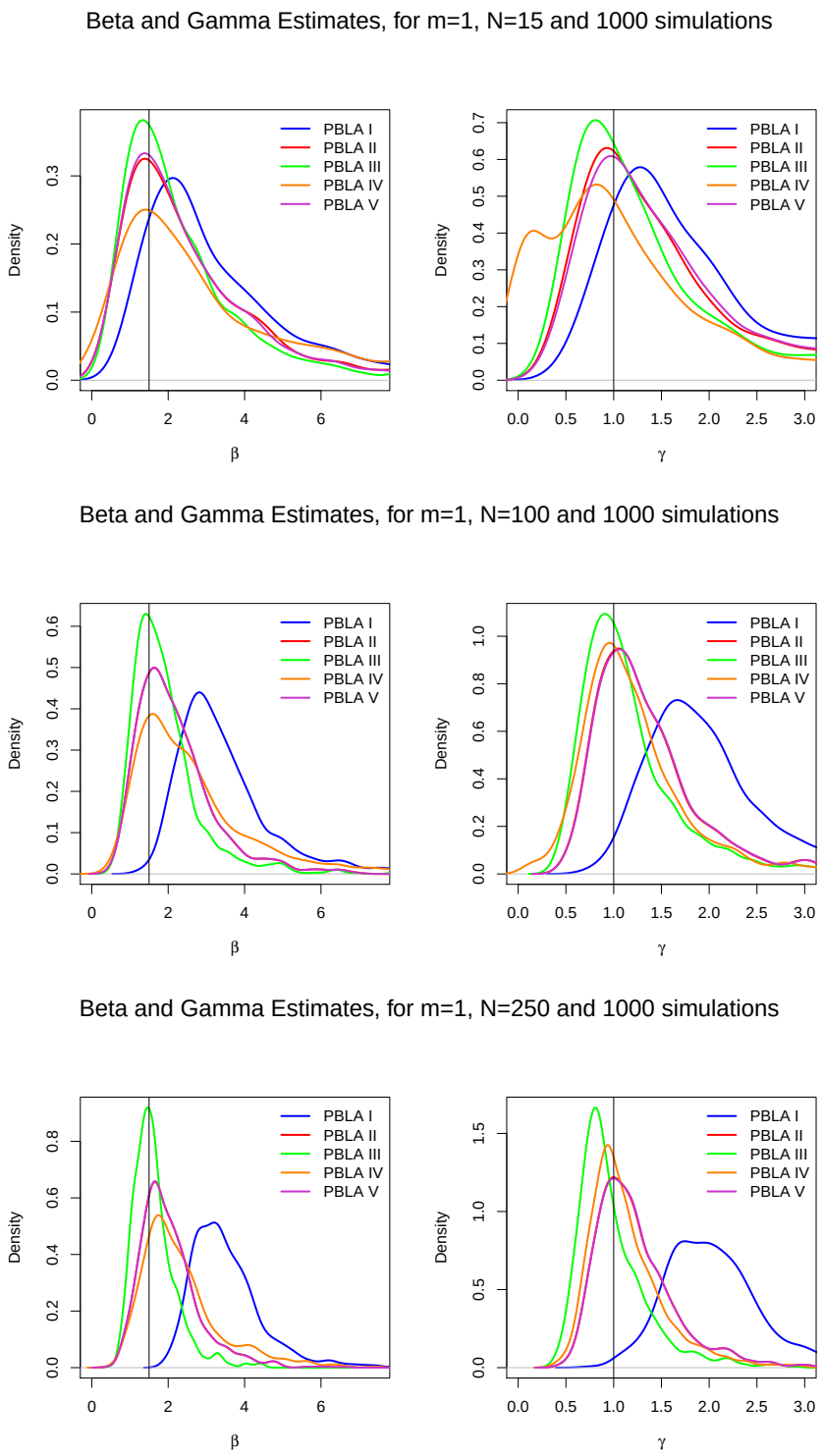


Figure 4.17: To compare the different PBLA versions, these figures show densities of MLEs for 1000 simulations over a range of population sizes with exponentially distributed infectious periods, where $\beta = 1.5$ and $\gamma = 1$. Note: PBLA II curve is almost exactly behind PBLA V.

the true likelihood when computation times are slow, so long as the population size is large enough for the CLT to hold. PBLA II and III, although slower than IV, are more widely applicable since they have no requirement for a large population size.

4.1.3.2 Gamma Infectious Periods

Moving on to gamma distributed infectious periods, we again simulate 1000 outbreaks for a series of different parameter values in order to perform parameter estimation using each of the PBLA versions. Simulation and estimation is performed as in the exponential case. We do not include PBLA IV in this section, since it is not suitable for gamma infectious periods.

For all sets of parameter values tested we find that PBLA II, III and V perform very similarly, and all offer better estimation than PBLA I. Figures 4.18, 4.19 and 4.20 display densities from 1000 simulations over a range of values for β, γ, m and N, under all of which we may make these same conclusions. Note that in these figures, PBLA versions II, III and V are so similar that the curves are almost indistinguishable.

As we have seen in the previous simulation studies, all of the PBLA versions somewhat struggle when the proportion of infectives is very large, as in the γ estimation in Figure 4.20 where $R_0 = 4$. This strengthens the proposition in Section 4.1.2.1 that the contribution from non-infective individuals (contained in ψ term) is key to the accuracy of the PBLA method, especially since most of the approximation is in the infective-to-infective pressure term.

PBLA V provides very similar estimation to PBLA II and III in all the cases we have examined. We might have expected it to perform better since we consider the infectious pressure from any individual j to individual k together with the pressure from k to j, rather than looking at these independently. As discussed in Section 3.4.9.3, it seems that this has little impact however. Regardless, PBLA V is considerably faster than PBLA II or III, so similar perfor-

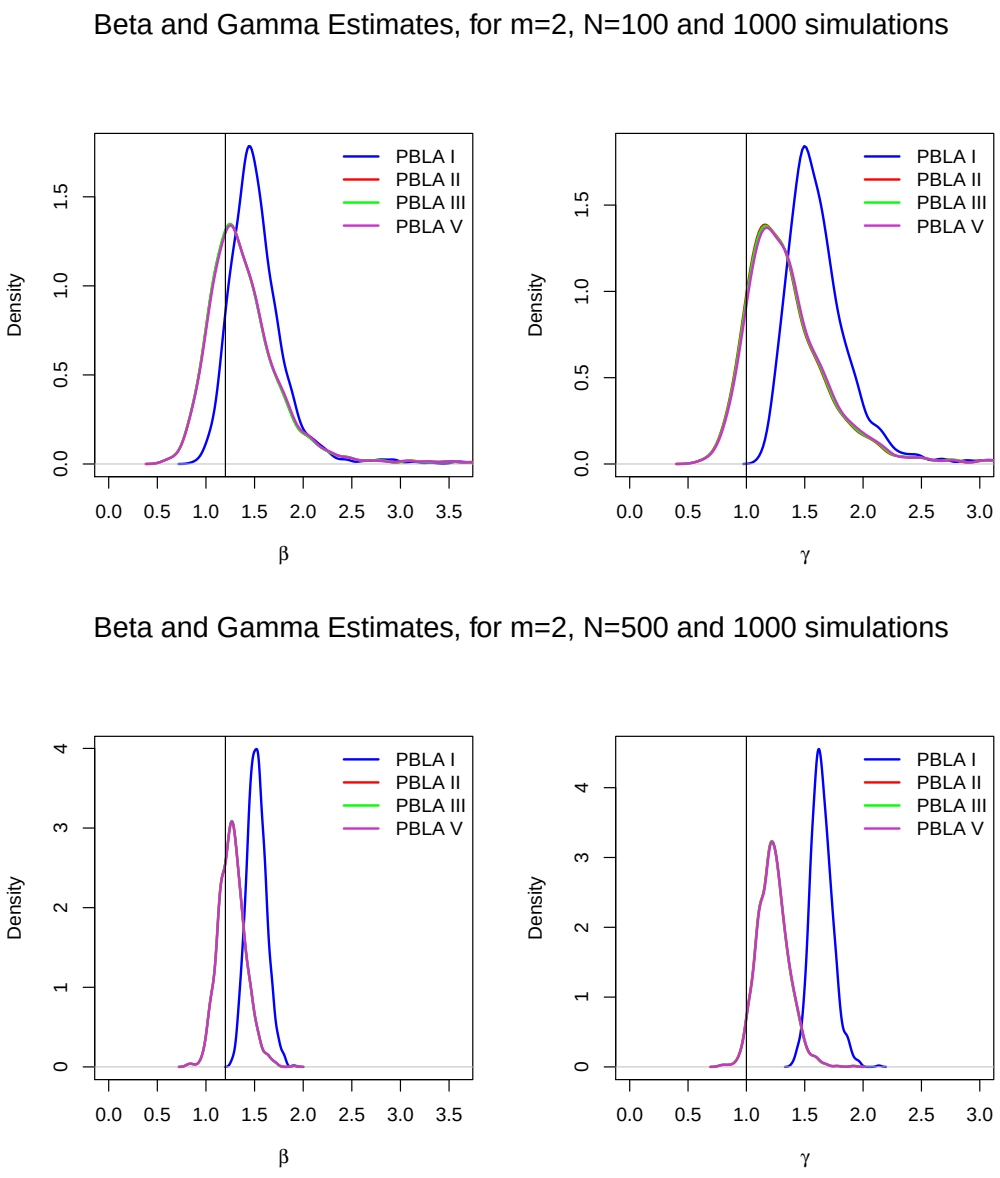


Figure 4.18: To compare the different PBLA versions, these figures show densities of MLEs for 1000 simulations with gamma distributed infectious periods. Respectively, true $\beta = 1.2$, $\gamma = 1$ and $N = 100$, and $\beta = 1.2$, $\gamma = 1$, $N = 500$. In both plots, $m = 2$. Note: PBLA II, III and V are almost exactly aligned.

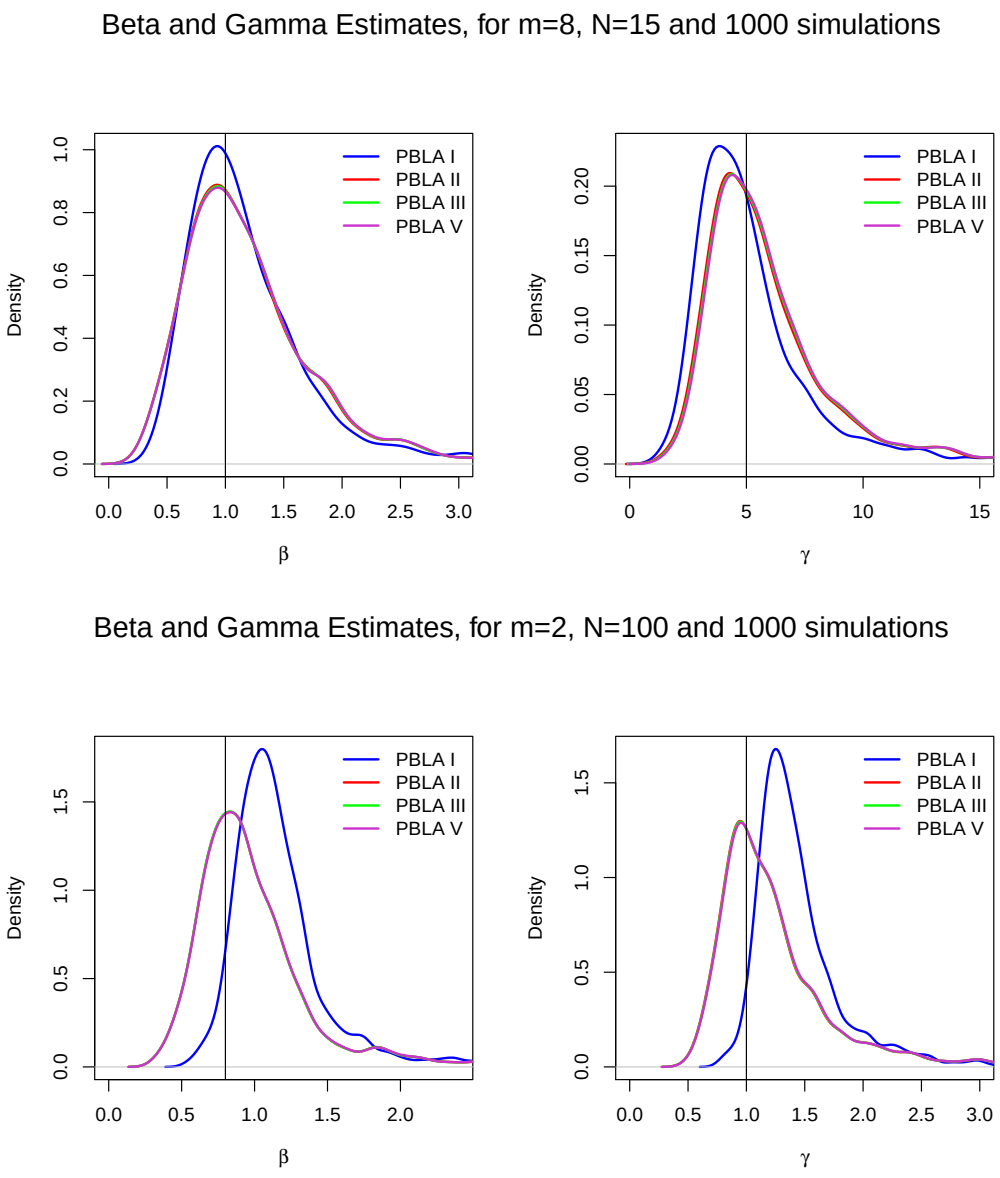


Figure 4.19: To compare the different PBLA versions, these figures show densities of MLEs for 1000 simulations with gamma distributed infectious periods. Respectively, true $\beta = 1$, $\gamma = 5$, $m = 8$ and $N = 15$, and $\beta = 0.8$, $\gamma = 1$, $N = 100$ and $m = 2$. Note: PBLA II-V are almost exactly aligned.

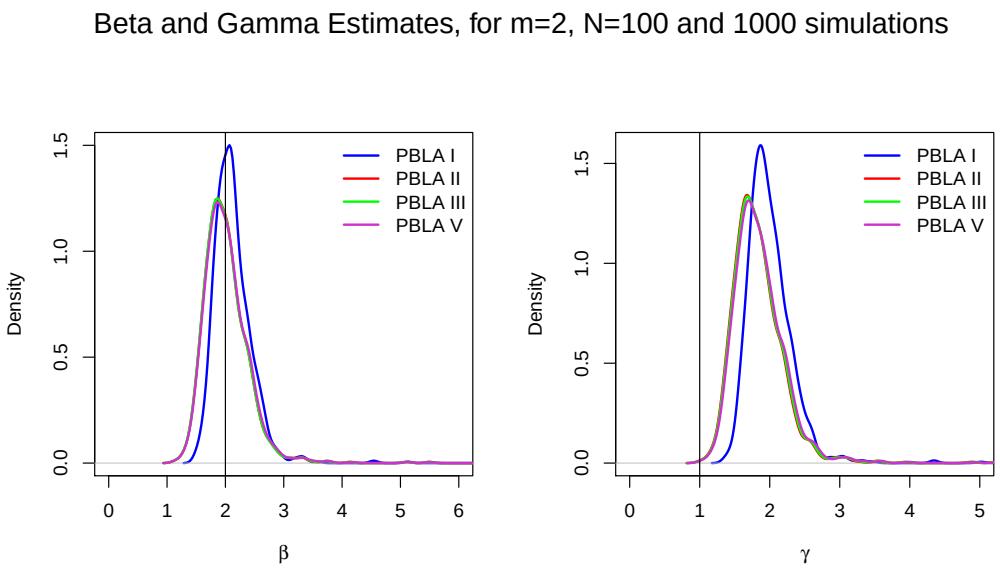


Figure 4.20: To compare the different PBLA versions, this figure shows densities of MLEs for 1000 simulations with gamma distributed infectious periods, where $\beta = 2$, $\gamma = 1$, $m = 2$ and $N = 100$. Note: PBLA II-V are almost exactly aligned.

mance means it could still be beneficial to implement this version.

4.1.3.3 Conclusion

In summary, we have found that, for both exponential and gamma distributed infectious periods, any of PBLA methods II through V offer similar performance, and improvement over PBLA I. For exponential infectious periods, PBLA IV may be considered a good choice since it offers comparable estimation at the lowest computational cost, so long as the population size is not too small. Otherwise, for gamma infectious periods, PBLA III and V provide the fastest computation time and closest estimation. Under the range of parameter values tested, each of methods II through V results in estimates close to the true values, though we continue to see these struggle when the proportion of infectives is very large. Future work could include the development of further PBLA versions which work better under these circumstances.

4.1.4 Computation time

So far we have compared the likelihood approximation methods to DA-MCMC in terms of the accuracy of parameter estimation. Although this is of course important, so is the computation time involved. If approximation methods do not offer an improvement over standard DA-MCMC with the true likelihood, then there will of course be less motivation to use them. In this section we perform a comparison of the computation time required for the PBLA method with MCMC, as compared to DA-MCMC. After our findings in sections 4.1.2 and 4.1.3, that PBLA generally outperforms the ED method and that PBLA III offers a good balance of computation time and wide applicability compared to the other PBLA versions, we will restrict our attention here to PBLA III.

4.1.4.1 Method

For this study, we will compare the time taken to obtain MCMC samples using both PBLA III with MCMC and standard DA-MCMC. We simulate a number of outbreaks with increasing population sizes N, using gamma distributed infectious periods with shape parameter $m = 1$ (corresponding to exponential infectious periods), $m = 2$ and $m = 5$. In all cases, we set infection rate parameter $\beta = 1.5$ and removal rate $\gamma = m$, to result in an R_0 value of 1.5. We simulate a single outbreak for each of $N = 20$, 50, 100, 200, 500, 1000 and 2000, for each set of (β, γ, m) values. We assume one initial infective for each simulation, and re-simulate any outbreaks of final size one.

We then perform parameter estimation for β and γ using PBLA with MCMC and standard DA-MCMC, assuming shape parameter m is fixed. We record the time taken to obtain a fixed number of samples n_s from the MCMC algorithms, where all parameters are updated at each iteration of the algorithms. For PBLA this is just β and γ, whereas for DA-MCMC we also update all infection times at each iteration. For PBLA we perform random walk updates, with low rate (10^{-4}) exponential priors. All analysis is coded in C and performed on the

same machine.

In reality, both methods would require a burn-in period for the Markov chain to move into equilibrium. The choice of initial values of the parameters therefore impacts the samples obtained. However, since DA-MCMC requires imputation of initial infection times but PBLA only requires initial β and γ values, there is no fair way to compare this burn-in between the two methods. We hence start all algorithms suitably tuned, with the chains already in equilibrium so that the comparison is fair regardless of the initial values selected.

We could perform a simple comparison of the time taken to obtain a fixed number of MCMC samples, but as discussed in Section 3.1 a motivating factor in using likelihood approximation methods is the high dependence between the infection times and γ, when using DA-MCMC. This may cause slow mixing of the Markov chain. We therefore seek some measure of computation time which includes the computational burden involved in obtaining independent samples.

Effective sample size (ESS) is one such measure which may be used, and is popular as an MCMC diagnostic (see e.g. Brooks et al. (2011) and Kass et al. (1998)). If we have obtained a number n_s of dependent samples from our MCMC, the ESS estimates the number of independent samples that this corresponds to (so in a completely independent chain, the ESS is equal to n_s). The ESS is defined as

$$\text{ESS} = \frac{n_s}{1 + 2\sum_{k=1}^{\infty} \rho(k)},$$

where $\rho(k)$ is the autocorrelation at lag k. In practice, we truncate the infinite sum to the first lag k where $\rho(k+1) < 0.05$.

Therefore, as an overall measure of the speed at which independent samples may be obtained, we compare the Effective sample size per second (ESS s^{-1}) of the PBLA MCMC and DA-MCMC algorithms, for both β and γ. This is defined simply as the ESS divided by the time taken to obtain those samples.

4.1.4.2 Results

Table 4.5 contains the results of this study, for $m = 1$, 2 and 5. For reproducibility, we also include the number of samples obtained for each N, and the values used for the tuning parameters σ_β and σ_γ, which are the variances of the Gaussian proposals for β and γ. The number of samples n_s decreases as N increases due to the greater computational burden of large populations, but since all chains are well-mixed when samples begin being collected this should not be impactful. Highlighted in the table are the lowest N values tested for which PBLA offered a greater ESS s^{-1} than DA-MCMC. Plots of the ESS s^{-1} against N are provided in Figures 4.21, 4.22 and 4.23, for $m = 1$, 2 and 5, respectively. These are plotted on a log scale for ease of exposition.

As we can see from Table 4.5 as well as Figures 4.21, 4.22 and 4.23, as the population size N increases, eventually PBLA will result in a greater ESS per second compared to DA-MCMC. In the case of exponential infectious periods ($m = 1$), PBLA is faster the DA-MCMC even for $N = 50$. As m increases, only after larger population sizes does it become computationally preferable. This is since the structure of the PBLA III likelihood requires multiple loops over m, and so the larger this is the longer the likelihood takes to compute, comparatively. The figures suggest that as the population size continues to increase, so does the computational advantage of PBLA compared to DA-MCMC.

4.1.4.3 Discussion

As in Section 4.1.1, this computation time study uses only one simulation per set of parameter values, so it may be that some simulations are not entirely representative of the average. Although this is important to note, it is reasonable to suppose that the results give a good overview of the difference in computation time between DA-MCMC and PBLA MCMC. Interestingly, we have seen overall in this chapter so far that for larger values of m, the PBLA method offers increased performance in terms of estimation, but does require larger

Table 4.5: Effective sample size per second obtained from DA-MCMC and PBLA MCMC, for a range of population sizes N. Values σ_β and σ_γ are the variances of the Gaussian proposals used for β and γ.

Panel A: **Shape parameter m=1**

			DA-MCMC	PBLA MCMC		
N	n	n_s	ESS s^{-1} for (β, γ)	σ_β	σ_γ	ESS s^{-1} for (β, γ)
20	16	10000	(783.56, 680.59)	1.0	0.6	(468.18, 507.41,)
50	42	3000	(66.35, 51.57)	1.0	0.6	(66.62, 75.50)
100	83	3000	(4.74, 4.46)	0.8	0.5	(9.46, 9.97,)
200	130	3000	(1.07, 1.03)	0.4	0.1	(3.99, 3.71,)
500	240	3000	(0.109, 0.108)	0.2	0.2	(1.00, 1.00)
1000	507	1000	(0.0167, 0.0169)	0.15	0.1	(0.219, 0.221)
2000	1077	1000	(0.00139, 0.00139)	0.15	0.1	(0.0432, 0.0433)

Panel B: **Shape parameter m=2**

			DA-MCMC	PBLA MCMC		
N	n	n_s	ESS s^{-1} for (β, γ)	σ_β	σ_γ	ESS s^{-1} for (β, γ)
20	11	10000	(1210.83, 861.07)	1.0	1.0	(144.38, 156.07)
50	34	3000	(75.67, 58.25)	1.0	1.0	(5.88, 5.87)
100	75	3000	(7.11, 5.92)	1.0	1.0	(0.89, 0.87)
200	106	3000	(1.89, 1.70)	0.4	0.4	(0.86, 0.82)
500	193	1000	(0.26, 0.24)	0.1	0.1	(0.15, 0.14)
1000	587	1000	(0.0089, 0.0086)	0.1	0.1	(0.015, 0.015)
2000	1134	1000	(0.0010, 0.0010)	0.1	0.1	(0.0043, 0.0043)

Panel C: **Shape parameter m=5**

			DA-MCMC	PBLA MCMC		
N	n	n_s	ESS s^{-1} for (β, γ)	σ_β	σ_γ	ESS s^{-1} for (β, γ)
20	16	10000	(257.91, 142.46)	1.0	1.0	(15.05, 12.77)
50	36	3000	(33.84, 24.34)	2.0	5.0	(1.77, 1.77)
100	70	3000	(5.87, 4.43)	0.5	1.0	(0.59, 0.57)
200	122	1000	(1.52, 1.17)	0.4	1.2	(0.13, 0.13)
500	194	1000	(0.23, 0.20)	0.1	0.5	(0.05, 0.05)
1000	565	1000	(0.00873, 0.00815)	0.1	0.1	(0.00518, 0.00485)
2000	1169	1000	(0.0098, 0.0093)	0.1	0.1	(0.0013, 0.0012)

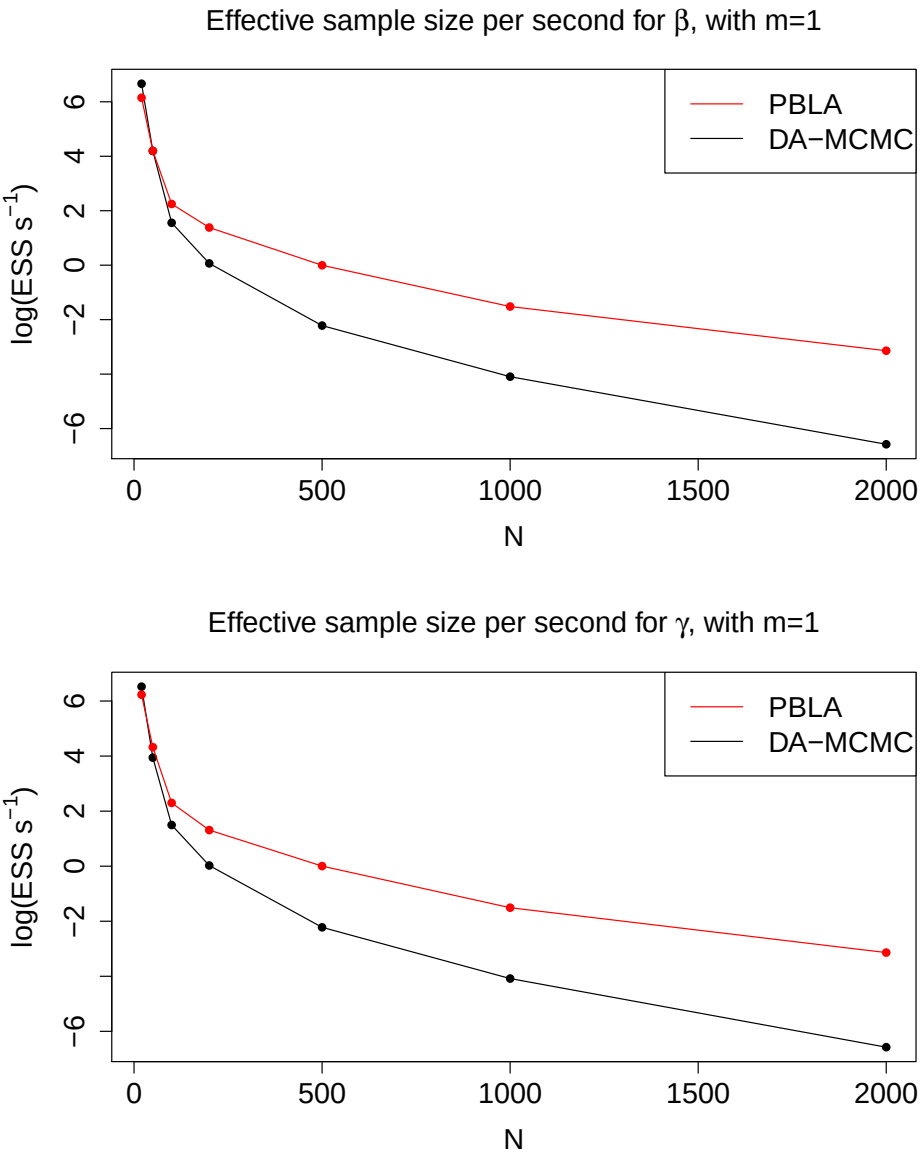


Figure 4.21: Plots of the log effective sample size per second of β and γ, obtained from PBLA MCMC and DA-MCMC for increasing values of population size N and fixed shape parameter $m = 1$. Note: the lines are for visualisation purposes, only N values at the marked points were tested.

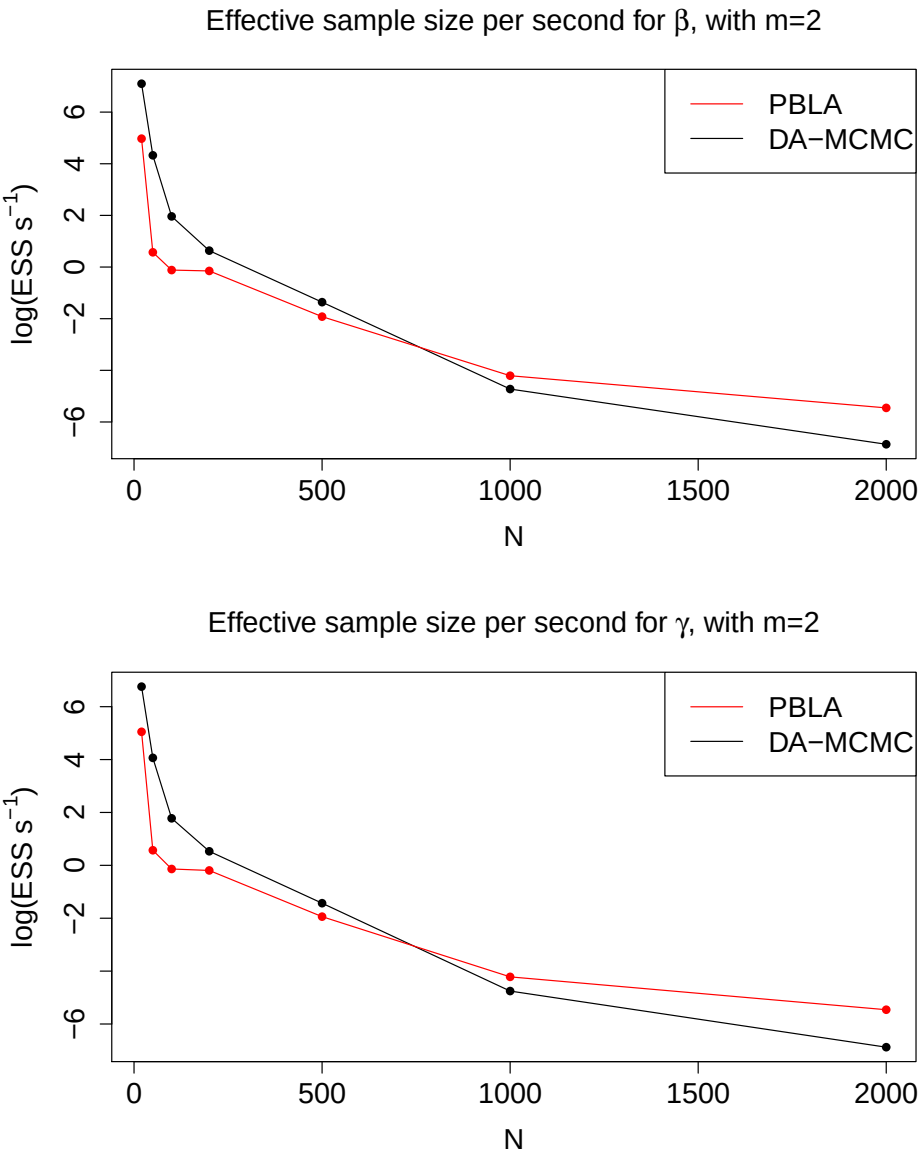


Figure 4.22: Plots of the log effective sample size per second of β and γ, obtained from PBLA MCMC and DA-MCMC for increasing values of population size N and fixed shape parameter $m = 2$. Note: the lines are for visualisation purposes, only N values at the marked points were tested.

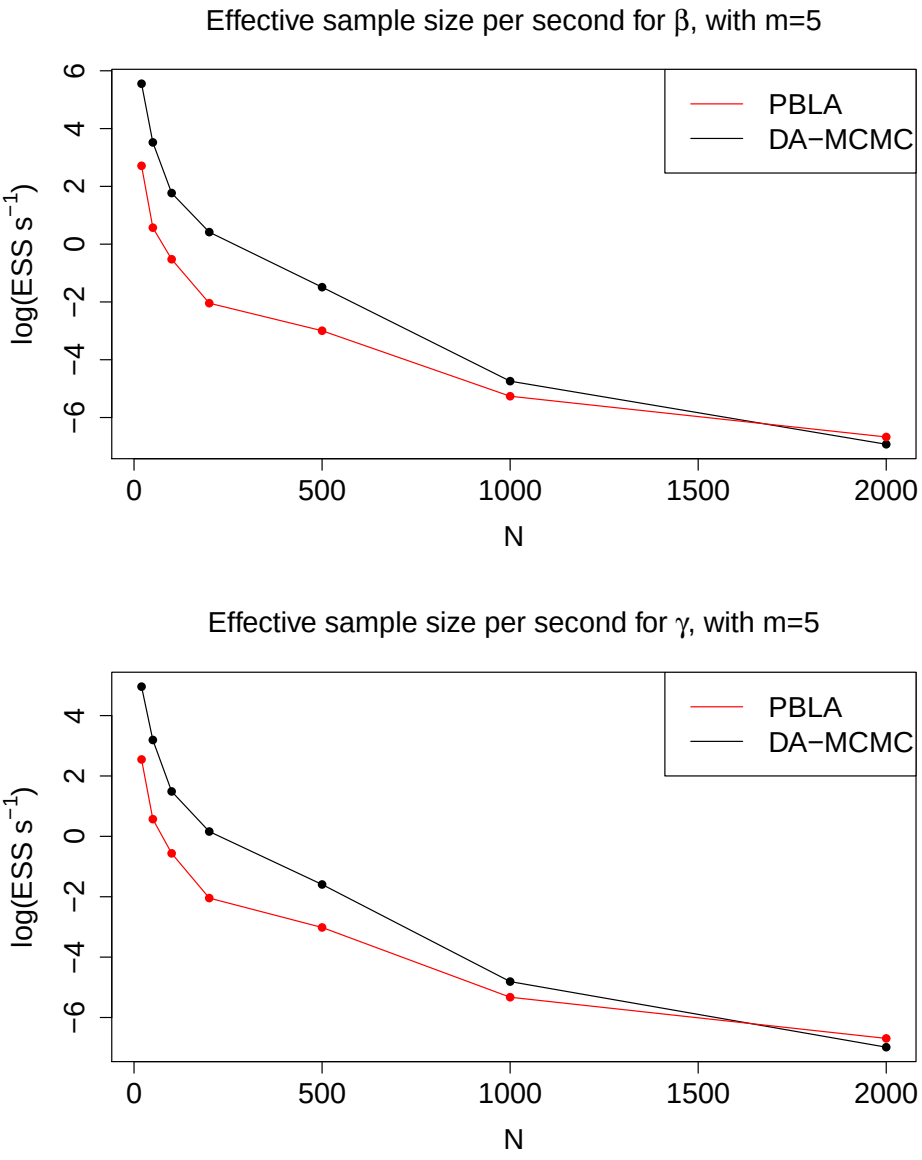


Figure 4.23: Plots of the log effective sample size per second of β and γ, obtained from PBLA MCMC and DA-MCMC for increasing values of population size N and fixed shape parameter $m = 5$. Note: the lines are for visualisation purposes, only N values at the marked points were tested.

population sizes to be faster than DA-MCMC. However, since it is large outbreaks which most motivate the use of likelihood approximations, this should not be a problem in practice. We also recall that this comparison has only included the speed of obtaining samples from an already well-mixed chain. This ignores the time taken to reach equilibrium, for which PBLA should be faster since it does not require updates of the infection times, though as discussed this is difficult to compare.

There is potential to increase the speed of the PBLA MCMC algorithm even further with future work. As well as being naturally parallelisable since the likelihood involves computation of the contributions from different pairs of individuals independently, the PBLA likelihood may also be used with maximum likelihood estimation, removing the need for MCMC altogether. For more details, we will explore an application using maximum likelihood estimation in Section 4.3. Even without these potential extensions, we have shown that for the larger population sizes for which PBLA is intended, it offers an improvement in computational speed compared to DA-MCMC. Combined with the comparable estimation which we have seen in sections 4.1.1 and 4.1.2, there is evidence that PBLA is a useful tool for analysis.

4.2 Applications: Tristan Da Cunha respiratory disease data

With the simulation studies complete, we now explore the application of likelihood approximations to real data. In the three analyses which follow we will use the PBLA III method for approximation, since it offers accurate estimation at relatively low computational cost, as well as being applicable to the widest range of models explored.

We begin with a data set from the remote South Atlantic island of Tristan Da Cunha. Known as the most remote inhabited island in the world, and with

almost zero immigration, Tristan Da Cunha is particularly of interest since the population is actually approximately closed. Its remoteness means that epidemics are almost always introduced from an external source such as the arrival of ships (Shibli et al., 1971), with the population too small to reasonably allow for reinfection. From 1963 to 1968, medical officers continually studied respiratory infections on the island in detail. These data have been analysed several times, including Hammond and Tyrrell (1971), Shibli et al. (1971), Becker and Hopper (1983), Hayakawa et al. (2003) and Xu et al. (2016). There were seven major outbreaks during the study period, but we focus on a particular outbreak from 1967, modelling the population with three types of individuals as categorised by age, and comparing our results to Hayakawa et al. (2003).

4.2.1 Data and Model

4.2.1.1 Data

Between October and November of 1967, 40 of the island's 255 inhabitants were infected with a respiratory disease. We consider the population segregated into three groups: (1) infants, aged 0-4, (2) children, aged 5-14, and (3) adults, aged 15 and above (at the time, children on Tristan Da Cunha attended school between the ages of 5 and 15). One case was unidentified, so we reduce the population size to 254.

To introduce the data, the population size is given by $N = 254$, of which $n = 40$ become infected. The initial population size of each group is $N_1 = 25$, $N_2 = 36$ and $N_3 = 193$, with the total number of cases for each group given by $n_1 = 9$, $n_2 = 6$ and $n_3 = 25$. The data contain the total number of cases identified each day, which we take to be equal to removal times. Table 4.6 displays this as daily data for simplicity, though for the PBLA method we convert it into an individual-based format. To do this (referring to Table 4.6), we set the removal time of the first infective as 1.0, the removal time of the second infective as 8.0

Table 4.6: Removal data from the 1967 respiratory disease outbreak on Tristan Da Cunha (from Hayakawa et al. (2003), and originally Becker and Hopper (1983)). Age groups defined as: infants aged 0-4, children aged 5-14, and adults aged 15 and above.

	Number of removals		
Day	Infants	Children	Adults
1	0	0	1
8	0	0	1
10	0	1	1
11	3	1	0
12	1	1	2
13	3	0	3
15	1	1	1
16	0	1	4
17	0	0	1
18	1	1	1
19	0	0	3
20	0	0	2
21	0	0	1
22	0	0	2
29	0	0	1
30	0	0	1
Total	9	6	25

and so on. To avoid the problems of equal removal times discussed in Section 3.4.10, we jitter all equal removals by increments of 0.1, which was found to have minimal impact on the likelihood values.

4.2.1.2 Model

We may next define the model to be used, which is the same as that in Hayakawa et al. (2003) for comparability. We assume that Tristan Da Cunha forms a closed population of N individuals labelled $1, 2, \ldots, N$, of whom $1, 2, \ldots, n$ become infected ($n \leq N$). All individuals are categorised by age into one of groups 1,

2 and 3, as defined above in section 4.2.1.1. We also assume external infection of the initial case. We will use a simple SIR stochastic transmission model for the disease, defined along much the same lines as previous models we have explored in this thesis. At any given time, each individual in the population will be either in state S (susceptible), state I (infective), or state R (removed). For all infectives $j = 1, 2, \ldots, n$, let i_j and r_j denote, respectively, their time of infection and removal. Any susceptible individual may become infected, as will be described below, and enter the I stage at which point they are able to infect others. They will later become removed (corresponding to detected) and enter the R stage. At this point, they are unable to cause infections and, since the length of the outbreak was too short to reasonably allow reinfection, are considered removed from the population.

The outbreak begins with the infection of the initial case κ at time i_κ, and ends when there are no infectives remaining in the population. The infection times $\mathbf{i} = \{i_j : j = 1, 2, \ldots, \kappa - 1, \kappa + 1, \ldots, n\}$ are unknown, and the data consist of removal times $\mathbf{r} = \{r_j : j = 1, 2, \ldots, n,$ where $r_1 < r_2 < \cdots < r_n\}$. During their infectious period, any infective will have a contacts with an individual from group i at times given by points of a Poisson process of rate β_i, where all Poisson processes are assumed mutually independent. That is to say, all individuals are equally infectious, but individuals in group i receive infectious pressure β_i to reflect their differing susceptibility to the disease. Any contact between an infective and a susceptible is assumed to result in immediate infection. The lengths of the infectious periods of different individuals are assumed mutually independent and exponentially distributed with parameter γ.

Hayakawa et al. (2003) used data augmentation and MCMC to estimate infection rates, infectious period parameters and reproduction numbers for this outbreak. We fit the same model but use the PBLA method with MCMC in order to compare our results.

Table 4.7: Gamma prior distributions for the infection parameters in the Tristan Da Cunha outbreak, as in Hayakawa et al. (2003).

	Mean	Variance
β_1	0.001	100
β_2	0.001	100
β_3	0.001	100
γ	0.1	100

4.2.2 Results

As in the simulation studies, we wish to compare PBLA to standard DA-MCMC by considering the difference in estimation of the model parameters. Under the Tristan Da Cunha model, these parameters are the infection rates for each age group β_1, β_2 and β_3, and infection rate γ. We will also consider estimation of the basic reproduction number $R_0 = \frac{\beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3}{\gamma}$. We perform MCMC with the PBLA III likelihood, and compare this to the DA-MCMC results from Hayakawa et al. (2003). The PBLA MCMC was coded in C and employs individual Gaussian random walk updates, with 100,000 samples taken after a burn-in of length 10,000. To maintain comparability, we used the same gamma distributed prior distributions as Hayakawa et al. (2003), with parameters given in Table 4.7.

The results of the analysis are shown in Figure 4.24. Plotted are histograms of MCMC samples from the PBLA likelihood for the model parameters as well as R_0. These samples are compared with the dotted lines in the figures, which represent the mean posterior estimates from Hayakawa et al. (2003). Table 4.8 contains the mean parameter estimates for both methods for comparison.

We see that the PBLA method provides estimates close to those of Hayakawa et al. (2003), despite not requiring the data augmentation of their analysis. This provides evidence that the PBLA method is an effective approximation, even for a multi-type epidemic model.

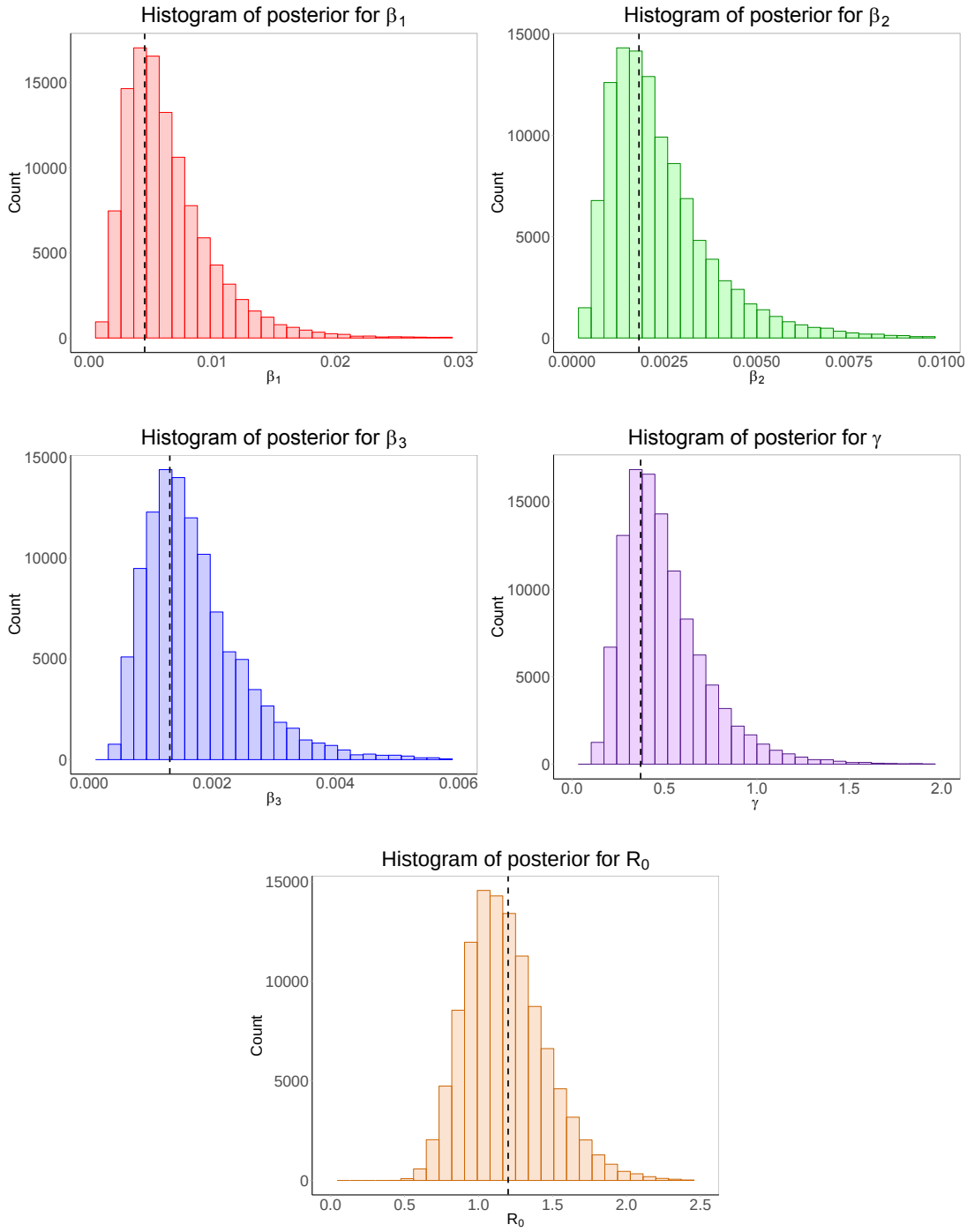


Figure 4.24: Histograms of parameter estimates for the Tristan Da Cunha data using the PBLA III approximation method. Dotted lines represent the mean estimates using DA-MCMC from Hayakawa et al. (2003).

Table 4.8: Mean parameter estimates for the the Tristan Da Cunha data using PBLA with MCMC and DA-MCMC (from Hayakawa et al. (2003)).

	DA-MCMC Mean	PBLA mean
β_1	0.00451	0.00641
β_2	0.00181	0.00239
β_3	0.00131	0.00171
γ	0.371	0.499
R_0	1.2	1.2

4.3 Applications: West African Ebola virus data

In our next analysis, we consider a data set from the 2014 Ebola virus epidemic in West Africa. The most widespread outbreak of the disease to date, the epidemic saw a case fatality of above 70% (of laboratory confirmed cases, see reference for Centers for Disease Control and Prevention (Accessed 2018-03-11)). This was mainly centred on the nations Guinea, Sierra Leone and Liberia, though minor outbreaks and isolated cases were also seen elsewhere across West Africa as well as the USA, UK, Spain and Italy. The outbreak was identified in Guinea in December of 2013, spreading to Liberia and then Sierra Leone over the next few months. As the number of infected cases began to rise, the Ebola outbreak gained much attention as the introduction of public health interventions failed to stop the increase in both cases and deaths.

There were 28,652 officially recorded cases in the outbreak (from December 2013 to April 2016), though we will focus on a smaller data set across only Guinea, Sierra Leone and Liberia, recorded from March 2014 to February 2016. This data set was obtained from the Centers for Disease Control and Prevention (Accessed 2018-03-11), though was originally collected by the World Health Organisation. We wish to compare analysis using the PBLA method to that from Althaus (2014), who used an ODE model with the Ebola data, to estimate transmission parameters for the disease. We wish to follow Althaus'

model as closely as possible to compare the results of their analysis with those using PBLA. Althaus described SEIR transmission dynamics with a set of ordinary differential equations; these were solved numerically and, assuming the cumulative numbers of cases and deaths to be Poisson distributed, maximum likelihood estimates for the parameters of interest were obtained. As we will see, this model is very different to those we have considered so far for PBLA, and so adjustments will need to be made to approximate it closely. It will be interesting to see how PBLA performs under these different conditions.

An important note is that we will use Althaus' model for comparison, but with a different data set than in the original paper. Althaus' data set was also obtained from the World Health Organisation, but since the paper was published in 2014 it does not contain data on the complete outbreak. Complete data is not required for Althaus' method, but is for the PBLA approach that we have adopted (in the sense that the entire length of the outbreak must have occurred, we may of course have unobserved event times). We will explore this in Section 4.3.3, but first we perform analysis with both methods on the same, larger data set.

4.3.1 Data and Models

4.3.1.1 Data

The Ebola outbreak data for Guinea, Sierra Leone and Liberia are given in Table 4.9. These have been adapted from the Centers for Disease Control and Prevention (Accessed 2018-03-11), though are originally from the World Health Organisation (Accessed 2017-11-27). This data set covers a much longer time frame than that used in the Althaus paper: from March 2014 to February 2016, which is the entire length of the outbreak for which detailed data was collected for these nations. We will consider both the Althaus method and PBLA method with this larger data set, to explore whether the parameter estimates are similar given a completed outbreak.

Both the data available from the Centers for Disease Control and Prevention (CDC) and the model used in Althaus' original analysis are not directly applicable to the PBLA method. They require some adaptation, but in order to make the most accurate assessment of PBLA we will adapt both, and re-run Althaus' analysis in its most comparable form. The original analysis was obtained from Althaus' Github repository (Accessed 2017-11-27)). The data as given in Table 4.9 are in this modified form.

The data consist of the number of deaths per day (confirmed, probable and suspected to be from Ebola), for each of the three countries. Although the data are shown in daily format for brevity, the PBLA method of course requires them as individual-based. The original Althaus data contain some days where the total number of observed cases decreases (presumably due to identification error), which our individual-based model cannot include. Hence we have modified the data to ignore any decreases in the number of cases.

Althaus' original analysis required imputation of both the cases identified per day and the deaths per day, whereas PBLA requires removal data. Case identification, of course, does not equal removal unless quarantine is immediate. We know from Althaus (2014) that control measures were put in place with the appearance of the index case. The date of this was identified as 2nd December 2013 by Baize et al. (2014), with the initial death being a one year old boy who lived in a village in Guinea and died December 6th 2013. However, details of the control measures are not provided and so we cannot be sure that quarantine was indeed immediate. On the other hand, taking the deaths data as the removals ignores all non-fatal cases. To remedy this issue, we choose to use the deaths data only in both analyses. Although this means we are discarding some of the available data, this will ensure that the PBLA parameter estimates remain comparable to those using Althaus' method. Our motivation here is to assess PBLA through comparison with Althaus' method, rather than perform a detailed data analysis. Our assumptions about, and adaptations of, the data are therefore not critical for this exercise.

The Althaus model takes longitudinal data as input, i.e. it considers the status of the system (the cumulative number of observed deaths in this case) at a discrete set of observed times. However, the PBLA method assumes that the time provided is the exact removal time for each individual. For example, from Table 4.9, that twelve individuals were infected on April 7th in Guinea, when in fact they probably became gradually infected since the last observation five days previous. To incorporate this into the data, for the PBLA method we evenly distribute all observed cases over the time period since the last observation. For computational reasons, we also scale the removal times by a factor of $\frac{1}{1000}$ for the PBLA analysis. Without this, some exponential expressions in the likelihood are too small to be calculated in R. Our parameter estimates in Section 4.3.2 will be scaled back up to be comparable to those from the Althaus method.

4.3.1.2 Model: Althaus analysis

Since we will be adapting the model used in Althaus (2014) for comparibility with the PBLA method, we will briefly describe this modified model. Further details can be found in the original paper.

The Ebola transmission will follow an SEIR model. As usual, susceptible individuals S enter the exposed class E upon infection, before moving to class I, and becoming able to infect others, at rate σ. With the deaths data as removals, individuals are moved to class R upon death at rate γ. This may be described by the following set of ordinary differential equations:

Table 4.9: WHO data concerning deaths in the West African Ebola epidemic of 2014, adapted from the Centers for Disease Control and Prevention (Accessed 2018-03-11). Details of the adaptation of the data are given in Section 4.3.1.1.

	Number of deaths				Number of deaths		
Day	Guinea	SL	Liberia	Day	Guinea	SL	Liberia
25/03/14	59	0	0	24/07/14	314	219	127
26/03/14	60	0	0	28/07/14	319	224	129
27/03/14	66	0	6	31/07/14	339	233	156
31/03/14	70	0	6	03/08/14	346	252	227
01/04/14	80	0	6	04/08/14	358	273	255
02/04/14	83	0	6	08/08/14	367	298	294
07/04/14	95	0	7	12/08/14	373	315	323
10/04/14	101	0	14	13/08/14	377	334	355
17/04/14	122	0	14	15/08/14	380	348	355
21/04/14	129	0	14	19/08/14	394	365	466
23/04/14	136	0	14	21/08/14	396	374	576
30/04/14	146	0	14	22/08/14	406	392	624
05/05/14	155	0	14	28/08/14	430	422	694
14/05/14	157	0	14	06/09/14	517	491	1089
23/05/14	174	0	14	08/09/14	555	509	1224
27/05/14	174	4	14	12/09/14	557	524	1224
28/05/14	186	5	14	16/09/14	595	562	1296
02/06/14	193	6	14	18/09/14	601	562	1459
05/06/14	215	7	14	22/09/14	632	593	1578
10/06/14	236	7	14	24/09/14	635	597	1677
11/06/14	241	19	14	26/09/14	648	605	1830
18/06/14	264	49	24	01/10/14	710	622	1998
24/06/14	270	49	34	03/10/14	739	623	2069
02/07/14	303	99	65	08/10/14	768	879	2210
07/07/14	305	101	75	10/10/14	778	930	2316
08/07/14	307	127	84	15/10/14	843	1183	2458
14/07/14	309	142	88	17/10/14	862	1200	2484
16/07/14	309	192	105	22/10/14	904	1359	2705
21/07/14	310	206	116	25/10/14	926	1359	2705

	Number of deaths		
Day	Guinea	SL	Liberia
29/10/14	997	1500	2705
31/10/14	1018	1510	2705
05/11/14	1041	1510	2705
07/11/14	1054	1510	2766
12/11/14	1054	1510	2836
14/11/14	1166	1510	2836
19/11/14	1192	1510	2964
21/11/14	1214	1510	2964
26/11/14	1260	1510	3016
28/11/14	1312	1530	3145
03/12/14	1327	1583	3145
10/12/14	1428	1768	3177
17/12/14	1525	2085	3290
24/12/14	1607	2582	3384
31/12/14	1708	2758	3423
07/01/15	1781	2943	3496
14/01/15	1814	3062	3538
21/01/15	1876	3145	3605
28/01/15	1910	3199	3686
04/02/15	1944	3276	3746
11/02/15	1995	3341	3826
18/02/15	2057	3408	3900
25/02/15	2091	3461	4037
04/03/15	2129	3546	4117
11/03/15	2170	3629	4162
18/03/15	2224	3691	4264
25/03/15	2263	3747	4301
01/04/15	2314	3799	4332
08/04/15	2333	3831	4408
15/04/15	2346	3857	4486
22/04/15	2358	3877	4573
29/04/15	2377	3899	4608
06/05/15	2386	3903	4716
13/05/15	2392	3904	4769
20/05/15	2407	3907	4806
27/05/15	2420	3908	4806
03/06/15	2429	3912	4806
10/06/15	2437	3915	4806
17/06/15	2444	3919	4806
24/06/15	2473	3928	4806
01/07/15	2482	3932	4806
08/07/15	2499	3940	4807
15/07/15	2506	3947	4808
22/07/15	2512	3949	4808
29/07/15	2520	3951	4808
05/08/15	2522	3951	4808
12/08/15	2524	3951	4808
19/08/15	2524	3952	4808
26/08/15	2527	3952	4808
03/09/15	2529	3953	4808
10/09/15	2530	3953	4808
17/09/15	2530	3953	4808
24/09/15	2532	3955	4808
01/10/15	2533	3955	4808
08/10/15	2534	3955	4808
15/10/15	2534	3955	4808
22/10/15	2535	3955	4808
29/10/15	2535	3955	4808
05/11/15	2536	3955	4808
11/11/15	2536	3955	4808

	Number of deaths		
Day	Guinea	SL	Liberia
18/11/15	2536	3955	4808
25/11/15	2536	3955	4808
02/12/15	2536	3955	4809
09/12/15	2536	3955	4809
16/12/15	2536	3955	4809
23/12/15	2536	3955	4809
30/12/15	2536	3955	4809
06/01/16	2536	3955	4809
13/01/16	2536	3955	4809
20/01/16	2536	3956	4809
Total	2536	3956	4809

$$\begin{aligned}
\frac{dS}{dt} &= -\beta(t)\frac{SI}{N} \\
\frac{dE}{dt} &= \beta(t)\frac{SI}{N} - \sigma E \\
\frac{dI}{dt} &= \sigma E - \gamma I \\
\frac{dR}{dt} &= \gamma I.
\end{aligned}$$

The averages of the latent and infectious period lengths are fixed to estimates from an outbreak of the same Ebola subtype in Congo in 1995 (Chowell et al., 2004), that is exposed period average length $\frac{1}{\sigma} = 5.3$ days and infectious period average length $\frac{1}{\gamma} = 5.61$ days. The infection rate $\beta(t)$, a function of time t which is normalised such that $t = 0$ corresponds to the first recorded death in the data for each country, is given here by

$$\beta(t) = b_0 \mathrm{e}^{-k(t+\tau_0)},$$

for constants b_0, k and τ_0, to be estimated. We define parameter τ_0 as the time of infection of the initial infective in the given country. For Guinea, Althaus fixes this to the known date of the start of the outbreak in the country: December 2nd 2013. For Sierra Leone and Liberia, this parameter is estimated.

We set the total population size for each country to be 10^6, as in the original analysis. Parameter estimates for b_0, k and τ_0 are obtained via maximum likelihood estimation, assuming the cumulative number of deaths is Poisson distributed.

4.3.1.3 Model: PBLA analysis

In order to ensure our analysis is comparable to that of Althaus, we use the most equivalent model possible for the PBLA analysis. We fit an SEIR model which is, so far as possible, the stochastic analogue of that described in Section 4.3.1.2 above.

We assume that each of the three countries form a closed population of $N = 10^6$ individuals labelled $1, 2, \ldots, N$, of whom $1, 2, \ldots, n$ become infected, where $n = 2536$, 3956 and 4809, respectively, in Guinea, Sierra Leone and Liberia. We assume that the initial case in each country was externally infected.

The SEIR transmission model will be much the same as for Althaus' analysis, though defined stochastically. At any given time, each individual in the population will be either in state S (susceptible), state E (exposed), state I (infective), or state R (removed). For each individual $j = 1, 2, \ldots, n$, e_j, i_j and r_j denote, respectively, their time of exposure (unknown), infectivity (unknown), and removal (known). Any susceptible individual may become exposed and enter the latent E stage during which they are infected, but not yet infectious. They next enter the I stage at which point they become able to infect others. They last become removed and enter the R stage, during which they are unable to cause new infections and are considered removed from the population.

The outbreak begins with the initial case κ entering their infectious period at time i_κ, since this is the point at which infectious pressure begins to be applied, and ends when there are no exposed or infective individuals remaining in the population. The exposure times $\mathbf{e} = \{e_j : j = 1, 2, \ldots, \kappa - 1, \kappa + 1, \ldots, n\}$ and infection times $\mathbf{i} = \{i_j : j = 1, 2, \ldots, n\}$ are unknown, and the data consist

of removal times $\mathbf{r} = \{r_j : j = 1, 2, \ldots, n, \text{ where } r_1 < r_2 < \cdots < r_n\}$. During their infectious period only, any infective j will have a contact with any other individual k at a time given by the point of a Poisson process of rate β_{jk}, where all Poisson processes are assumed mutually independent. Any contact between an infective and a susceptible is assumed to result in immediate infection. We also assume that the infectious periods are mutually independent and exponentially distributed with known parameter $\gamma = \frac{1}{5.61}$ so that the infectious periods are of mean length 5.61 days, as well as the latent periods being of fixed length $\frac{1}{\sigma} = 5.3$ days.

Since the PBLA framework does not allow for an infection rate dependent upon time as in Althaus' analysis, we define β_{jk} as a proxy-time-dependent infection rate:

$$\beta_{jk} = b_0 \mathrm{e}^{-k(T_{jk}+\tau_0)}.$$

Here, T_{jk} approximates the midpoint of the time period for which there is infectious pressure between individuals j and k, which under the PBLA method is the only time we consider quantity β_{jk}. For infective j and individual k who also eventually becomes infected, T_{jk} is therefore defined as:

$$\begin{aligned} T_{jk} &= \frac{1}{2}\Big(\mathbb{E}[r_j \wedge e_k] + \mathbb{E}[i_j \wedge e_k]\Big) \\ &= \begin{cases} r_k - \frac{1}{\gamma} - \frac{1}{\sigma} - \frac{1}{4\gamma}\mathrm{e}^{-\gamma(r_j - r_k + \frac{1}{\sigma})} & \text{if } r_j > r_k - \frac{1}{\sigma}, \\ r_j - \frac{1}{2\gamma} + \frac{3}{4\gamma}\mathrm{e}^{-\gamma(r_k - r_j - \frac{1}{\sigma})} & \text{if } r_j \leq r_k - \frac{1}{\sigma}. \end{cases} \end{aligned}$$

For all individuals k in the population who did not become infected, we define

$$T_{jk} = r_j - \frac{1}{2\gamma},$$

which again approximates the midpoint of the time period for which infective j put infectious pressure on k. There are many other possible versions of T_{jk} that could be considered, representing different stages through the time at which there is pressure between j and k, but a brief investigation of the most extreme possibilities revealed that this choice did not significantly impact the results.

In β_{jk}, we scale the infection rates by fixing τ_0 to the corresponding estimates from the Althaus analysis for the outbreak start date in each country. This leads to a heterogeneously mixing population where we only need to estimate quantities b_0 and k.

Although we could use MCMC with the PBLA likelihood to find parameter estimates, in this case we will optimise the likelihood, using the *optim* function in R. This results in even faster computation, and is comparable to the MLEs obtained with Althaus' method.

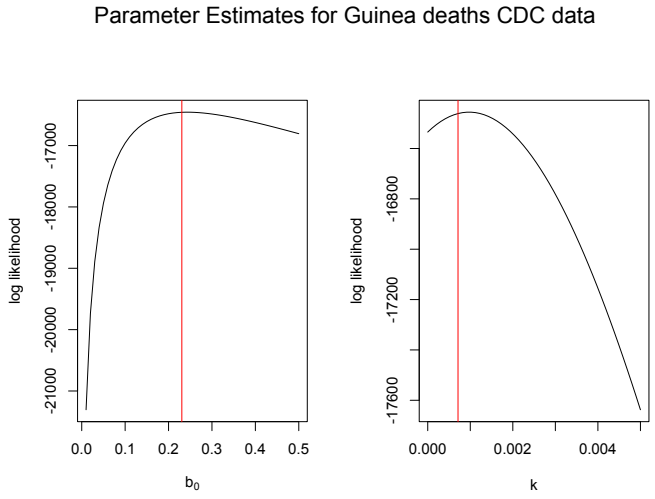
4.3.2 Results

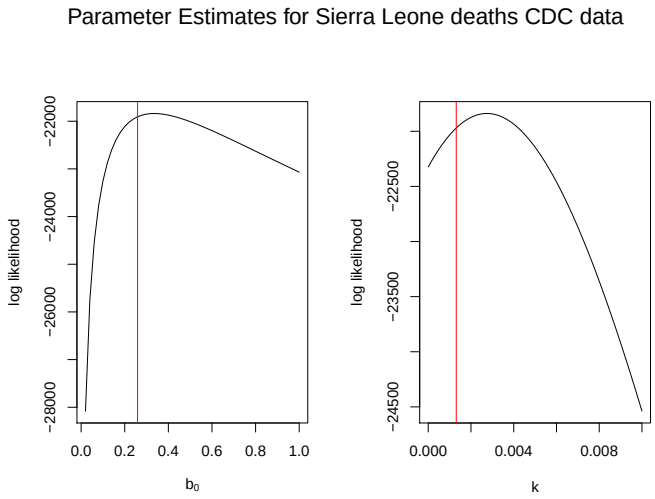
With both models defined, we may progress to comparing results from both methods of analysis. Table 4.10 provides the MLEs of b_0 and k obtained under each method and for each of the three countries. We see that the estimates obtained are, in general, very similar under both models/methods. In the Althaus analysis we also estimate τ_0, and then use the respective estimate for each country in the PBLA likelihood. This corresponds to $\tau_0 = 113, 58$, and 0 in Guinea, Sierra Leone and Liberia, respectively.

Figure 4.25 contains profile likelihoods for parameters b_0 and k using the PBLA log-likelihood, where the red lines provide the corresponding MLE from the Althaus method. Figure 4.26 then shows contour plots for these parameters, with the Althaus MLEs shown by the black points. We see again that both methods provide similar estimation, indicating that the PBLA approximation is performing well. This also indicates that the proxy-time-dependent infection rate used with PBLA is a good approximation to the original rate used in the Althaus analysis.

4.3.3 Original Althaus data analysis

As we have discussed, one fundamental difference between the Althaus method and PBLA is that the Althaus method uses the data in a longitudinal format,





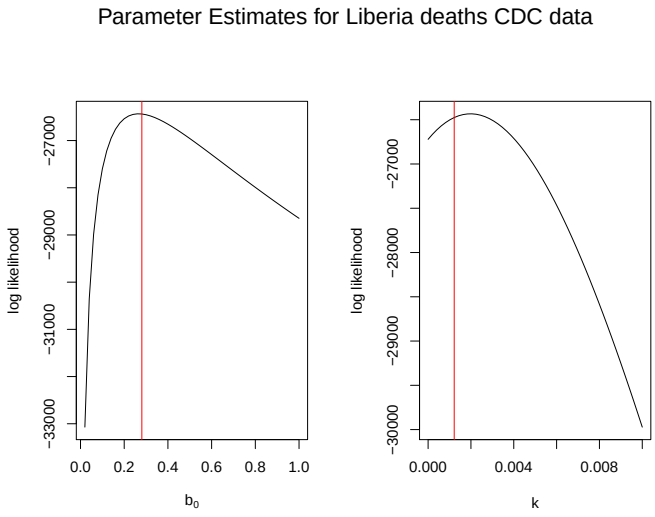


Figure 4.25: Profile likelihoods for b_0 and k under the PBLA log-likelihood, using the full CDC data. Red lines display the corresponding MLEs using the Althaus method.

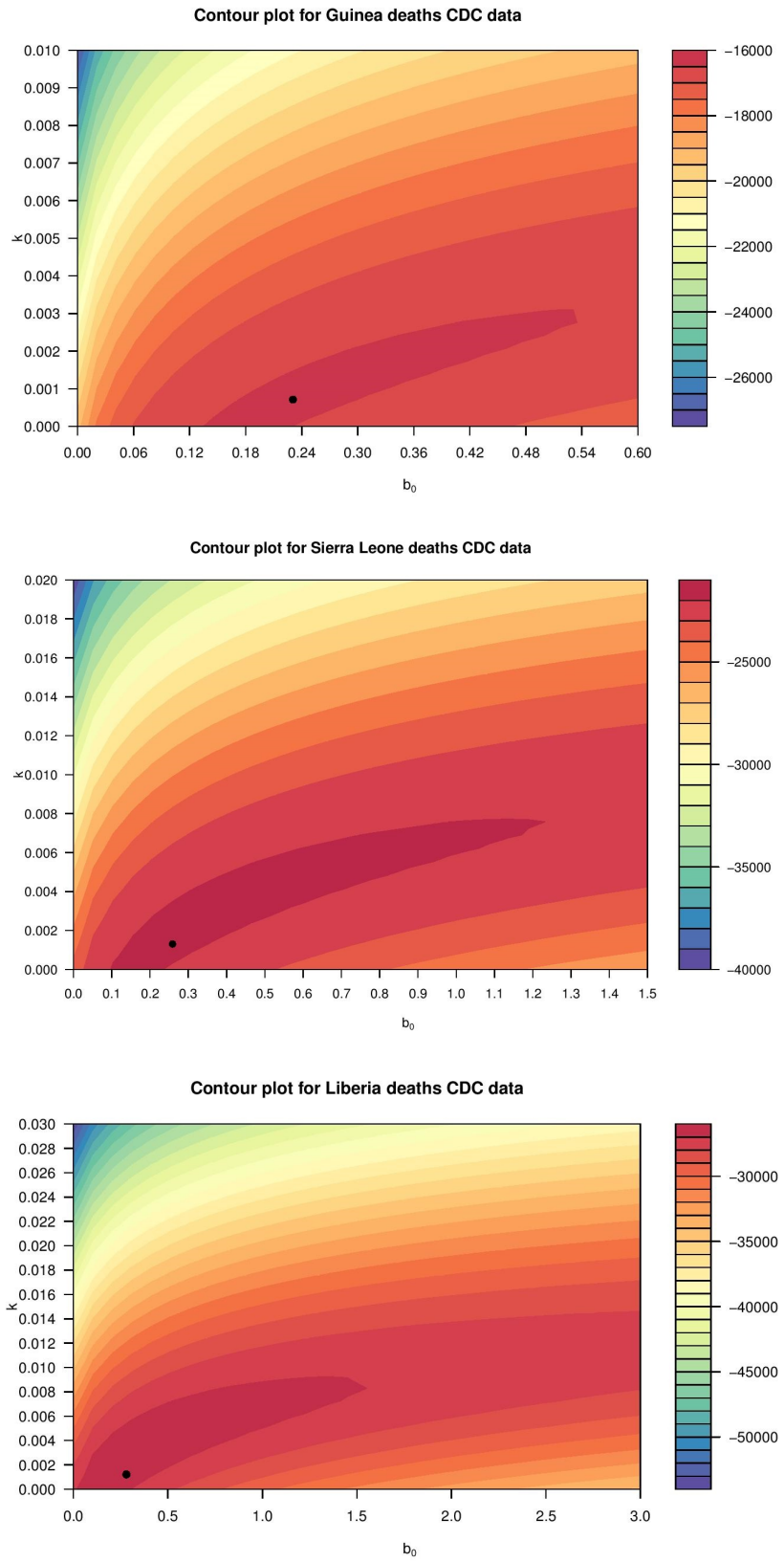


Figure 4.26: Contour plots for b_0 and k under the PBLA log-likelihood, using the full CDC data. Black points indicate the corresponding MLEs using the Althaus method.

Table 4.10: MLEs for the West African Ebola outbreak, using the full CDC data with the Althaus and PBLA methods.

		b_0	k
Althaus	Guinea	0.2306	0.0007118
	Sierra Leone	0.2766	0.001797
	Liberia	0.3028	0.002509
PBLA	Guinea	0.2429	0.001054
	Sierra Leone	0.3347	0.002890
	Liberia	0.2659	0.002141

whereas PBLA takes individual-based observations and, as part of this, assumes that we have observed all removals until the end of the outbreak. This is not necessary for the Althaus method which simply checks the status of the population at some discrete set of times, whether or not the last time represents the end of the outbreak. In order to avoid use of incomplete outbreak data we have performed the analysis so far with the larger data set from the CDC. Next, however, we will re-perform the analysis with the original Althaus data to demonstrate what happens to the PBLA likelihood when it is applied to an incomplete outbreak.

4.3.3.1 Data and models

The data are given in Table 4.11, adapted from Althaus (2014) though again originally from the World Health Organisation (Accessed 2017-11-27). The data consist of the number of deaths per day (confirmed, probable and suspected to be from Ebola), for each of the three countries. We make the same adaptations to the data as before. We ignore any decreases in the observed number of deaths, with this modification already applied to the data set given in Table 4.11. We also take only the deaths data, scaled by a factor of $\frac{1}{1000}$ for the PBLA analysis as before, and smooth out the deaths equally over the region during which they may have occurred. Again, for the PBLA analysis we

convert these data to an individual based format.

Comparing the CDC data in Table 4.9 to the Althaus data in Table 4.11 over the region which they both include, we see that the two data sets are generally very similar, but with some small differences. Notably, we see that the dates have been shifted by one or two days. This may be explained as the Althaus data providing the date on which the death was predicted to have happened, and the CDC data providing the date on which the WHO report was published. Since the scale is equivalent in both cases (just shifted) and we will use the same data set with both methods of analysis, we make no attempt to align the two.

Using the same models as in Section 4.3.1 for both the Althaus and PBLA analysis, we may now obtain parameter estimates and investigate the impact of an incomplete outbreak.

4.3.3.2 Results

To begin our investigation we plot the observed cumulative deaths in the Althaus data over time, and seek to compare this to Althaus' model simulation for the same period. As discussed in Section 4.3.1.2, Althaus assumes the observed number of deaths is Poisson distributed with parameter equal to the expected number of deaths from the ODE model. As in, with mean and variance equal to this expected number of deaths. We therefore fit the corresponding curve of Althaus' ODE model estimates (for each day in the outbreak) over the observed data, along with error bars to display this assumed variability.

These plots are given in Figure 4.27 for each of the affected countries. We include two plots for Liberia, the first over a shorter time where we can compare the behaviour of the Althaus simulation over the time frame of the outbreak, and the second where we consider the entire simulated outbreak. We see that, in general, the behaviour of the simulations is very similar to that observed during the course of the observed outbreak, but that for Guinea and

Table 4.11: WHO data concerning deaths in the West African Ebola epidemic of 2014, adapted from Althaus (2014) though originally from the CDC. Details of the data adaptation are given in Section 4.3.3.1.

	Number of deaths		
Day	Guinea	Sierra Leone	Liberia
22/03	29	0	0
24/03	30	0	0
25/03	1	0	0
26/03	2	0	0
27/03	4	0	0
28/03	4	0	0
31/03	10	0	0
01/04	3	0	0
04/04	3	0	0
07/04	9	0	0
09/04	6	0	0
11/04	5	0	0
14/04	2	0	0
16/04	14	0	0
17/04	7	0	0
20/04	7	0	0
23/04	5	0	0
01/05	8	0	0
03/05	6	0	0
05/05	2	0	0
07/05	1	0	0
10/05	0	0	0
12/05	13	0	0
18/05	5	0	0
23/05	0	0	0
27/05	10	5	0
28/05	7	0	0
29/05	0	1	0
01/06	15	0	0
⋮	⋮	⋮	⋮

	Number of deaths		
Day	Guinea	Sierra Leone	Liberia
⋮	⋮	⋮	⋮
03/06	7	0	0
05/06	11	0	0
06/06	0	1	0
15/06	37	39	0
16/06	1	0	24
17/06	0	3	0
20/06	6	0	0
22/06	0	0	10
30/06	33	50	31
02/07	2	2	10
06/07	2	26	9
08/07	2	15	4
12/07	0	52	17
14/07	1	3	1
17/07	0	9	10
20/07	4	13	11
23/07	5	5	2
27/07	20	9	27
30/07	7	19	71
01/08	12	21	28
04/08	5	13	27
06/08	4	12	12
09/08	6	17	29
11/08	4	19	32
13/08	3	14	58
15/08	0	0	53
16/08	14	17	0
18/08	2	9	110
20/08	10	18	48
Total	406	392	624

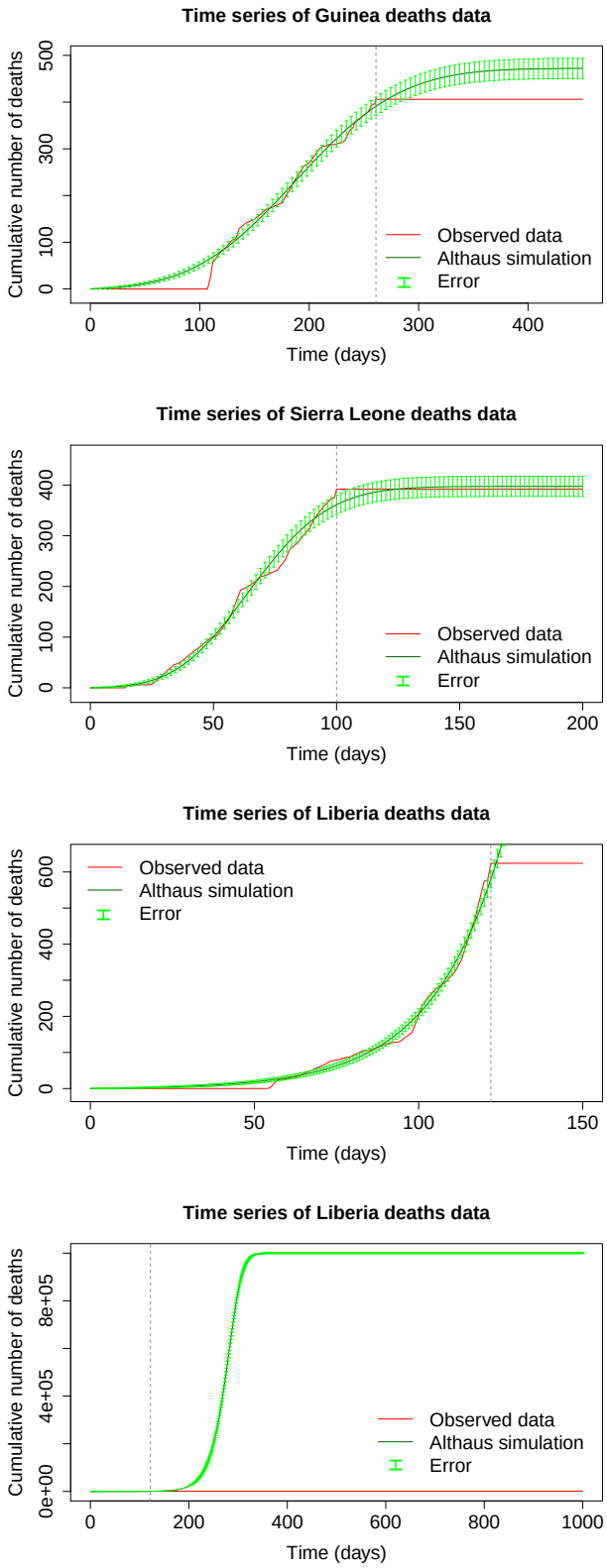


Figure 4.27: Cumulative observed deaths in each country overlaid with equivalent estimations from the Althaus simulation method, with error bars displaying the assumed variability. Dashed lines show the end of the observed data.

Liberia, there are many more future deaths predicted. This is particularly true for Liberia where, under the parameter values as estimated, it is predicted that all 10^6 individuals will eventually be infected in an outbreak spanning around a year. This might indicate that the PBLA method will not result in similar parameter estimates to the Althaus method for this shorter dataset, since for that to occur in Liberia, for example, under the PBLA assumption of a completed outbreak we would expect to have seen a final size much closer to N.

Further investigating this, we proceed to our full analysis using both the PBLA and Althaus methods with the original Althaus data. Table 4.12 contains parameter estimates for b_0 and k. We again estimate τ_0 in the Althaus analysis, and then use these in the PBLA analysis. For this data, τ_0 was estimated as 110, 15 and 57 for Guinea, Sierra Leone and Liberia, respectively. Figure 4.28 displays profile likelihoods for b_0 and k for each country using the PBLA method, and Figure 4.29 displays corresponding contour plots for these parameters, comparing the results to the MLEs from the Althaus method in each case. As we can see from the table as well as the plots, the parameter estimates from PBLA are somewhat similar to the Althaus estimates, but certainly not as close as when using the larger data set. Estimation is fairly similar for Sierra Leone, but for Guinea and particularly Liberia (where we saw the least complete outbreak), we are not so able to accurately estimate the parameters. For Liberia especially, we note that the k estimate is considerably different and that, from Figure 4.29, the contour is very flat for b_0.

We may conclude that the assumption of a complete outbreak is especially important when using the PBLA method, since under an incomplete outbreak the method is far less able to achieve accurate parameter estimation. However, as we have seen, with the larger, complete data set PBLA offers very similar performance to the Althaus method.

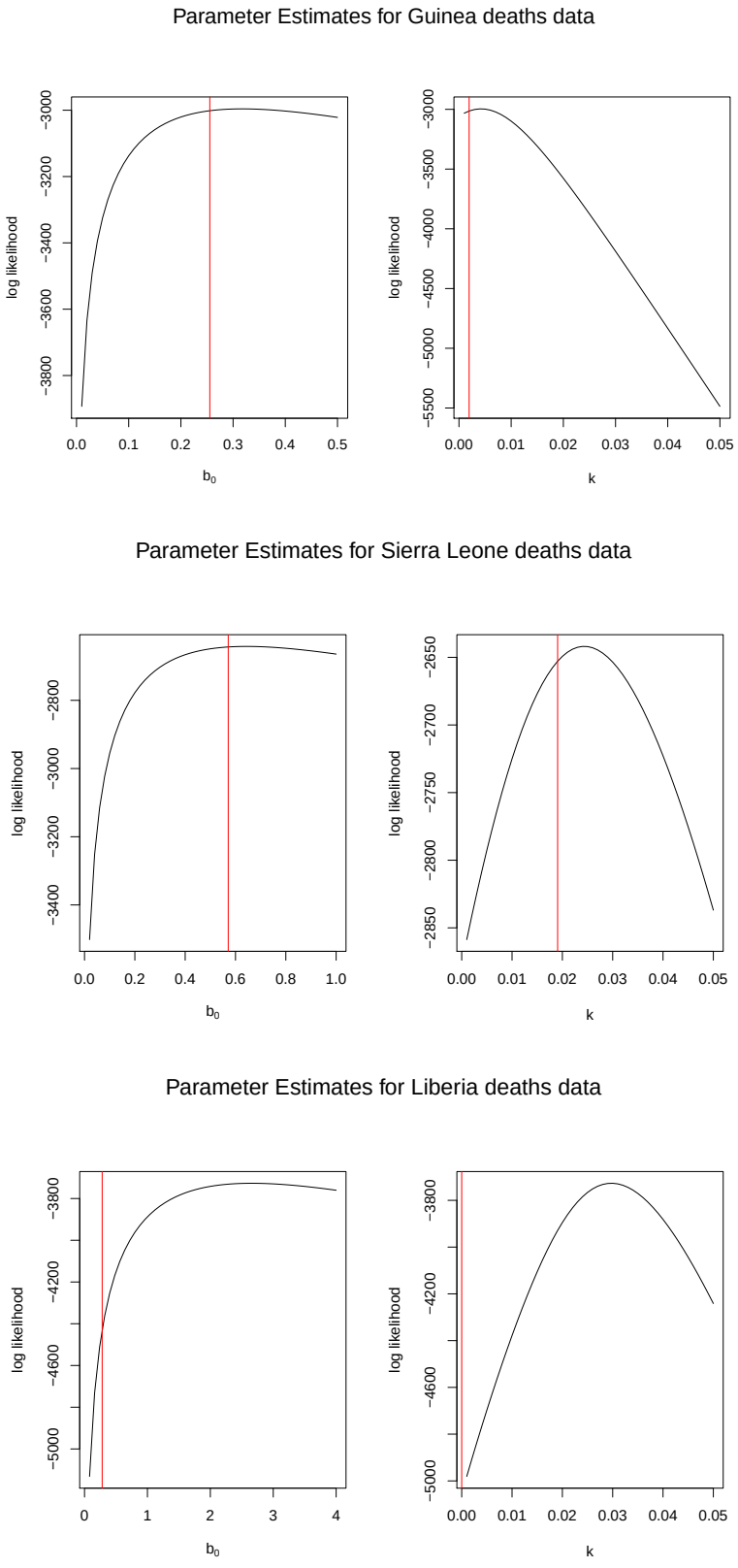


Figure 4.28: Profile likelihoods for b_0 and k under the PBLA log-likelihood, using the Althaus data. Red lines display the corresponding MLEs from the Althaus method.

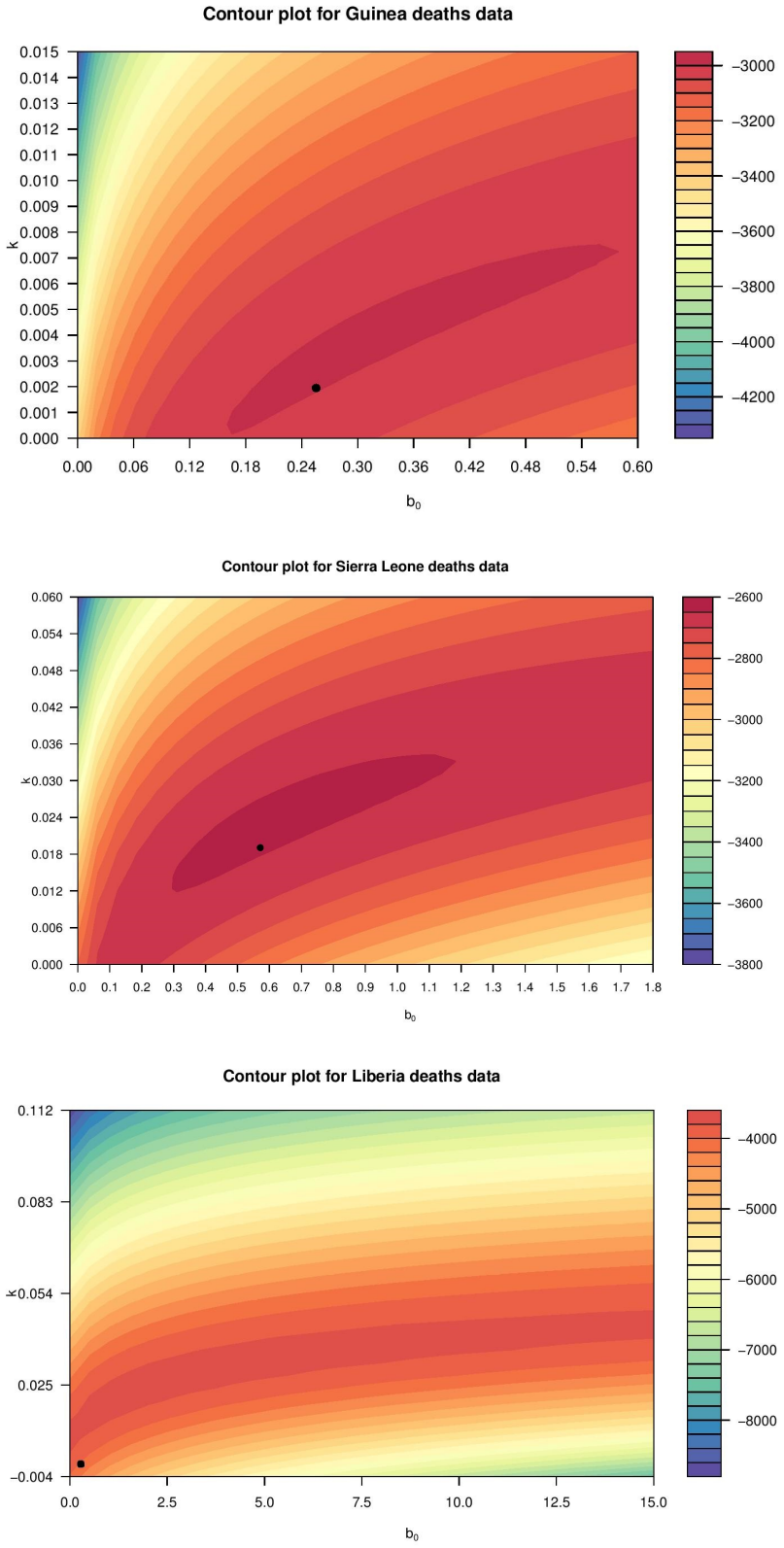


Figure 4.29: Contour plots for b_0 and k under the PBLA log-likelihood, using the Althaus data. Black points indicate the corresponding MLEs using the Althaus method.

Table 4.12: Parameter estimates for the West African Ebola outbreak. The τ_0 estimates from the Althaus method were used in the PBLA analysis.

		b_0	k
Althaus	Guinea	0.2554	0.001948
	Sierra Leone	0.5711	0.01906
	Liberia	0.2806	0.0000002079
PBLA	Guinea	0.3173	0.004224
	Sierra Leone	0.6436	0.002434
	Liberia	2.6555	0.02974

4.4 Applications: 2001 UK Foot and Mouth outbreak data

The final application to be explored concerns data from the 2001 UK Foot-and-Mouth disease (FMD) epidemic.

FMD is a viral infection which affects mainly hoofed livestock (cows, sheep, pigs), but has also been known to affect animals such as antelope and hedgehogs as well as, occasionally, humans. Although FMD is rarely fatal, it causes significant drops in the dairy production of cattle, as well as slow weight gain and blisters which may cause lameness (Alexandersen et al., 2003). The disease is highly infectious, and may be spread not just through animal-animal contact but also contact with farming equipment, vehicles and feed. There is hence considerable focus on containment of outbreaks, so as to avoid large economic loss through trade restrictions and culling. This containment involves vaccination strategies and strict monitoring of farms (Grubman and Baxt, 2004). There are difficulties in vaccination, however, due to large variation between different serotypes of the disease, with no cross-protection between these (Brown, 1992).

The 2001 UK outbreak of FMD had a significant impact on the country, seeing widespread culling as well as non-farming related impacts such as the cancellation of sporting/leisure events and the postponement of the general election. The entire outbreak saw more than 2000 identified cases, and around 10,000,000 cattle and sheep were eventually culled to stop the outbreak with a total estimated cost of £8 billion to the economy.

The Department for Environment, Food and Rural Affairs (DEFRA) collected detailed information on the FMD outbreak, which was released on their website (**www.defra.gov.uk**) in 2003. There have been numerous mathematical analyses of this data, including Morris et al. (2001), Ferguson et al. (2001) and Diggle (2006). We will focus on the analysis of data from Cumbria and the surrounding area, which was the county most severely affected by the outbreak. We will compare our analysis with that of Kypraios (2007), who used a true likelihood-based approach to infer infectivity and susceptibility parameters. We will use the PBLA III method, with the same SIR and heterogeneous mixing model as Kypraios (2007) for comparison. This will involve a spatial component in the infection rate, so that the contact rate between farms depends on the geographic distance between them.

We will first describe this data set, and then fully define the model to be used in this analysis.

4.4.1 Data and Model

4.4.1.1 Data

The data we will use are the same as those used in Kypraios (2007). These were made available by Kypraios who obtained them from the file "DataForModellersOct03.xls" which was previously available on the DEFRA website but has since been removed. In our previous analyses, the individuals in the population have all been humans, but in this analysis each individual represents a Cumbrian farm. The data include dates of culling of all infected farms,

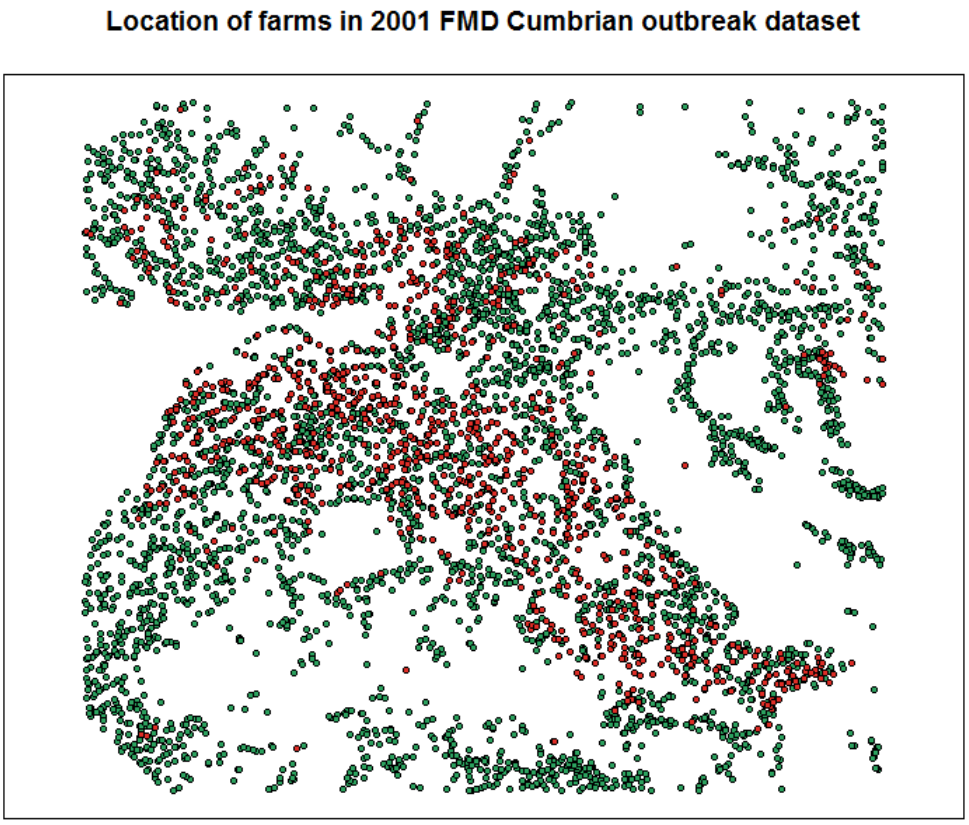


Figure 4.30: Geographical locations of Cumbrian and surrounding farms included in the 2001 Foot and Mouth disease dataset. Red points represents infected farms and green points represent those that were not infected.

which we take to be equal to removal data in the same way as Kypraios (2007). The infection times are then unknown.

As well as the removal data, DEFRA made available the geographic location of all Cumbrian farms, given as x and y coordinates of each. Figure 4.30 shows a plot of these. Some farms in the data are actually in the counties surrounding Cumbria, but we include these in the analysis for consistency with Kypraios (2007).

Information was also available on the total numbers of cows and sheep on each farm, which we will incorporate. There was additional data on the numbers of pigs, goats and deer on each farm, but to maintain comparability with

Kypraios (2007) we ignore these, and also exclude any farms with no cows or sheep.

This results in a total of $N = 5378$ farms, of which $n = 1021$ were culled. Some of these were culled without knowledge of their infection status, say if they were considered to have had dangerous contacts. However, no information is available as to which farms these were, and so we consider all cullings as the removal of infected cases, as in Kypraios (2007).

4.4.1.2 Model

The model used for analysing the FMD data will be the same as in Kypraios (2007) to allow for the best comparison of the PBLA method, and we briefly describe this here. We assume a closed population of N farms labelled $1, \ldots, N$, of which $1, \ldots, n$ become infected and where the initial case is assumed to have been infected externally. We use an SIR model as defined in Section 3.2, so that at any time t, each farm will be either susceptible (S), infective (I) or removed (R). For all infectives $j = 1, \ldots, n$, i_j and r_j denote their time of infection and removal, respectively.

The initial case κ (unknown) becomes infected at i_κ, starting the outbreak. This then ends when no infectives remain in the population. Infection times $\mathbf{i} = \{i_j : j = 1, 2, \ldots, \kappa - 1, \kappa + 1, \ldots, n\}$ are unknown as usual, and the data consist of removal times $\mathbf{r} = \{r_j : j = 1, 2, \ldots, n$ where $r_1 < r_2 < \cdots < r_n\}$, in addition to the geographical location and the number of cows and sheep of each farm.

A heterogeneously mixing population structure is assumed, such that the infection rate from farm i to farm j is given by β_{ij}. Therefore, during their infectious period an infective i will have a contact with an individual j at a time given by the point of a Poisson process of rate β_{ij}, where all such Poisson processes are assumed mutually independent. Contacts between an infective and a susceptible are assumed to result in immediate infection. We assume that

the infectious periods are independent and gamma distributed, with shape parameter m and rate parameter γ, so that $r_j - i_j \sim \Gamma(m, \gamma)$. We also assume that the shape parameter is known and fixed to $m = 4$.

To incorporate the different aspects of the data, the infection rate β_{ij} is defined as

$$\beta_{ij} = \beta_0 \times K(i,j) \times \left(\epsilon(n_i^c)^\zeta + (n_i^s)^\zeta\right) \times \left(\xi(n_j^c)^\zeta + (n_j^s)^\zeta\right),$$

where

$$K(i,j) = \frac{v}{\rho(i,j)^2 + v^2}.$$

Here, β_0 is a constant representing the overall average infection rate. Parameters ϵ and ξ represent the relative infectiousness and susceptibility, respectively, of cows to sheep, where n_i^c and n_i^s are the known numbers of cows and sheep on farm i. The level of linearity (or sub-linearity) of the infectivity/susceptibility of each farm to the number of animals is given by ζ. Lastly, $K(i,j)$ is a Cauchy kernel determining the spatial aspect of the infection rate. With $\rho(i,j)$ defined as the Euclidean distance between farms i and j, v is then the parameter defining this spatial component. Kypraios (2007) discusses the use of other potential measures of distance than the Euclidean distance (for example the minimum walking distance), but as usual we keep our model as similar as possible to theirs for comparability.

We hence have a six parameter model, with parameters $(\beta_0, \gamma, v, \epsilon, \xi, \zeta)$ which we wish to obtain estimates for. We will perform maximum likelihood estimation and MCMC with the PBLA likelihood, and compare the results to those obtained in Kypraios (2007) via DA-MCMC.

4.4.2 Results

We perform parameter estimation using the PBLA likelihood with both maximum likelihood estimation and MCMC, which we may compare with the results of a partially non-centered DA-MCMC algorithm from Kypraios (2007). We use the same gamma-distributed prior distributions for the model param-

Table 4.13: Prior distributions used for FMD data analysis. For each parameter, we use a gamma distributed prior with shape parameter m and rate parameter λ.

	m	λ
β_0	0.001	0.001
γ	0.001	0.001
v	1	0.1
ϵ	1	0.001
ξ	1	0.001
ζ	1	0.001

eters as Kypraios (2007), and these are given in Table 4.13. MCMC was coded in C, and 50,000 samples were obtained for each parameter. We performed Gaussian random walk updates for each of the six parameters separately. Maximum likelihood estimation was performed in R using the *NLM* (Non-Linear Minimisation) optimisation package. The posterior distribution has many local maxima and is of course six dimensional, leading to a challenging maximisation problem, but upon testing of a variety of maximisation functions, *NLM* was found to perform best.

Table 4.14 contains the MLEs and MCMC posterior means obtained using the PBLA likelihood, in addition to the posterior means from Kypraios' DA-MCMC. Figure 4.31 contains profile likelihoods for each of these six model parameters, when the other parameters are fixed to their MLEs. These are compared to the posterior means from Kypraios (2007). Lastly, Figure 4.32 shows trace plots for the PBLA MCMC samples alongside the DA-MCMC samples used in Kypraios (2007). We see that generally estimation is very similar between PBLA and DA-MCMC, the only parameter with potentially significant underestimation being v. In the MCMC trace plots, both methods seem to be exploring the same areas of the parameter space, except for v and γ. It is interesting that the areas of high density for γ under each method appear so

Table 4.14: MLEs and posterior means for FMD data model parameters using the PBLA likelihood, compared with DA-MCMC posterior means from Kypraios (2007). NOTE: the β_0 estimate was not provided in Kypraios (2007), but was obtained from the author.

	PBLA MLE	PBLA MCMC posterior mean	DA-MCMC posterior mean
β_0	7.02×10^{-7}	7.92×10^{-7}	6.07×10^{-7}
γ	0.445	0.448	0.52
v	0.00477	0.00474	0.0065
ϵ	1.576	1.652	1.45
ξ	2.389	2.383	2.32
ζ	0.318	0.315	0.32

distinct in Figure 4.32, when in Table 4.14 and Figure 4.31 the estimates appear fairly similar. One other notable conclusion is the similarity in the MLEs and MCMC means using the PBLA likelihood, highlighting that maximum likelihood estimation methods are able to obtain very similar estimates, but at a much smaller computational cost.

With five of the model parameters $(\beta, v, \epsilon, \xi, \zeta)$ involved in defining the overall infection rate β_{ij}, we also consider how well the PBLA method estimates this as a whole compared to DA-MCMC. Figure 4.33 includes histograms of β_{ij} estimates, obtained from the MCMC samples. Since β_{ij} obviously depends on the pair of farms (i, j), we have selected six such pairs at random. For each pair, we then calculate β_{ij} using the set of parameter samples at each iteration of the MCMC algorithm, for both PBLA MCMC and DA-MCMC. As we can see from Figure 4.33, PBLA tends to underestimate β_{ij} slightly, though in general is fairly similar to DA-MCMC. Combined with the underestimation of gamma, this is similar to what we saw in Section 4.1; that although PBLA often results in some underestimation of β and γ, estimation of their ratio (R_0) will usually still be good. For the pairs (i, j) in Figure 4.33, Figure 4.34 includes

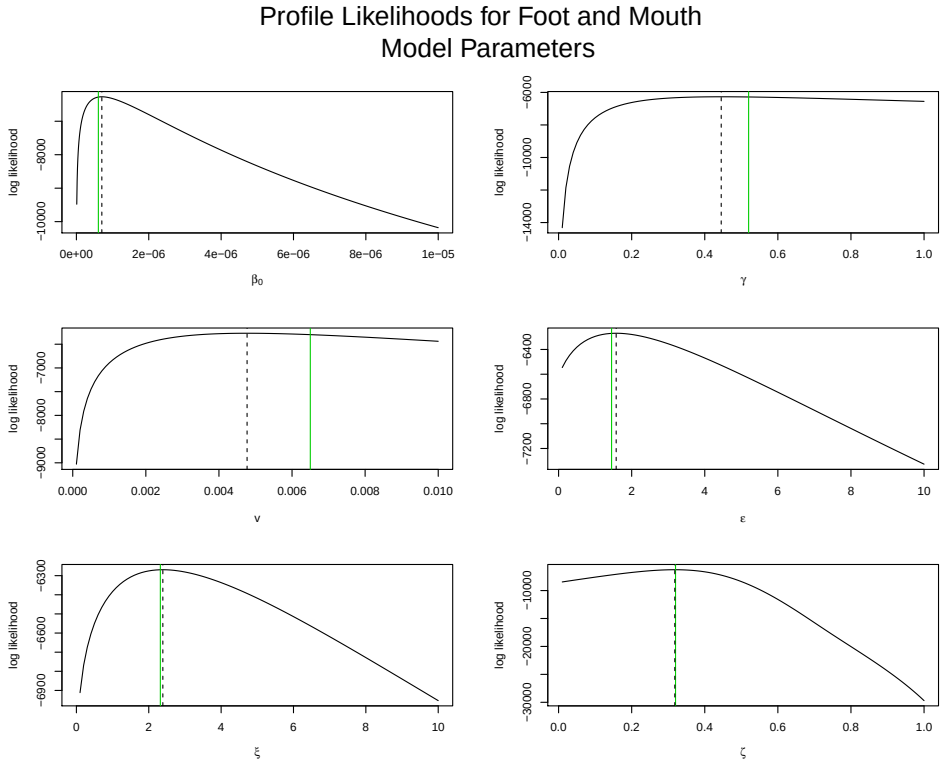


Figure 4.31: Profile likelihoods for the FMD model parameters. All parameters not being profiled are fixed to their PBLA MLEs. Dotted lines mark the MLE for the parameter in question, and green lines provide posterior mean estimates from Kypraios (2007).

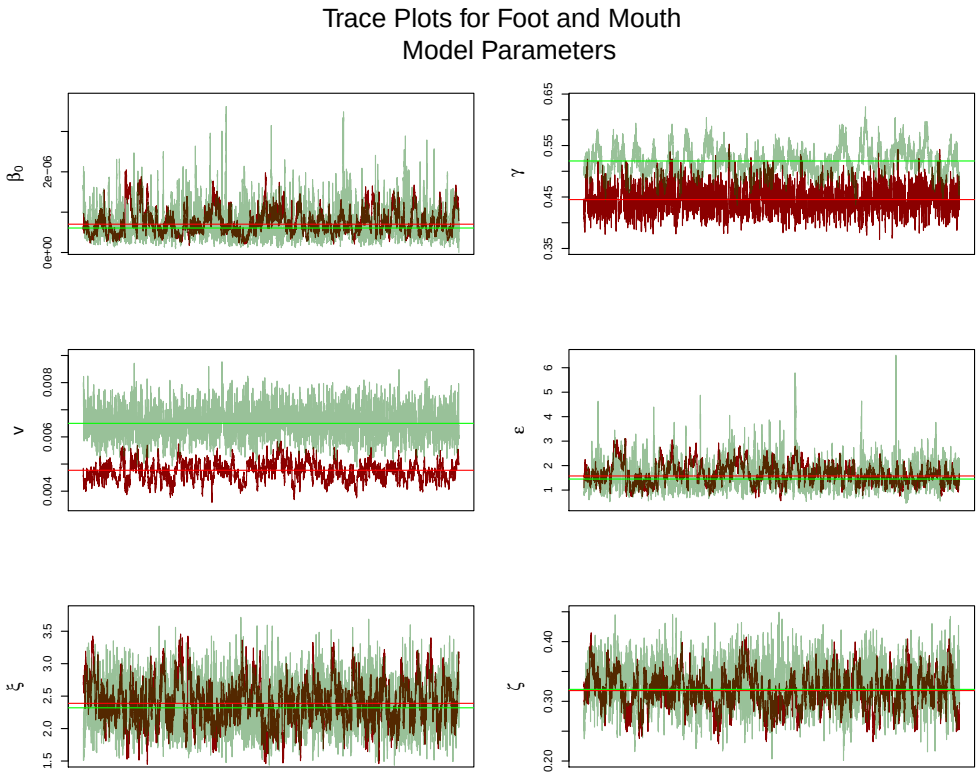


Figure 4.32: Trace plots of MCMC samples for the six FMD model parameters, both from PBLA MCMC (red) and DA-MCMC (green). DA-MCMC samples were not provided in Kypraios (2007), but were obtained from the author. Red horizontal lines mark the PBLA MCMC posterior mean and green horizontal lines mark the DA-MCMC posterior mean, for each parameter.

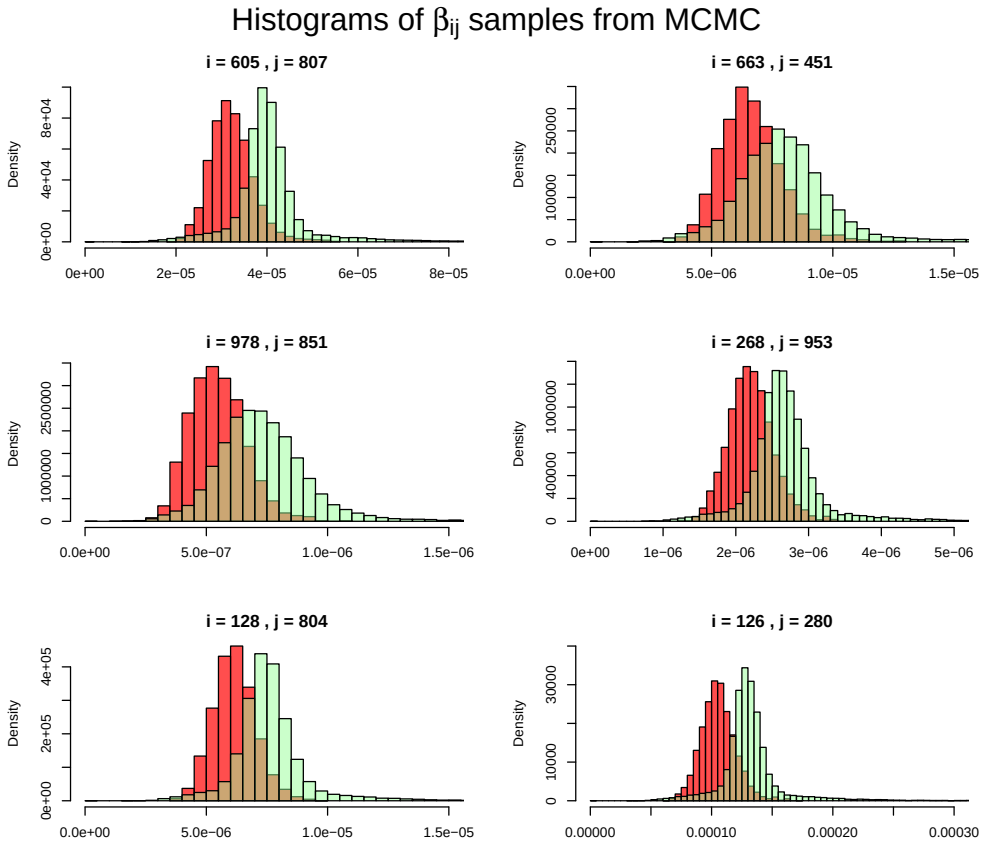


Figure 4.33: Histograms of estimates of β_{ij} for the FMD data using MCMC samples, for six randomly selected pairs of farms (i, j). Red bars are PBLA MCMC samples, and green are from DA-MCMC.

the corresponding estimates of $R_0^{ij} = \frac{\beta_{ij}}{\gamma}$. As a reproduction number this is less interpretable than R_0 for a homogeneously mixing epidemic, since it depends on the given pair (i, j) (and hence does not represent the overall infectivity of the disease). However, for our purposes here we demonstrate that the ratio of β and γ is estimated well. Overall, PBLA offers generally similar estimation to DA-MCMC for the FMD data, despite the approximation involved.

4.5 Conclusions

In this chapter we have assessed the performance of the ED and, more thoroughly, the PBLA methods, using both simulated and real data.

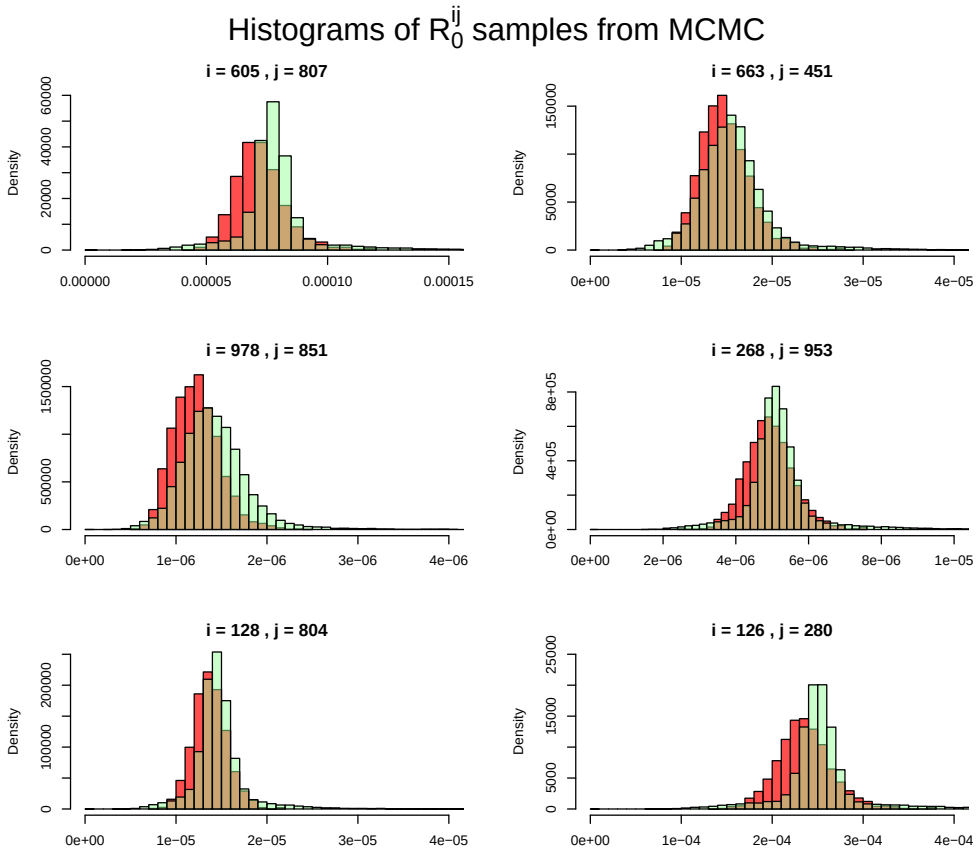


Figure 4.34: Histograms of estimates of $R_0^{ij} = \frac{\beta_{ij}}{\gamma}$ for the FMD data using MCMC samples, for six randomly selected pairs of farms (i, j). Red bars are PBLA MCMC samples, and green are from DA-MCMC.

The simulation studies of Section 4.1 highlighted where the likelihood approximation methods provide similar estimation to DA-MCMC. We saw in Section 4.1.1 that generally the PBLA III method provides similar estimation to DA-MCMC, and outperforms the ED method. Both methods perform better with gamma distributed infectious periods than exponential, though they struggle when very large or small proportions of the population are infected. This was corroborated by our findings in Section 4.1.2, where we suggested that the poor performance of PBLA when the proportion of infectives is large may be due to the ψ term in the likelihood. This term involves the most approximation, and becomes much larger compared to the other terms when the proportion of infectives is high. In this section, we also confirmed our findings that the values of β, γ and N do not affect the accuracy of the approximation methods, as well as that both methods perform considerably better with gamma infectious periods compared to exponential, even for shape parameter values as low as $m = 2$.

In Section 4.1.3 we compared the performance of the different PBLA versions as defined in Chapter 3. Under both exponential and gamma infectious periods as usual, we assessed the ability of each the PBLA versions I through V to recover true parameter values from simulated data. We found that for exponential periods where PBLA IV is applicable, this offers the fastest option at comparable accuracy to the other versions (so long as the outbreak is large enough for the central limit theorem approximation to hold). Otherwise, PBLA III offered consistently good estimation. For gamma infectious periods, PBLA III and V were found to be preferable.

We concluded the simulation studies with an analysis of computation time in Section 4.1.4. Here we found that for larger population sizes, PBLA offers improved computation time, compared to DA-MCMC. The larger the value of shape parameter m for gamma distributed infectious periods, the larger a population size is required for PBLA to be faster. As the population size increases, PBLA offers greater and greater computational advantage over DA-MCMC,

which is particularly useful in practice since there is most motivation for using likelihood approximations for larger populations.

Sections 4.2, 4.3 and 4.4 explored the application of the PBLA method to a variety of real data sets. We began with a study of respiratory disease on the island of Tristan Da Cunha. Data from this island have received considerable attention from the epidemic modelling community, since Tristan Da Cunha represents an (approximately) closed population on which detailed information has been recorded on outbreaks and the structure of its population. We compared analysis of an outbreak from 1967 using PBLA to that of Hayakawa et al. (2003), who used DA-MCMC. This involved a model where the infection rate differed with age, with individuals categorised into one of three age groups. We found that the PBLA method resulted in parameter estimates very similar to DA-MCMC.

Section 4.3 concerned analysis of data from the 2014 Ebola virus outbreak in West Africa, specifically in Guinea, Sierra Leone and Liberia. This outbreak received much attention worldwide, especially due to Ebola's high fatality and lack of cure. We compared analysis using the PBLA method with maximum likelihood estimation to that of Althaus (2014), who used an ODE model combined with maximum likelihood estimation. Since these methods are innately different (for instance, the ODE model requires daily data rather than individual based), some adjustments to the data were required. We also introduced a proxy-time-dependent infection rate to best match the time dependent rate used in Althaus (2014). This replaces time with an estimate of the mid-point of the period during which there is infectious pressure between any individuals j and k, since this is the only period in which we require the infection rate to be defined. Analysis was performed on data concerning the entirety of the outbreak in Guinea, Sierra Leone and Liberia obtained from the CDC, though we compared this with analysis using Althaus' original data set which did not include the end of the outbreak. This demonstrated the need of complete data for the PBLA method to perform well. When complete data was

used however, PBLA offered very similar estimates to Althaus (2014), despite the considerable approximations and adjustments to the model that were required.

The final application was explored in Section 4.4, and this involved analysis of data from the 2001 UK Foot and Mouth disease epidemic. We compared analysis using PBLA with both MCMC and maximum likelihood estimation to analysis with DA-MCMC in Kypraios (2007). We used the same model as Kypraios, which involved the use of PBLA with a spatial likelihood component in which the infection rate depends on the geographical distance between each pair of farms. We found that both maximum likelihood estimation and MCMC with the PBLA likelihood resulted in similar parameter estimates, which also generally well agreed with those from DA-MCMC.

CHAPTER 5

Conclusions

5.1 Overview

Bayesian statistics has received much attention within the stochastic epidemic modelling community over the past three decades. It allows incorporation of data and prior knowledge into statistical models and, combined with data augmentation techniques, allows for analysis of data which are only partially observed. Although DA-MCMC has become a standard tool for computational analysis of partially observed disease outbreak data, it often suffers from a number of problems.

High posterior correlation between unknown infection times and the infectious period parameters, especially for larger outbreak sizes, can lead to slow mixing of the Markov chain. Current methods for handling this issue have limitations, for example non-centred parameterisations are limited in their applicability, and thinning of the Markov chain, although reducing the autocorrelation between successive samples, increases the number of iterations of the algorithm that need to be implemented to result in the same overall sample size. This can be problematic for likelihoods which are computationally demanding to compute, particularly, for example, for large population sizes.

There is a growing demand for fast, often real-time, analyses, particularly with the increasing use of mathematical modelling by more applied scientists and

public health officials. There is hence considerable motivation to develop computational methods for Bayesian analysis of stochastic epidemic data which overcome these problems. In this thesis, we have explored existing methods for analysis and introduced a series of novel likelihood approximation methods.

5.2 Abakaliki data analysis

Chapter 2 described an analysis of the Abakaliki smallpox data. Despite this data set receiving much attention in the epidemic modelling literature, this is the first Bayesian analysis of the full data to have been completed. We used DA-MCMC to estimate key transmission parameters, which we compared to a previous analysis by Eichner and Dietz (2003). We also performed model assessment using simulation based methods, and our Bayesian approach allowed us to perform novel estimation of the infection pathway. We found that the parameter estimates using DA-MCMC were very similar to those in Eichner and Dietz (2003), despite the fact that they used an approximate likelihood. This motivated the rest of this thesis, in exploring the development of further approximation methods.

The analysis of the Abakaliki data set was largely self-contained, but there are some potential directions for future work. There could be benefit in the use of model assessment tools other than the simulation-based approach taken in Section 2.7, since we cannot be confident that the choice of summary statistics is entirely appropriate. For example, when comparing the incidence curves of simulated outbreaks to the true data, we chose to restrict our attention to only simulations of the same final size as the data, despite the fact that the vast majority of simulations were not of this size. Lau et al. (2014) suggest model diagnostic tools for spatio-temporal transmission models, which apply classical techniques within a Bayesian framework. These use non-centred parameterizations to construct residuals for model assessment, and could potentially

be applied within the Abakaliki model framework. Further work could also involve adaptation of the model used for analysis. Whilst the disease progression model used was relatively representative of smallpox in reality, future work could involve implementation of a more realistic mixing structure for the 30,000 people outside of the affected compounds, rather than assuming homogeneous mixing. Although for our analysis we wanted to use Eichner and Dietz' model to ensure comparability, such steps could be taken if further analysis of the Abakaliki outbreak was performed.

5.3 Likelihood Approximation Methods

The second part of this thesis focused on likelihood approximation methods for infectious disease data. Chapter 3 first explored a generalised version of the Eichner and Dietz approximation method from Chapter 2, and then introduced a series of new approximation methods called PBLA. These essentially assumed independence between the likelihood contributions of different pairs of individuals, in order to obtain a likelihood expression independent of the unknown infection times. These expressions then do not require data augmentation for use with MCMC algorithms, and may also be used for maximum likelihood estimation. We explored various different versions of the PBLA likelihood, aiming to offer e.g. improved accuracy (PBLA II/III) or improved computational speed for certain infectious period distributions (PBLA IV).

After defining these approximation methods in Chapter 3, Chapter 4 gave examples of their application. We compared parameter estimation using the ED approximation, the various PBLA methods and standard MCMC in a series of simulation studies. Here we identified the situations, in terms of infectious period distribution, population sizes and outbreak sizes, in which the methods well approximate DA-MCMC with the true likelihood, as well as comparing the computation time required. The second half of Chapter 4 then described the application of the PBLA method to three real data sets. We compared pa-

rameter estimation with PBLA to existing published analyses, each with specific modelling requirements. We found that, in general, PBLA offered very similar parameter estimation to the previous analyses. This highlights the potential of the method to be used successfully in practice, particularly by applied scientists for whom fast inference is key. Analysis of the computation time required for PBLA compared to standard MCMC revealed that, particularly for increasingly large population sizes, PBLA may offer comparable accuracy in estimation at a substantially lower computational cost. One key hurdle to overcome in the adoption of PBLA by applied scientists might be its limited accessibility compared to currently employed methods such as ABC, which are perhaps easier to interpret. However, there is scope to create a package for implementing PBLA which does not require such an in-depth knowledge of the theory behind it, and we propose that this would help achieve wider adoption of the method.

There is considerable potential for future development of the PBLA method. In Chapter 3 we explored various versions of the PBLA likelihood, but it would certainly be possible to find further versions offering improved estimation. Recalling the true likelihood defined in Section 1.3.5, a key aspect of the current PBLA versions we have explored is that they separate the L_1 term (Equation (1.3.3)) from the L_2 term (Equation (1.3.4)), despite the fact that they both describe the infection component of the likelihood. A future PBLA version might try to combine these terms, in order to obtain a more accurate approximation. Another aspect of the true likelihood is that the product term will be zero for an 'impossible' outbreak, that is an outbreak for which each infectee does not have a potential infector at their time of infection. The PBLA likelihood does not have an explicit equivalent term. This can allow, for exponential infectious periods for example, the estimate of the removal rate γ to become too large. Future PBLA versions might seek to further constrain the infection times in order to avoid this.

The PBLA method may also be extended in terms of its framework. So far

we have focused on exponential and gamma infectious periods, but it may also be possible to calculate the likelihood terms for other distributions. In Section 3.4.11 we discussed the extension of the PBLA method to SEIR models, but found that for non-fixed latent periods this did not result in particularly simple expressions. It may be possible, however, to extend PBLA to SEIR or other compartmental models using slightly different techniques. Lastly, there is potential to extend PBLA for use with non-fixed population size models. These arise in, for example, analysis of hospital infection data where the total population of a ward varies over time (see e.g. Worby et al., 2016).

As discussed, we have compared PBLA to DA-MCMC in Section 4.1.4 and found that for larger population sizes with both exponential and gamma distributed infectious periods, PBLA offers a larger effective sample size per second than DA-MCMC. However, there is scope to improve the efficiency of the PBLA method even further. For example, since PBLA calculates the contribution from different pairs of individuals independently, it is naturally parallelisable. This may increase the computational advantages of the method, at no cost of accuracy. In turn, this could help establish PBLA as a serious alternative to existing approaches such as MCMC and ABC, for ensuring fast and efficicient analysis.

APPENDIX A

Appendix for Abakaliki data analysis: full conditional distributions

In order to implement the MCMC algorithm for analysis of the Abakaliki smallpox data described in Chapter 2, the full conditional distribution (the distribution of a single parameter conditional on all of the others) of each parameter is required. We use the full conditionals since sampling from the full posterior is computationally demanding. The parameters to be updated are t_q, λ_a, λ_f, λ_h, v, b, $\tilde{\mathbf{p}}$, $\mathbf{s}^u$ and the exposure, fever, removal and quarantine times, and we provide the log full conditionals for each here up to proportionality.

In the following, recall $\log(L)$ defined in Equation 2.5.12. We include an additive constant c in the log full conditionals, since the full conditionals are multiplicatively proportional to the expressions shown.

$$\log(\pi(t_q \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, b, v, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})) =$$
$$\log\left(\prod_{j \in \mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt - \log(L) + \log(\pi(t_q)) + c.$$

$$\log(\pi(b \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})) = \log\left(\prod_{j \in \mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt - \log(L) + \log(\pi(b)) + c.$$

$$\log(\pi(v \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, b, \lambda_a, \lambda_f, \lambda_h, \mathbf{s})) = -\log(L) + \sum_{r=0}^{n_{com}-1} p_r \mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}} \log(v) + \sum_{r=0}^{n_{com}-1} (1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}} \log(1-v) + \sum_{\substack{r=n_{com} \\ r \in \mathcal{N}_{n-inf}}}^{N-1} (1-p_r)s_r \log(1-v) + \log(\pi(v)) + c.$$

$$\log(\pi(\lambda_a \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, b, \lambda_f, \lambda_h, \mathbf{s})) = \log\left(\prod_{j \in \mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt - \log(L) + \log(\pi(\lambda_a)) + c.$$

$$\log(\pi(\lambda_f \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, b, \lambda_a, \lambda_h, \mathbf{s})) = \log\left(\prod_{j \in \mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt - \log(L) + \log(\pi(\lambda_f)) + c.$$

$$\log(\pi(\lambda_h \mid \mathbf{r}, \boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}, \kappa, e_\kappa, t_q, v, b, \lambda_f, \lambda_a, \mathbf{s})) = \log\left(\prod_{j \in \mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt + \log(\pi(\lambda_h)) + c.$$

For any $i = 0, 1, ..., n_{com} - 1$, defining $\tilde{\mathbf{p}}_{-i} = (p_0, p_1, ..., p_{i-1}, p_{i+1}, ..., p_{n_{com}-1})$,

$$\log(\pi(\tilde{p}_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \tilde{\mathbf{p}}_{-i}, \mathbf{s}^u)) = -\int_{e_\kappa}^{T} \Lambda_{CN}(t)dt + p_i \log(v) + (1-p_i)\log(1-v) + c.$$

For any element of $\mathbf{s}^u$,

$$\begin{aligned}\log(\pi(s_i^u \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \tilde{\mathbf{p}}, \mathbf{s}^u_{-i})) = \\ -\int_{e_\kappa}^{T} \Lambda_{CN}(t)dt + \sum_{r=0}^{n_{com}-1} p_r \mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}} \log(v) + \\ \sum_{r=0}^{n_{com}-1} (1-p_r)\mathbb{1}_{\{s_r=1 \text{ or } s_r^u=1\}} \log(1-v) + \\ \sum_{\substack{r=n_{com} \\ r\in\mathcal{N}_{inf}}}^{N-1} (1-p_r)s_r \log(1-v) + c,\end{aligned}$$

where $\mathbf{s}^u_{-i} = (s_0^u, s_1^u, ..., s_{i-1}^u, s_{i+1}^u, ..., s_{n_{com}-1}^u)$.

We define $\mathbf{e}_{-i} = (e_0, e_1, ..., e_{i-1}, e_{i+1}, ..., e_{N-1})$ for all infectives $i = 0, 1, ..., N-1$, and similarly for $\mathbf{i}_{-i}$, $\mathbf{q}_{-i}$ and $\boldsymbol{\tau}_{-i}$, then

$$\begin{aligned}\log(\pi(e_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}_{-i}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}, \tilde{\mathbf{p}}, \mathbf{s}^u)) = \\ \log\left(\prod_{j\in\mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt - \log(L) + c.\end{aligned}$$

$$\begin{aligned}\log(\pi(i_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}, \mathbf{i}_{-i}, \mathbf{q}, \boldsymbol{\tau}, \tilde{\mathbf{p}}, \mathbf{s}^u)) = \\ \log\left(\prod_{j\in\mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt - \log(L) + c.\end{aligned}$$

$$\begin{aligned}\log(\pi(q_i \mid \mathbf{r}, \tilde{\boldsymbol{\Phi}}, \mathbf{e}, \mathbf{i}, \mathbf{q}_{-i}, \boldsymbol{\tau}, \tilde{\mathbf{p}}, \mathbf{s}^u)) = \\ \log\left(\prod_{j\in\mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt - \log(L) + c.\end{aligned}$$

$$\begin{aligned}\log(\pi(\tau_i \mid \mathbf{r}, \boldsymbol{\Phi}, \mathbf{e}, \mathbf{i}, \mathbf{q}, \boldsymbol{\tau}_{-i}, \tilde{\mathbf{p}}, \mathbf{s}^u)) = \\ \log\left(\prod_{j\in\mathcal{N}_{inf}} \Lambda_j(e_j-)\right) - \int_{e_\kappa}^{T} \Lambda_{CN}(t) + \Lambda_{CC}(t)dt - \log(L) + c.\end{aligned}$$

We update the exposure times $\mathbf{e}$ as detailed above, but this may in turn alter the exposure time of the initial infective e_κ, or indeed κ itself if a different

individual becomes the initial infective. We do not include the full conditional distribution for this, but in practice we simply update κ and e_κ as required by the **e** update.

APPENDIX B

Appendix for PBLA: likelihood calculations for gamma infectious periods

In this appendix we derive the likelihood expressions for the Pair Based Likelihood Approximation (version I) with Gamma distributed infectious periods, using both integration and probabilistic arguments. We recall that we define gamma distributed infectious periods such that $f_I(r_j - i_j \mid m, \gamma) = \frac{\gamma^m}{\Gamma(m)}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)}$, and we restrict shape parameter m to integer values. As in the exponential case, we will derive expressions for $\mathbb{E}[\chi_j\phi_j]$ and $\mathbb{E}[\psi_j]$, first using the full integration method and then with probabilistic arguments.

B.1 Integration method

Expression one: $\mathbb{E}[\chi_j\phi_j]$

Recall that

$$\mathbb{E}[\chi_j\phi_j] = \sum_{\substack{k=1,\\ k\neq j}}^{n} \beta_{kj}\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}}\mathrm{e}^{-B_j(r_j - i_j)}\right],$$

so that we must calculate the expectation for all pairs j, k. As in the exponential

case,

$$\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j-i_j)}\right] = \int_{-\infty}^{r_k}\int_{-\infty}^{r_j} \mathbb{1}_{i_k<i_j<r_k} \mathrm{e}^{-B_j(r_j-i_j)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_j\ di_k, \quad \text{(B.1.1)}$$

where it has been assumed that i_j and i_k are independent. For a given j and k, this integral will take one of two forms, determined by the values of r_k and r_j.

Case (i): $r_k \geq r_j$

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j-i_j)}\right] \\
&= \int_{-\infty}^{r_j}\int_{i_k}^{r_j} \mathrm{e}^{-B_j(r_j-i_j)} \frac{\gamma^m}{\Gamma(m)}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\frac{\gamma^m}{\Gamma(m)}(r_k-i_k)^{m-1} \\
&\quad \times \mathrm{e}^{-\gamma(r_k-i_k)}\ di_j\ di_k \\
&= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\int_{i_k}^{r_j}\mathrm{e}^{-B_j(r_j-i_j)}(r_j-i_j)^{m-1} \\
&\quad \times \mathrm{e}^{-\gamma(r_j-i_j)}\ di_j\ di_k \\
&= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\int_{0}^{r_j-i_k}\mathrm{e}^{-y(\gamma+B_j)}y^{m-1}\ dy\ di_k,
\end{aligned}$$

where $y = r_j - i_j$. Then the inner integral is proportional to a gamma CDF, which we denote as $F_{k,\theta}(x)$. Since we take only positive integer values for the shape parameter, the distribution function can be expressed as

$$F_{k,\theta}(x) = 1 - \sum_{l=0}^{k-1}\frac{1}{l!}(\theta x)^l \mathrm{e}^{-\theta x}, \quad \text{(B.1.2)}$$

for shape k and rate θ. Substituting this for the integral,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j-i_j)}\right] \\
&= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}F_{m,\gamma+B_j}(r_j-i_k)\frac{\Gamma(m)}{(\gamma+B_j)^m}\ di_k \\
&= \frac{\gamma^{2m}}{\Gamma(m)(\gamma+B_j)^m}\int_{-\infty}^{r_j}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)} \\
&\quad \times \left(1 - \sum_{l=0}^{m-1}\frac{1}{l!}(\gamma+B_j)^l(r_j-i_k)^l\mathrm{e}^{-(\gamma+B_j)(r_j-i_k)}\right)\ di_k,
\end{aligned}$$

which we will split into two integrals.

(i) The first integral is given by

$$\begin{aligned}\int_{-\infty}^{r_j}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\ di_k &= \int_{r_k-r_j}^{\infty} y^{m-1}\mathrm{e}^{-\gamma y}\ dy \\ &= (1-F_{m,\gamma}(r_k-r_j))\frac{\Gamma(m)}{\gamma^m}. \end{aligned} \tag{B.1.3}$$

(ii) The second integral is then equal to

$$\begin{aligned}&\int_{-\infty}^{r_j}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\sum_{l=0}^{m-1}\frac{1}{l!}(\gamma+B_j)^l(r_j-i_k)^l\mathrm{e}^{-(\gamma+B_j)(r_j-i_k)}\ di_k \\ &=\sum_{l=0}^{m-1}\frac{(\gamma+B_j)^l}{l!}\int_{-\infty}^{r_j}\mathrm{e}^{-\gamma(r_k-i_k)}\mathrm{e}^{-(\gamma+B_j)(r_j-i_k)}(r_k-i_k)^{m-1}(r_j-i_k)^l\ di_k \\ &=\sum_{l=0}^{m-1}\frac{(\gamma+B_j)^l}{l!}\mathrm{e}^{-\gamma(r_k-r_j)}\int_0^{\infty}(r_k-r_j+y)^{m-1}y^l\mathrm{e}^{-y(2\gamma+B_j)}\ dy,\end{aligned}$$

where $y = r_j - i_k$. Note that the integral takes the form of the expectation of $(r_k - r_j + Y)^{m-1}$, where Y is gamma distributed with shape $l+1$ and rate $2\gamma + B_j$. This expression is therefore equal to

$$\sum_{l=0}^{m-1}\frac{(\gamma+B_j)^l}{l!}\mathrm{e}^{-\gamma(r_k-r_j)}\frac{\Gamma(l+1)}{(2\gamma+B_j)^{l+1}}\mathbb{E}\big[(r_k-r_j+Y)^{m-1}\mid Y\sim\Gamma(l+1,2\gamma+B_j)\big]. \tag{B.1.4}$$

Combining Equations (B.1.3) and (B.1.4) with the constant term $\frac{\gamma^{2m}}{\Gamma(m)(\gamma+B_j)^m}$, we obtain the result that, for $r_k \geq r_j$,

$$\begin{aligned}&\mathbb{E}\big[\mathbb{1}_{\{k\text{ infective at }i_j\}}\mathrm{e}^{-B_j(r_j-i_j)}\big] = \\ &\qquad\Big(\frac{\gamma}{\gamma+B_j}\Big)^m(1-F_{m,\gamma}(r_k-r_j)) - \sum_{l=0}^{m-1}\frac{\gamma^{2m}}{(2\gamma+B_j)^{l+1}}\frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{(\gamma+B_j)^{m-l}\Gamma(m)} \\ &\qquad\times\mathbb{E}\big[(r_k-r_j+Y)^{m-1}\mid Y\sim\Gamma(l+1,2\gamma+B_j)\big],\end{aligned} \tag{B.1.5}$$

where

$$\mathbb{E}\big[(r+X)^l\mid X\sim\Gamma(m,\gamma)\big] = \sum_{p=0}^{l}\binom{l}{p}r^{l-p}\frac{(m+p-1)_p}{\gamma^p},$$

using the Pochhammer symbol, $(x)_p = \binom{x}{p}p!$, also known as the falling factorial.

Case (ii): $r_k < r_j$

Taking now the case $r_k < r_j$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j-i_j)}\right] \\
&\quad= \int_{-\infty}^{r_k}\int_{i_k}^{r_k} \mathrm{e}^{-B_j(r_j-i_j)} \frac{\gamma^m}{\Gamma(m)}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\frac{\gamma^m}{\Gamma(m)}(r_k-i_k)^{m-1} \\
&\qquad\times \mathrm{e}^{-\gamma(r_k-i_k)}\ di_j\ di_k \\
&\quad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_k}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\int_{i_k}^{r_k}\mathrm{e}^{-(\gamma+B_j)(r_j-i_j)}(r_j-i_j)^{m-1}\ di_j\ di_k \\
&\quad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_k}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\int_{r_j-r_k}^{r_j-i_k}\mathrm{e}^{-(\gamma+B_j)y}y^{m-1}\ dy\ di_k,
\end{aligned}$$

where $y = r_j - i_j$. As in the case $r_k \geq r_j$, this takes the form of a CDF for the gamma distribution, and so using Equation (B.1.2) we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j-i_j)}\right] \\
&\quad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_k}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\frac{\Gamma(m)}{(\gamma+B_j)^m} \\
&\qquad\times\left(F_{m,\gamma+B_j}(r_j-i_k)-F_{m,\gamma+B_j}(r_j-r_k)\right)\ di_k. \qquad \text{(B.1.6)}
\end{aligned}$$

Again, we will split this into two integrals.

(i) The first integral is equal to

$$\begin{aligned}
&\int_{-\infty}^{r_k}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}F_{m,\gamma+B_j}(r_j-i_k)\ di_k \\
&=\int_{-\infty}^{r_k}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\left(1-\sum_{l=0}^{m-1}\frac{(\gamma+B_j)^l}{l!}(r_j-i_k)^l\mathrm{e}^{-(\gamma+B_j)(r_j-i_k)}\right)\ di_k \\
&=\int_{-\infty}^{r_k}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)} \\
&\qquad\qquad-(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\sum_{l=0}^{m-1}\frac{(\gamma+B_j)^l}{l!}(r_j-i_k)^l\mathrm{e}^{-(\gamma+B_j)(r_j-i_k)})\ di_k \\
&=\int_0^\infty \mathrm{e}^{-\gamma y}y^{m-1}\ dy \\
&\qquad\qquad-\sum_{l=0}^{m-1}\frac{(\gamma+B_j)^l}{l!}\int_0^\infty \mathrm{e}^{-\gamma y}\mathrm{e}^{-(\gamma+B_j)(r_j-r_k+y)}y^{m-1}(r_j-r_k+y)^l\ dy,
\end{aligned}$$

where $y = r_k - i_k$. With the first term proportional to a gamma PDF and the

second to an expectation as in the $r_k \geq r_j$ case, the equation reduces to

$$\frac{\Gamma(m)}{\gamma^m} - \sum_{l=0}^{m-1} \frac{(\gamma + B_j)^l}{l!} \frac{\Gamma(m)}{(2\gamma + B_j)^m} \times \mathrm{e}^{-(\gamma+B_j)(r_j - r_k)} \mathbb{E}\big[(r_j - r_k + Y)^l \mid Y \sim \Gamma(m, 2\gamma + B_j)\big].$$

Combining this with the constant term $\frac{\gamma^{2m}}{\Gamma(m)(\gamma+B_j)^m}$ from Equation (B.1.6), we obtain for integral (i):

$$\Big(\frac{\gamma}{\gamma + B}\Big)^m - \frac{\gamma^{2m}}{(\gamma + B_j)^m (2\gamma + B_j)^m} \mathrm{e}^{-(\gamma+B_j)(r_j - r_k)} \times \sum_{l=0}^{m-1} \frac{(\gamma + B_j)^l}{l!} \mathbb{E}\big[(r_j - r_k + Y)^l \mid Y \sim \Gamma(m, 2\gamma + B_j)\big]. \tag{B.1.7}$$

(ii) Moving on to the second integral,

$$\begin{aligned} &\int_{-\infty}^{r_k} (r_k - i_k)^{m-1} \mathrm{e}^{-\gamma(r_k - i_k)} F_{m,\gamma+B_j}(r_j - r_k) \; di_k \\ =&F_{m,\gamma+B_j}(r_j - r_k) \int_0^{\infty} \mathrm{e}^{-\gamma y} y^{m-1} \; di_k \\ =&\frac{\Gamma(m)}{\gamma^m} F_{m,\gamma+B_j}(r_j - r_k). \end{aligned}$$

Combined with the constant term $\frac{\gamma^{2m}}{\Gamma(m)(\gamma+B_j)^m}$ from Equation (B.1.6), the second integral is equal to

$$\frac{\gamma^m}{(\gamma + B_j)^m} F_{m,\gamma+B_j}(r_j - r_k). \tag{B.1.8}$$

Combining Equations (B.1.7) and (B.1.8), the entire expression in the case $r_k < r_j$ is

$$\mathbb{E}\big[\mathbb{1}_{\{k \text{ infective at } i_j\}} \mathrm{e}^{-B_j(r_j - i_j)}\big] = \Big(\frac{\gamma}{\gamma + B_j}\Big)^m \Big(1 - F_{m,\gamma+B_j}(r_j - r_k)\Big) - \Big(\frac{\gamma}{\gamma + B_j}\Big)^m \Big(\frac{\gamma}{2\gamma + B_j}\Big)^m \times \mathrm{e}^{-(\gamma+B_j)(r_j - r_k)} \sum_{l=0}^{m-1} \frac{(\gamma + B_j)^l}{l!} \mathbb{E}\big[(r_j - r_k + Y)^l \mid Y \sim \Gamma(m, 2\gamma + B_j)\big]. \tag{B.1.9}$$

Then, combining Equations (B.1.5) and (B.1.9), we obtain the full expression for $\mathbb{E}[\chi_j\phi_j]$ with gamma distributed infectious periods:

$$\mathbb{E}[\chi_j\phi_j] = \sum_{\substack{k=1,\\k\neq j}}^{n} \beta_{kj} \begin{cases} \left(\frac{\gamma}{\gamma+B_j}\right)^m (1 - F_{m,\gamma}(r_k - r_j)) \\ \qquad - \sum_{l=0}^{m-1} \frac{\gamma^{2m}}{(2\gamma+B_j)^{l+1}} \frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{(\gamma+B_j)^{m-l}\Gamma(m)} \\ \qquad \times \mathbb{E}\left[(r_k - r_j + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma+B_j)\right] \\ \qquad\qquad\qquad\qquad \text{if } r_k \geq r_j, \\ \left(\frac{\gamma}{\gamma+B_j}\right)^m \left(1 - F_{m,\gamma+B_j}(r_j - r_k)\right) \\ \qquad - \left(\frac{\gamma}{\gamma+B_j}\right)^m \left(\frac{\gamma}{2\gamma+B_j}\right)^m \mathrm{e}^{-(\gamma+B_j)(r_j-r_k)} \sum_{l=0}^{m-1} \frac{(\gamma+B_j)^l}{l!} \\ \qquad \times \mathbb{E}\left[(r_j - r_k + Y)^l \mid Y \sim \Gamma(m, 2\gamma+B_j)\right] \\ \qquad\qquad\qquad\qquad \text{if } r_k < r_j, \end{cases}$$

where $B_j = \sum_{l=n+1}^{N} \beta_{jl}$ and

$$\mathbb{E}\left[(r+X)^l \mid X \sim \Gamma(m,\gamma)\right] = \sum_{p=0}^{l} \binom{l}{p} r^{l-p} \frac{(m+p-1)_p}{\gamma^p},$$

with $(x)_p = \binom{x}{p} p!$.

Expression two: $\mathbb{E}[\psi_j]$

As before, we begin with the definition

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1\\k\neq j}}^{n} \mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right].$$

Each term in this product will take the form

$$\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right] \;=\; \int\int \mathrm{e}^{-\beta(r_k \wedge i_j - i_k \wedge i_j)} f_I(r_j - i_j) f_I(r_k - i_k)\; di_j\; di_k,$$

where $\beta_{kj} = \beta$, for simplicity. Again we split this integral into sections, conditional upon the values of r_j and r_k.

Case (i): $r_k \geq r_j$

As in the exponential case,

$$r_k \wedge i_j - i_k \wedge i_j = \begin{cases} i_j - i_k & \text{if } i_k < i_j < r_k, \\ 0 & \text{otherwise.} \end{cases}$$

Then, similarly to before,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta(r_k\wedge i_j - i_k\wedge i_j)} \mid r_k \geq r_j\right] &= \int_{-\infty}^{r_j}\int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\ &\quad + \int_{-\infty}^{r_j}\int_{i_j}^{r_k} 1 \times f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j,\end{aligned}$$

where it has been assumed that i_j and i_k are independent. We have switched the order of integration for the gamma case since it results in simpler calculation. We will calculate each integral individually.

(i) To begin,

$$\begin{aligned}&\int_{-\infty}^{r_j}\int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\ =&\int_{-\infty}^{r_j}\int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j-i_k)} \frac{\gamma^m}{\Gamma(m)}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)} \frac{\gamma^m}{\Gamma(m)}(r_k-i_k)^{m-1} \\ &\times \mathrm{e}^{-\gamma(r_k-i_k)}\ di_k\ di_j \\ =&\frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\int_{-\infty}^{i_j}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\mathrm{e}^{-\beta(i_j-i_k)}\ di_k\ di_j \\ =&\frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\int_{r_k-i_j}^{\infty} y^{m-1}\mathrm{e}^{-\gamma y}\mathrm{e}^{-\beta(i_j-r_k+y)}\ dy\ di_j,\end{aligned}$$

where $y = r_k - i_k$. Then,

$$\begin{aligned}&\int_{-\infty}^{r_j}\int_{-\infty}^{i_j} \mathrm{e}^{-B(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\ =&\frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\mathrm{e}^{-\beta(i_j-r_k)}(1-F_{m,\gamma+\beta}(r_k-i_j))\frac{\Gamma(m)}{(\gamma+\beta)^m}\ di_j \\ =&\frac{\gamma^{2m}}{\Gamma(m)(\gamma+\beta)^m}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\mathrm{e}^{-\beta(i_j-r_k)} \\ &\qquad\times\Big(\sum_{l=0}^{m-1}\frac{1}{l!}\mathrm{e}^{-(\gamma+\beta)(r_k-i_j)}(\gamma+\beta)^l(r_k-i_j)^l\Big)\ di_j \\ =&\frac{\gamma^{2m}}{\Gamma(m)(\gamma+\beta)^m}\sum_{l=0}^{m-1}\frac{(\gamma+\beta)^l}{l!}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}(r_k-i_j)^l \\ &\qquad\times\mathrm{e}^{-\gamma(r_j-i_j)}\mathrm{e}^{-\beta(i_j-r_k)}\mathrm{e}^{-(\gamma+\beta)(r_k-i_j)}\ di_j \\ =&\frac{\gamma^{2m}}{\Gamma(m)(\gamma+\beta)^m}\sum_{l=0}^{m-1}\frac{(\gamma+\beta)^l}{l!}\int_{0}^{\infty} y^{m-1}(r_k-r_j+y)^l\mathrm{e}^{-\gamma y} \\ &\qquad\times\mathrm{e}^{-\beta(r_j-r_k-y)}\mathrm{e}^{-(\gamma+\beta)(r_k-r_j+y)}\ dy,\end{aligned}$$

where $y = r_j - i_j$. Simplifying this integral,

$$\begin{aligned}
&\int_{-\infty}^{r_j}\int_{-\infty}^{i_j} \mathrm{e}^{-B(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\
&\quad= \sum_{l=0}^{m-1} \frac{(\gamma+\beta)^l}{(\gamma+\beta)^m}\frac{\gamma^{2m}}{\Gamma(m)l!}\int_0^{\infty}(r_k-r_j+y)^l y^{m-1}\mathrm{e}^{-2\gamma y}\ dy \\
&\qquad\times \mathrm{e}^{-\beta(r_j-r_k)}\mathrm{e}^{-(\gamma+\beta)(r_k-r_j)} \\
&\quad= \sum_{l=0}^{m-1} \frac{(\gamma+\beta)^l}{(\gamma+\beta)^m}\frac{\gamma^{2m}}{\Gamma(m)l!}\mathrm{e}^{-\gamma(r_k-r_j)}\mathbb{E}\left[(r_k-r_j+Y)^l \mid Y\sim\Gamma(m,2\gamma)\right]\frac{\Gamma(m)}{(2\gamma)^m},
\end{aligned}$$

since the integral was proportional to the expectation of $(r_k - r_j + y)^l$ for a $\Gamma(m, 2\gamma)$ distributed random variable y. Overall, we find that

$$\begin{aligned}
&\int_{-\infty}^{r_j}\int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\
&\quad= \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{2^m l!}\left(\frac{\gamma}{\gamma+\beta}\right)^m(\gamma+\beta)^l\mathbb{E}\left[(r_k-r_j+Y)^l \mid Y\sim\Gamma(m,2\gamma)\right].
\end{aligned} \tag{B.1.10}$$

(ii) The second integral is equal to

$$\begin{aligned}
&\int_{-\infty}^{r_j}\int_{i_j}^{r_k} 1\times f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\
&\quad= \int_{-\infty}^{r_j}\int_{i_j}^{r_k} \frac{\gamma^m}{\Gamma(m)}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\frac{\gamma^m}{\Gamma(m)}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\ di_k\ di_j \\
&\quad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\int_{i_j}^{r_k}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\ di_k\ di_j \\
&\quad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\int_0^{r_k-i_j} y^{m-1}\mathrm{e}^{-\gamma y}\ dy\ di_j,
\end{aligned}$$

where $y = r_k - i_k$. Then,

$$\begin{aligned}
&\int_{-\infty}^{r_j}\int_{i_j}^{r_k} 1\times f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\
&\qquad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\frac{\Gamma(m)}{\gamma^m}F_{m,\gamma}(r_k-i_j)\ di_j,
\end{aligned}$$

by the definition of a gamma CDF. Using Equation (B.1.2),

$$\begin{aligned}
&\int_{-\infty}^{r_j}\int_{i_j}^{r_k} 1\times f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\
&\qquad= \frac{\gamma^m}{\Gamma(m)}\int_{-\infty}^{r_j}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\left(1-\sum_{l=0}^{m-1}\frac{\gamma^l}{l!}\mathrm{e}^{-\gamma(r_k-i_j)}(r_k-i_j)^l\right)\ di_j.
\end{aligned}$$

(ii.1) Splitting this into two integrals, we first take $y = r_j - i_j$, so that

$$\begin{aligned}
&\frac{\gamma^m}{\Gamma(m)} \int_{-\infty}^{r_j} (r_j - i_j)^{m-1} \mathrm{e}^{-\gamma(r_j - i_j)} \; di_j \\
=&\frac{\gamma^m}{\Gamma(m)} \int_0^{\infty} y^{m-1} \mathrm{e}^{-\gamma y} \; dy \\
=&\frac{\gamma^m}{\Gamma(m)} \frac{\Gamma(m)}{\gamma^m} \\
=&1. \qquad \text{(B.1.11)}
\end{aligned}$$

(ii.2) Taking the second integral,

$$\begin{aligned}
&\frac{\gamma^m}{\Gamma(m)} \int_{-\infty}^{r_j} (r_j - i_j)^{m-1} \mathrm{e}^{-\gamma(r_j - i_j)} \sum_{l=0}^{m-1} \frac{\gamma^l}{l!} \mathrm{e}^{-\gamma(r_k - i_j)} (r_k - i_j)^l \; di_j \\
=&\frac{\gamma^m}{\Gamma(m)} \sum_{l=0}^{m-1} \frac{\gamma^l}{l!} \int_{-\infty}^{r_j} (r_j - i_j)^{m-1} (r_k - i_j)^l \mathrm{e}^{-\gamma(r_j - i_j)} \mathrm{e}^{-\gamma(r_k - i_j)} \; di_j \\
=&\frac{\gamma^m}{\Gamma(m)} \sum_{l=0}^{m-1} \frac{\gamma^l}{l!} \int_0^{\infty} y^{m-1} (r_k - r_j + y)^l \mathrm{e}^{-\gamma(r_k - r_j + 2y)} \; dy
\end{aligned}$$

where $y = r_j - i_j$. Then note, as in the calculations for $\mathbb{E}[\chi_j \phi_j]$, that the integral is proportional to the expectation of $(r_k - r_j + y)^l$, where y is gamma distributed. Therefore,

$$\begin{aligned}
&\frac{\gamma^m}{\Gamma(m)} \sum_{l=0}^{m-1} \frac{\gamma^l}{l!} \int_0^{\infty} y^{m-1} (r_k - r_j + y)^l \mathrm{e}^{-\gamma(r_k - r_j + 2y)} \; dy \\
=&\frac{\gamma^m}{\Gamma(m)} \sum_{l=0}^{m-1} \frac{\gamma^l}{l!} \mathrm{e}^{-\gamma(r_k - r_j)} \mathbb{E}\left[((r_k - r_j + Y)^l) \mid Y \sim \Gamma(m, 2\gamma) \right] \frac{\Gamma(m)}{(2\gamma)^m} \\
=&\sum_{l=0}^{m-1} \frac{\gamma^l}{l! 2^m} \mathrm{e}^{-\gamma(r_k - r_j)} \mathbb{E}\left[((r_k - r_j + Y)^l) \mid Y \sim \Gamma(m, 2\gamma) \right]. \qquad \text{(B.1.12)}
\end{aligned}$$

Combining Equations (B.1.11) and (B.1.12), the second part of $\mathbb{E}[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}]$ is given by

$$\begin{aligned}
&\int_{-\infty}^{r_j} \int_{i_j}^{r_k} 1 \times f_I(r_j - i_j) f_I(r_k - i_k) \; di_k \; di_j \\
&= 1 - \sum_{l=0}^{m-1} \frac{\gamma^l}{l! 2^m} \mathrm{e}^{-\gamma(r_k - r_j)} \mathbb{E}\left[((r_k - r_j + Y)^l) \mid Y \sim \Gamma(m, 2\gamma) \right]. \qquad \text{(B.1.13)}
\end{aligned}$$

Overall, adding together Equations (B.1.10) and (B.1.13) we obtain

$$\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k\wedge i_j - i_k\wedge i_j)} \mid r_k \geq r_j\right] = 1 + \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\gamma(r_k - r_j)}}{l!2^m} \times \mathbb{E}\left[((r_k - r_j + \Upsilon)^l) \mid \Upsilon \sim \Gamma(m, 2\gamma)\right] \left(\left(\frac{\gamma}{\gamma+\beta_{kj}}\right)^m (\gamma+\beta_{kj})^l - \gamma^l \right). \tag{B.1.14}$$

Case (ii): $r_k < r_j$

In this case,

$$r_k \wedge i_j - i_k \wedge i_j = \begin{cases} r_k - i_k & \text{if } r_k < i_j, \\ i_j - i_k & \text{if } i_k < i_j < r_k, \\ 0 & \text{otherwise.} \end{cases}$$

Then as before,

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\beta(r_k\wedge i_j - i_k\wedge i_j)} \mid r_k < r_j\right] &= \int_{r_k}^{r_j}\int_{-\infty}^{r_k} \mathrm{e}^{-\beta(r_k - i_k)} f_I(r_j - i_j) f_I(r_k - i_k)\ di_k\ di_j \\ &\quad + \int_{-\infty}^{r_k}\int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j - i_k)} f_I(r_j - i_j) f_I(r_k - i_k)\ di_k\ di_j \\ &\quad + \int_{-\infty}^{r_k}\int_{i_j}^{\infty} 1 \times f_I(r_j - i_j) f_I(r_k - i_k)\ di_k\ di_j, \end{aligned}$$

where i_j and i_k are assumed independent, and we will calculate each integral individually.

(i) Firstly,

$$\begin{aligned} &\int_{r_k}^{r_j}\int_{-\infty}^{r_k} \mathrm{e}^{-\beta(r_k - i_k)} f_I(r_j - i_j) f_I(r_k - i_k)\ di_k\ di_j \\ &= \int_{r_k}^{r_j}\int_{-\infty}^{r_k} \mathrm{e}^{-\beta(r_k - i_k)} \frac{\gamma^m}{\Gamma(m)}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)} \frac{\gamma^m}{\Gamma(m)}(r_k - i_k)^{m-1} \\ &\quad \times \mathrm{e}^{-\gamma(r_k - i_k)}\ di_k\ di_j \\ &= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{r_k}^{r_j}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)} \int_{-\infty}^{r_k}(r_k - i_k)^{m-1}\mathrm{e}^{-\gamma(r_k - i_k)} \\ &\quad \times \mathrm{e}^{-\beta(r_k - i_k)}\ di_k\ di_j \\ &= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{r_k}^{r_j}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)} \int_{0}^{\infty} y^{m-1}\mathrm{e}^{-(\gamma+\beta)y}\ dy\ di_j, \end{aligned}$$

where $y = r_k - i_k$. Continuing,

$$\int_{r_k}^{r_j} \int_{-\infty}^{r_k} \mathrm{e}^{-\beta(r_k - i_k)} f_I(r_j - i_j) f_I(r_k - i_k) \; di_k \; di_j$$
$$= \frac{\gamma^{2m}}{\Gamma(m)^2} \int_{r_k}^{r_j} (r_j - i_j)^{m-1} \mathrm{e}^{-\gamma(r_j - i_j)} \frac{\Gamma(m)}{(\gamma+\beta)^m} \; di_j,$$

since the integral is proportional to the PDF of a gamma distribution. Then

$$\int_{r_k}^{r_j} \int_{-\infty}^{r_k} \mathrm{e}^{-\beta(r_k - i_k)} f_I(r_j - i_j) f_I(r_k - i_k) \; di_k \; di_j$$
$$= \frac{\gamma^{2m}}{(\gamma+\beta)^m \Gamma(m)} \int_{r_k}^{r_j} (r_j - i_j)^{m-1} \mathrm{e}^{-\gamma(r_j - i_j)} \; di_j$$
$$= \frac{\gamma^{2m}}{(\gamma+\beta)^m \Gamma(m)} \int_{0}^{r_j - r_k} y^{m-1} \mathrm{e}^{-\gamma y} \; dy,$$

where $y = r_j - i_j$. Finally,

$$\int_{r_k}^{r_j} \int_{-\infty}^{r_k} \mathrm{e}^{-\beta(r_k - i_k)} f_I(r_j - i_j) f_I(r_k - i_k) \; di_k \; di_j$$
$$= \frac{\gamma^{2m}}{(\gamma+\beta)^m \Gamma(m)} F_{m,\gamma}(r_j - r_k) \frac{\Gamma(m)}{\gamma^m}$$
$$= \left(\frac{\gamma}{\gamma+\beta}\right)^m F_{m,\gamma}(r_j - r_k), \tag{B.1.15}$$

where we have used the definition of the gamma CDF from Equation (B.1.2).

(ii) The second integral is equal to

$$\int_{-\infty}^{r_k} \int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j - i_k)} f_I(r_j - i_j) f_I(r_k - i_k) \; di_k \; di_j$$
$$= \int_{-\infty}^{r_k} \int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j - i_k)} \frac{\gamma^m}{\Gamma(m)} (r_j - i_j)^{m-1} \mathrm{e}^{-\gamma(r_j - i_j)} \frac{\gamma^m}{\Gamma(m)} (r_k - i_k)^{m-1}$$
$$\times \mathrm{e}^{-\gamma(r_k - i_k)} \; di_k \; di_j$$
$$= \frac{\gamma^{2m}}{\Gamma(m)^2} \int_{-\infty}^{r_k} (r_j - i_j)^{m-1} \mathrm{e}^{-\gamma(r_j - i_j)} \int_{-\infty}^{i_j} (r_k - i_k)^{m-1} \mathrm{e}^{-\gamma(r_k - i_k)}$$
$$\times \mathrm{e}^{-\beta(i_j - i_k)} \; di_k \; di_j$$
$$= \frac{\gamma^{2m}}{\Gamma(m)^2} \int_{-\infty}^{r_k} (r_j - i_j)^{m-1} \mathrm{e}^{-\gamma(r_j - i_j)} \int_{r_k - i_j}^{\infty} y^{m-1} \mathrm{e}^{-\gamma y} \mathrm{e}^{-\beta(i_j - r_k + y)} \; dy \; di_j,$$

where $y = r_k - i_k$. Then,

$$\begin{aligned}
&\int_{-\infty}^{r_k}\int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\
&\quad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_k} (r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\mathrm{e}^{-\beta(i_j-r_k)}(1-F_{m,\gamma+\beta}(r_k-i_j)) \\
&\qquad\times \frac{\Gamma(m)}{(\gamma+\beta)^m}\ di_j \\
&\quad= \frac{\gamma^{2m}}{(\gamma+\beta)^m\Gamma(m)}\int_{-\infty}^{r_k} (r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\mathrm{e}^{-\beta(i_j-r_k)} \\
&\qquad\times \sum_{l=0}^{m-1}\frac{1}{l!}\mathrm{e}^{-(\gamma+\beta)(r_k-i_j)}(\gamma+\beta)^l(r_k-i_j)^l\ di_j \\
&\quad= \frac{\gamma^{2m}}{(\gamma+\beta)^m\Gamma(m)}\sum_{l=0}^{m-1}\frac{(\gamma+\beta)^l}{l!}\int_{-\infty}^{r_k}(r_j-i_j)^{m-1}(r_k-i_j)^l\mathrm{e}^{-\gamma(r_j+r_k-2i_j)}\ di_j \\
&\quad= \frac{\gamma^{2m}}{(\gamma+\beta)^m\Gamma(m)}\sum_{l=0}^{m-1}\frac{(\gamma+\beta)^l}{l!}\int_{0}^{\infty}(r_j-r_k+y)^{m-1}y^l\mathrm{e}^{-\gamma(r_j-r_k+2y)}\ dy,
\end{aligned}$$

where $y = r_k - i_j$. The integral is proportional to the expectation of $(r_j - r_k + y)^{m-1}$, where y is gamma distributed, and so we obtain

$$\begin{aligned}
&\int_{-\infty}^{r_k}\int_{-\infty}^{i_j} \mathrm{e}^{-\beta(i_j-i_k)} f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\
&\quad= \frac{\gamma^{2m}}{(\gamma+\beta)^m\Gamma(m)}\sum_{l=0}^{m-1}\frac{(\gamma+\beta)^l}{l!}\mathrm{e}^{-\gamma(r_j-r_k)} \\
&\qquad\times \mathbb{E}\left[(r_j-r_k+Y)^{m-1}\mid Y\sim\Gamma(l+1,2\gamma)\right]\frac{\Gamma(l+1)}{(2\gamma)^{l+1}} \\
&\quad= \sum_{l=0}^{m-1}\frac{\mathrm{e}^{-\gamma(r_j-r_k)}}{2^{l+1}}\frac{\gamma^{m-1}}{\Gamma(m)}\mathbb{E}[(r_j-r_k+Y)^{m-1}\mid Y\sim\Gamma(l+1,2\gamma)] \\
&\qquad\times\left(\frac{\gamma}{\gamma+\beta}\right)^m\left(\frac{\gamma+\beta}{\gamma}\right)^l. \qquad \text{(B.1.16)}
\end{aligned}$$

(iii) Moving on to the third integral,

$$\begin{aligned}
&\int_{-\infty}^{r_k}\int_{i_j}^{\infty} 1\times f_I(r_j-i_j) f_I(r_k-i_k)\ di_k\ di_j \\
&\quad= \int_{-\infty}^{r_k}\int_{i_j}^{\infty}\frac{\gamma^m}{\Gamma(m)}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\frac{\gamma^m}{\Gamma(m)}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\ di_k\ di_j \\
&\quad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_k}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\int_{i_j}^{\infty}(r_k-i_k)^{m-1}\mathrm{e}^{-\gamma(r_k-i_k)}\ di_k\ di_j \\
&\quad= \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_k}(r_j-i_j)^{m-1}\mathrm{e}^{-\gamma(r_j-i_j)}\int_{-\infty}^{r_k-i_j} y^{m-1}\mathrm{e}^{-\gamma y}\ dy\ di_j,
\end{aligned}$$

where $y = r_k - i_k$. Then, using the fact the integral is proportional to a gamma CDF and Equation (B.1.2),

$$\begin{aligned}
&\int_{-\infty}^{r_k}\int_{i_j}^{\infty} 1 \times f_I(r_j - i_j) f_I(r_k - i_k)\ di_k\ di_j \\
&\quad = \frac{\gamma^{2m}}{\Gamma(m)^2}\int_{-\infty}^{r_k}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)}F_{m,\gamma}(r_k - i_j)\frac{\Gamma(m)}{\gamma^m}\ di_j \\
&\quad = \frac{\gamma^m}{\Gamma(m)}\int_{-\infty}^{r_k}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)}\Big(1 - \sum_{l=0}^{m-1}\frac{1}{l!}\mathrm{e}^{-\gamma(r_k - i_j)}\gamma^l(r_k - i_j)^l\Big)\ di_j.
\end{aligned}$$

We will divide this into two integrals.

(iii.1) Firstly, setting $y = r_j - i_j$,

$$\begin{aligned}
\frac{\gamma^m}{\Gamma(m)}\int_{-\infty}^{r_k}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)}\ di_j &= \frac{\gamma^m}{\Gamma(m)}\int_{r_j - r_k}^{\infty} y^{m-1}\mathrm{e}^{-\gamma y}\ dy \\
&= 1 - F_{m,\gamma}(r_j - r_k).
\end{aligned}$$

(iii.2) Secondly,

$$\begin{aligned}
&\frac{\gamma^m}{\Gamma(m)}\int_{-\infty}^{r_k}(r_j - i_j)^{m-1}\mathrm{e}^{-\gamma(r_j - i_j)}\sum_{l=0}^{m-1}\frac{1}{l!}\mathrm{e}^{-\gamma(r_k - i_j)}\gamma^l(r_k - i_j)^l\ di_j \\
&\quad = \frac{\gamma^m}{\Gamma(m)}\sum_{l=0}^{m-1}\frac{\gamma^l}{l!}\int_{-\infty}^{r_k}(r_j - i_j)^{m-1}(r_k - i_j)^l\mathrm{e}^{-\gamma(r_j - i_j)}\mathrm{e}^{-\gamma(r_k - i_j)}\ di_j \\
&\quad = \frac{\gamma^m}{\Gamma(m)}\sum_{l=0}^{m-1}\frac{\gamma^l}{l!}\int_{0}^{\infty}(r_j - r_k + y)^{m-1}y^l\mathrm{e}^{-\gamma(r_j - r_k + 2y)}\ dy,
\end{aligned}$$

where $y = r_k - i_j$. Noting that this integral is proportional to the expectation of $(r_j - r_k + y)^{m-1}$, where y is gamma distributed, the expression becomes

$$\begin{aligned}
&= \frac{\gamma^m}{\Gamma(m)}\sum_{l=0}^{m-1}\frac{\gamma^l}{l!}\mathrm{e}^{-\gamma(r_j - r_k)}\int_{0}^{\infty}(r_j - r_k + y)^{m-1}y^l\mathrm{e}^{-2\gamma y}\ dy \\
&= \frac{\gamma^m}{\Gamma(m)}\sum_{l=0}^{m-1}\frac{\gamma^l}{l!}\mathrm{e}^{-\gamma(r_j - r_k)}\mathbb{E}\big[(r_j - r_k + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma)\big] \\
&\quad \times \frac{\Gamma(l+1)}{(2\gamma)^{l+1}} \\
&= \sum_{l=0}^{m-1}\frac{\gamma^{m-1}\mathrm{e}^{-\gamma(r_j - r_k)}}{2^{l+1}\Gamma(m)}\mathbb{E}\big[(r_j - r_k + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma)\big].
\end{aligned}$$

Recombining the two integrals, we obtain overall for integral (iii),

$$\begin{aligned}\int_{-\infty}^{r_k}\int_{i_j}^{\infty} 1 \times f_I(r_j - i_j) f_I(r_k - i_k)\ di_k\ di_j \\ = 1 - F_{m,\gamma}(r_j - r_k) - \sum_{l=0}^{m-1} \frac{\gamma^{m-1}\mathrm{e}^{-\gamma(r_j - r_k)}}{2^{l+1}\Gamma(m)} \\ \times \mathbb{E}\big[(r_j - r_k + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma)\big]. \end{aligned} \tag{B.1.17}$$

We combine Equations (B.1.15), (B.1.16) and (B.1.17) to obtain

$$\begin{aligned}&\mathbb{E}\big[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)} \mid r_k < r_j\big] \\ &\quad = 1 - F_{m,\gamma}(r_j - r_k)\Big(1 - \big(\frac{\gamma}{\gamma+\beta_{kj}}\big)^m\Big) + \sum_{l=0}^{m-1} \frac{\gamma^{m-1}\mathrm{e}^{-\gamma(r_j - r_k)}}{2^{l+1}\Gamma(m)} \\ &\quad \times \mathbb{E}\big[(r_j - r_k + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma)\big]\left(\Big(\frac{\gamma}{\gamma+\beta_{kj}}\Big)^m\Big(\frac{\gamma+\beta_{kj}}{\gamma}\Big)^l - 1\right).\end{aligned} \tag{B.1.18}$$

Finally, combining Equations (B.1.14) and (B.1.18), we obtain the full expression for $\mathbb{E}[\psi_j]$ with gamma distributed infectious periods:

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1\\k\neq j}}^{n} \begin{cases} 1 + \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\gamma(r_k - r_j)}}{l!2^m}\mathbb{E}\big[(r_k - r_j + Y)^l \mid Y \sim \Gamma(m, 2\gamma)\big] \\ \qquad \times\left(\Big(\frac{\gamma}{\gamma+\beta_{kj}}\Big)^m(\gamma+\beta_{kj})^l - \gamma^l\right) & \text{if } r_k \geq r_j, \\ 1 - F_{m,\gamma}(r_j - r_k)\Big(1 - \big(\frac{\gamma}{\gamma+\beta_{kj}}\big)^m\Big) + \sum_{l=0}^{m-1} \frac{\gamma^{m-1}\mathrm{e}^{-\gamma(r_j - r_k)}}{2^{l+1}\Gamma(m)} \\ \qquad \times\mathbb{E}\big[(r_j - r_k + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma)\big] \\ \qquad \times\Big(\big(\frac{\gamma}{\gamma+\beta_{kj}}\big)^m\big(\frac{\gamma+\beta_{kj}}{\gamma}\big)^l - 1\Big) & \text{if } r_k < r_j, \end{cases} \tag{B.1.19}$$

with

$$\mathbb{E}\big[(r + X)^l \mid X \sim \Gamma(m, \gamma)\big] = \sum_{p=0}^{l} \binom{l}{p} r^{l-p} \frac{(m+p-1)_p}{\gamma^p}.$$

and $(x)_p = \binom{x}{p} p!$.

Again, we may now calculate the likelihood for any given choice of prior probability mass function for the initial infective and prior probability density/mass function for the initial infective's infection time. In summary,

$$\pi(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \Big(\prod_{j=1}^{n} \mathbb{E}[\chi_j \phi_j] \mathbb{E}[\psi_j]\Big) \sum_{\kappa=1}^{n} \frac{\pi(\kappa)\mathbb{E}[\phi_\kappa \pi(i_\kappa \mid \kappa)]}{\mathbb{E}[\chi_\kappa \phi_\kappa]\mathbb{E}[\psi_\kappa]},$$

where

$$\mathbb{E}[\chi_j \phi_j] = \sum_{\substack{k=1,\\ k\neq j}}^{n} \beta_{kj} \begin{cases} \Big(\frac{\gamma}{\gamma+B_j}\Big)^m (1 - F_{m,\gamma}(r_k - r_j)) \\ \qquad - \sum_{l=0}^{m-1} \frac{\gamma^{2m}}{(2\gamma+B_j)^{l+1}} \frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{(\gamma+B_j)^{m-l}\Gamma(m)} \\ \qquad \times \mathbb{E}\big[(r_k - r_j + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma + B_j)\big] \\ \qquad\qquad \text{if } r_k \geq r_j, \\ \Big(\frac{\gamma}{\gamma+B_j}\Big)^m \Big(1 - F_{m,\gamma+B_j}(r_j - r_k)\Big) \\ \qquad - \Big(\frac{\gamma}{\gamma+B_j}\Big)^m \Big(\frac{\gamma}{2\gamma+B_j}\Big)^m \mathrm{e}^{-(\gamma+B_j)(r_j-r_k)} \sum_{l=0}^{m-1} \frac{(\gamma+B_j)^l}{l!} \\ \qquad \times \mathbb{E}\big[(r_j - r_k + Y)^l \mid Y \sim \Gamma(m, 2\gamma + B_j)\big] \\ \qquad\qquad \text{if } r_k < r_j, \end{cases}$$

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1\\ k\neq j}}^{n} \begin{cases} 1 + \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{l!2^m} \mathbb{E}\big[(r_k - r_j + Y)^l \mid Y \sim \Gamma(m, 2\gamma)\big] \\ \qquad \times \left(\Big(\frac{\gamma}{\gamma+\beta_{kj}}\Big)^m (\gamma + \beta_{kj})^l - \gamma^l \right) \\ \qquad\qquad \text{if } r_k \geq r_j, \\ 1 - F_{m,\gamma}(r_j - r_k)\Big(1 - \big(\frac{\gamma}{\gamma+\beta_{kj}}\big)^m\Big) + \sum_{l=0}^{m-1} \frac{\gamma^{m-1}\mathrm{e}^{-\gamma(r_j-r_k)}}{2^{l+1}\Gamma(m)} \\ \qquad \times \mathbb{E}\big[(r_j - r_k + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma)\big] \\ \qquad \times \Big(\big(\frac{\gamma}{\gamma+\beta_{kj}}\big)^m \big(\frac{\gamma+\beta_{kj}}{\gamma}\big)^l - 1 \Big) \\ \qquad\qquad \text{if } r_k < r_j. \end{cases}$$

Recall that $B_j = \sum_{l=n+1}^{N} \beta_{jl}$,

$$F_{k,\theta}(x) = 1 - \sum_{l=0}^{k-1} \frac{1}{l!} (\theta x)^l \mathrm{e}^{-\theta x},$$

and

$$\mathbb{E}\big[(r + X)^l \mid X \sim \Gamma(m, \gamma)\big] = \sum_{p=0}^{l} \binom{l}{p} r^{l-p} \frac{(m+p-1)_p}{\gamma^p},$$

with $(x)_p = \binom{x}{p} p!$.

B.2 Probabilistic method

We are able to obtain the same likelihood expression using probabilistic arguments. In the case of exponential infectious periods, we argued using the fact that the probability of an event occurring is independent of time. Here we use the method of stages to split a $\Gamma(m, \gamma)$ time period into m exponentially distributed sections, so that similar arguments may be used.

Expression one: $\mathbb{E}[\chi_j \phi_j]$

Consider first $\mathbb{E}[\chi_j \phi_j]$. We must, for any given infectives j and k, calculate $\mathbb{E}\left[e^{-B_j(r_j - i_j)} \mathbb{1}_{\{i_k < i_j < r_k\}}\right]$, where $B_j = \sum_{l=n+1}^{N} \beta_{jl}$. We split this into cases depending on which is greater, r_k or r_j.

Case (i): $r_k \geq r_j$

We begin by noting that, given $i_j < r_j < r_k$, there are 3 possible locations for i_k and hence

$$\mathbb{1}_{\{r_j < i_k < r_k\}} + \mathbb{1}_{\{i_j < i_k < r_j\}} + \mathbb{1}_{\{i_k < i_j < r_k\}} = 1.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[e^{-B_j(r_j - i_j)} \mathbb{1}_{\{i_k < i_j < r_k\}}\right] &= \mathbb{E}\left[e^{-B_j(r_j - i_j)} (1 - \mathbb{1}_{\{r_j < i_k < r_k\}} - \mathbb{1}_{\{i_j < i_k < r_j\}})\right] \\
&= \mathbb{E}\left[e^{-B_j(r_j - i_j)}\right] - \mathbb{E}\left[e^{-B_j(r_j - i_j)} \mathbb{1}_{\{r_j < i_k < r_k\}}\right] \\
&\quad - \mathbb{E}\left[e^{-B_j(r_j - i_j)} \mathbb{1}_{\{i_j < i_k < r_j\}}\right] \\
&= \left(\frac{\gamma}{\gamma + B_j}\right)^m - \left(\frac{\gamma}{\gamma + B_j}\right)^m F_{m, \gamma + B_j}(r_k - r_j) \\
&\quad - \mathbb{E}\left[e^{-B_j(r_j - i_j)} \mathbb{1}_{\{i_j < i_k < r_j\}}\right]. \qquad \text{(B.2.1)}
\end{aligned}$$

This is since the first term takes the form of a Gamma moment generating function. The second term represents the probability of there being no points in a Poisson process of rate B_j between i_j and r_j as well as m exponential γ stages for individual k in (r_j, r_k) (given by the indicator function). $F_{m,\gamma}(\cdot)$ represents a gamma CDF with parameters m and γ. Although not written for simplicity, all the terms here are of course conditional on $r_k \geq r_j$

What remains is to calculate the final expectation term in the expression above,

namely

$$\mathbb{E}\left[\mathrm{e}^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_j<i_k<r_j\}} \mid r_k \geq r_j\right].$$

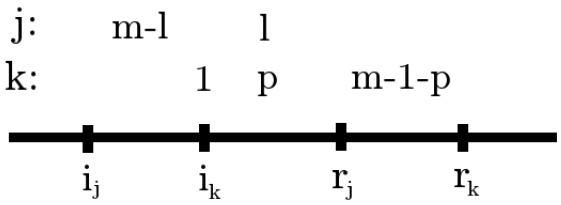


Figure B.1: For $r_k \geq r_j$.

We proceed as in the exponential case by working backwards in time from the final event r_k. We condition on the number of $\Gamma(m, \gamma)$ stages for individual k in each time period. Specifically, we assume $m-1-p$ stages in the period (r_j, r_k), p in the period (i_k, r_j) and one final stage exactly at i_k. Clearly $p \in (0, ..., m-1)$. This is shown in Figure B.1. Now,

$$\mathbb{P}(m-1-p \text{ stages in } (r_j, r_k)) = \mathrm{e}^{-\gamma(r_k-r_j)}\frac{\gamma^{m-1-p}}{(m-1-p)!}(r_k-r_j)^{m-1-p},$$

since the number of stages that occur follows a Poisson distribution with parameter γ.

Next, moving in reverse time from r_j, we must also consider $\Gamma(m, \gamma)$ stages for individual j. Since a total of m of these must occur between i_j and r_j, we split them such that l occur in (i_k, r_j) and the remaining $m-l$ occur in (i_j, i_k), where $l \in (0, ..., m-1)$.

Therefore, for the time period (i_k, r_j) we must allocate p stages for individual k and l stages for individual j. The number of ways to allocate the p stages among a total of $l+p$ is simply given by $\binom{l+p}{p}$. For time period (i_j, i_k), only stages for j continue to occur and hence we do not need to order these.

The last part to consider is the $\mathrm{e}^{-B_j(r_j-i_j)}$ term in the expectation. Note that this is equal to the probability of there being no points in a Poisson process of rate B_j in (i_j, r_j), and so we must condition on all stages of our infectious

processes for j and k occurring before a point in this B_j process. We require $l+p+1$ stages in (i_k, r_j) (including the stage for k which occurs exactly at i_k). Since there is an infectious process for k occurring with rate γ and another for j with rate γ, which are independent, we may sum them to a Poisson process occurring with rate 2γ. Hence, the probability of $l+p+1$ stages occurring before a Poisson process of rate B_j is given by $\left(\frac{\gamma}{2\gamma+B_j}\right)^{l+p+1}$. Similarly, in (i_j, i_k) we require $m-l$ stages for j before any events in the Poisson process of rate B_j (events for k are no longer occurring), which is given by $\left(\frac{\gamma}{\gamma+B_j}\right)^{m-l}$.

Combining these arguments, we find

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_j<i_k<r_j\}} \mid r_k \geq r_j\right] &= \sum_{l=0}^{m-1}\sum_{p=0}^{m-1} \mathrm{e}^{-\gamma(r_k-r_j)}\frac{\gamma^{m-1-p}}{(m-1-p)!} \\ &\quad \times (r_k-r_j)^{m-1-p}\binom{l+p}{p}\left(\frac{\gamma}{2\gamma+B_j}\right)^{l+p+1} \\ &\quad \times \left(\frac{\gamma}{\gamma+B_j}\right)^{m-l},\end{aligned}$$

which can be rearranged to

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_j<i_k<r_j\}} \mid r_k \geq r_j\right] &= \sum_{l=0}^{m-1} \frac{\gamma^{2m}}{(2\gamma+B_j)^{l+1}}\frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{(\gamma+B_j)^{m-l}\Gamma(m)} \\ &\quad \times \sum_{p=0}^{m-1}\binom{m-1}{p}(r_k-r_j)^{m-1-p}\frac{(l+p)_p}{(2\gamma+B_j)^p}.\end{aligned}$$

Putting this back together with the rest of expression from Equation (B.2.1), we obtain

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_k<i_j<r_k\}} \mid r_k \geq r_j\right] &= \left(\frac{\gamma}{\gamma+B_j}\right)^m (1-F_{m,\gamma+B_j}(r_k-r_j)) \\ &\quad - \sum_{l=0}^{m-1}\frac{\gamma^{2m}}{(2\gamma+B_j)^{l+1}}\frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{(\gamma+B_j)^{m-l}\Gamma(m)} \\ &\quad \times \sum_{p=0}^{m-1}\binom{m-1}{p}(r_k-r_j)^{m-1-p} \\ &\quad \frac{(l+p)_p}{(2\gamma+B_j)^p},\end{aligned} \tag{B.2.2}$$

which is the same expression found via direct integration methods in Equation (3.4.12).

Case (ii): $r_k < r_j$

Moving on to the case $r_k < r_j$, we again calculate $\mathbb{E}\left[e^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_k<i_j<r_k\}}\right]$. The arguments required are similar here, for example we now consider the possible locations of i_j rather than i_k:

$$\mathbb{1}_{\{i_k<i_j<r_k\}} + \mathbb{1}_{\{i_j<i_k<r_k\}} + \mathbb{1}_{\{r_k<i_j<r_j\}} = 1.$$

Then, as before with all terms conditional on $r_k < r_j$,

$$\begin{aligned}
\mathbb{E}\left[e^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_k<i_j<r_k\}}\right] &= \mathbb{E}\left[e^{-B_j(r_j-i_j)}(1-\mathbb{1}_{\{r_k<i_j<r_j\}}-\mathbb{1}_{\{i_j<i_k<r_k\}})\right] \\
&= \mathbb{E}\left[e^{-B_j(r_j-i_j)}\right] - \mathbb{E}\left[e^{-B_j(r_j-i_j)}\mathbb{1}_{\{r_k<i_j<r_j\}}\right] \\
&\quad -\mathbb{E}\left[e^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_j<i_k<r_k\}}\right] \\
&= \left(\frac{\gamma}{\gamma+B_j}\right)^m - \left(\frac{\gamma}{\gamma+B_j}\right)^m F_{m,\gamma+B_j}(r_j-r_k) \\
&\quad -\mathbb{E}\left[e^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_j<i_k<r_j\}}\right]. \qquad \text{(B.2.3)}
\end{aligned}$$

Again we must calculate $\mathbb{E}\left[e^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_j<i_k<r_j\}} \mid r_k < r_j\right]$, though now clearly the conditioning argument is reversed for j and k. We condition again on the number of stages of $\Gamma(m,\gamma)$, this time for individual j. We set $m-l$ stages in (i_j, i_k), p stages in (i_k, r_k) and $l-p$ stages in (r_k, r_j), totalling m stages overall if $l \in (0,...,m-1)$ and $p \in (0,...,l)$. This is shown in Figure B.2.

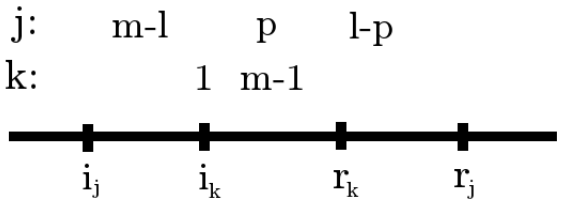


Figure B.2: For $r_k < r_j$.

As opposed to the previous case $r_k \geq r_j$, here the term $e^{-B_j(r_j-i_j)}$ corresponds to there being no points in a Poisson process over the whole time period. Hence, through the aggregation of Poisson processes, the probability of $l-p$ stages

for individual j in (r_k, r_j) is equal to the probability of $l-p$ stages of a Poisson process of rate $\gamma + B_j$. This is given by $\mathrm{e}^{-(\gamma+B_j)(r_j-r_k)}\frac{(\gamma+B_j)^{l-p}}{(l-p)!}(r_j-r_k)^{l-p}$. In addition, we require these $l-p$ stages to occur before any points in the B_j Poisson process, the probability of which is given by $\left(\frac{\gamma}{\gamma+B_j}\right)^{l-p}$.

For the time period (i_k, r_k), we require p stages for j as well as m stages for k. The number of ways to allocate j's stages among $m-1$ for k is therefore $\binom{m+p-1}{p}$ ($m-1$ since the final stage exactly at i_k must clearly be for individual k. The probability of all $m+p$ stages occurring before any points in the B_j Poisson process is $\left(\frac{\gamma}{2\gamma+B_j}\right)^{m+p}$.

Lastly, for time period (i_j, i_k) we have the remaining $m-l$ stages for individual k. The probability that these all occur before any points in the B_j process is similarly given by $\left(\frac{\gamma}{\gamma+B_j}\right)^{m-l}$.

Combining these arguments:

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_j<i_k<r_j\}} \mid r_k < r_j\right] &= \sum_{l=0}^{m-1}\sum_{p=0}^{l}\mathrm{e}^{-(\gamma+B_j)(r_j-r_k)}\frac{(\gamma+B_j)^{l-p}}{(l-p)!} \\ &\quad\times(r_j-r_k)^{l-p}\binom{m+p-1}{p}\left(\frac{\gamma}{\gamma+B_j}\right)^{l-p} \\ &\quad\times\left(\frac{\gamma}{2\gamma+B_j}\right)^{m+p}\left(\frac{\gamma}{\gamma+B_j}\right)^{m-l}.\end{aligned}$$

Substituting this into Equation (B.2.3) and rearranging then gives

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-B_j(r_j-i_j)}\mathbb{1}_{\{i_k<i_j<r_k\}} \mid r_k < r_j\right] &= \left(\frac{\gamma}{\gamma+B_j}\right)^{m}(1-F_{m,\gamma+B_j}(r_j-r_k)) \\ &\quad-\Bigg(\left(\frac{\gamma}{(\gamma+B_j)}\right)^{m}\left(\frac{\gamma}{(2\gamma+B_j)}\right)^{m} \\ &\quad\times\mathrm{e}^{-(\gamma+B_j)(r_j-r_k)}\sum_{l=0}^{m-1}\frac{(\gamma+B_j)^{l}}{l!} \\ &\quad\times\sum_{p=0}^{l}\binom{l}{p}(r_j-r_k)^{l-p} \\ &\quad\frac{(m+p-1)_p}{(2\gamma+B_j)^p}\Bigg),\end{aligned} \tag{B.2.4}$$

as was found previously in Equation (B.1.9).

Overall, combining Equations (B.2.2) and (B.2.4),

$$\mathbb{E}[\chi_j\phi_j] = \sum_{\substack{k=1,\\k\neq j}}^{n} \beta_{kj} \begin{cases} \left(\frac{\gamma}{\gamma+B_j}\right)^m (1 - F_{m,\gamma}(r_k - r_j)) \\ \qquad - \sum_{l=0}^{m-1} \frac{\gamma^{2m}}{(2\gamma+B_j)^{l+1}} \frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{(\gamma+B_j)^{m-l}\Gamma(m)} \\ \qquad \times \mathbb{E}\left[(r_k - r_j + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma+B_j)\right] \\ \qquad\qquad\qquad\qquad \text{if } r_k \geq r_j, \\ \left(\frac{\gamma}{\gamma+B_j}\right)^m \left(1 - F_{m,\gamma+B_j}(r_j - r_k)\right) \\ \qquad - \left(\frac{\gamma}{\gamma+B_j}\right)^m \left(\frac{\gamma}{2\gamma+B_j}\right)^m \mathrm{e}^{-(\gamma+B_j)(r_j-r_k)} \sum_{l=0}^{m-1} \frac{(\gamma+B_j)^l}{l!} \\ \qquad \times \mathbb{E}\left[(r_j - r_k + Y)^l \mid Y \sim \Gamma(m, 2\gamma+B_j)\right] \\ \qquad\qquad\qquad\qquad \text{if } r_k < r_j, \end{cases}$$

where

$$\mathbb{E}\left[(r+X)^l \mid X \sim \Gamma(m,\gamma)\right] = \sum_{p=0}^{l} \binom{l}{p} r^{l-p} \frac{(m+p-1)_p}{\gamma^p},$$

in agreement with the result of Equation (3.4.12).

Expression two: $\mathbb{E}[\psi_j]$

Secondly, we use probability arguments to obtain our expression for $\mathbb{E}[\psi_j]$. Recall that

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1\\k\neq j}}^{n} \mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right].$$

For each pair j,k we seek an expression for $\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}\tau_{kj}}\right]$, where $\tau_{kj} = r_k \wedge i_j - i_k \wedge i_j$. Again, for ease of exposition we write $\beta_{kj} = \beta$ in this section. We condition as usual on which of r_k and r_j is greater.

Case (i): $r_k \geq r_j$

For $r_k \geq r_j$,

$$\tau_{kj} = \begin{cases} i_j - i_k & \text{if } i_k < i_j, \\ 0 & \text{otherwise,} \end{cases}$$

therefore we require,

$$\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)}\right] = \mathbb{E}\left[\mathrm{e}^{-\beta(i_j - i_k)}\mathbb{1}_{i_k < i_j} + \mathrm{e}^{-\beta(0)}\mathbb{1}_{i_j < i_k}\right]. \tag{B.2.5}$$

Taking the first case in Equation (B.2.5), we must calculate

$$\mathbb{E}\left[e^{-\beta(i_j-i_k)}\mathbb{1}_{i_k<i_j}\right] = \mathbb{E}\left[e^{-\beta(i_j-i_k)} \mid i_k < i_j\right]\mathbb{P}[i_k < i_j]. \tag{B.2.6}$$

This case is shown in Figure B.3.

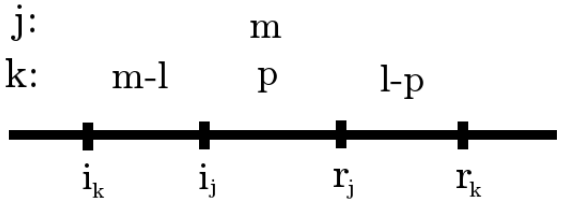


Figure B.3: For $r_k \geq r_j$, case 1.

Along similar lines to the calculation of $\mathbb{E}[\chi_j\phi_j]$, we start by working backwards from r_k. We condition on the number of stages of $\Gamma(m,\gamma)$ for individual k which occur in each time period. Specifically, we assume $l-p$ stages in (r_j, r_k) and p stages in (i_j, r_j), where $l \in (0,1,\ldots,m-1)$ and $p \in (0,1,\ldots,l)$. The probability of $l-p$ stages occurring in (r_j, r_k) is given by

$$\mathbb{P}(l-p \text{ stages in } (r_j, r_k)) = e^{-\gamma(r_j-r_k)}\frac{\gamma^{l-p}}{(l-p)!}(r_k-r_j)^{l-p},$$

since the number of stages in (r_j, r_k) is Poisson distributed with parameter γ.

Next, we must consider the period (i_j, r_j), where we must allocate the remaining p stages of $\Gamma(m,\gamma)$ for k as well as the total m stages of $\Gamma(m,\gamma)$ for j. We construct this as a Poisson process of rate 2γ, choosing an individual k or j at each stage event with probability $\frac{1}{2}$. This occurs with overall probability $(\frac{1}{2})^{m+p}$.

In addition we require the number of ways to allocate the p stages among the total $m-1+p$ (since the last event in the total m for j must occur at i_j). This is given by $\binom{m-1+p}{p}$.

Finally we consider time period (i_k, i_j). We have $m-l$ stages remaining for k, which must all occur before a Poisson process of rate β (as given by the expectation in Equation (B.2.6)), giving $\left(\frac{\gamma}{\gamma+\beta}\right)^{m-l}$.

Combining all of the constituent parts, we obtain

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta(i_j-i_k)}\mathbb{1}_{i_k<i_j}\right] &= \sum_{l=0}^{m-1}\sum_{p=0}^{l}\mathrm{e}^{-\gamma(r_j-r_k)}\frac{\gamma^{l-p}}{(l-p)!}(r_k-r_j)^{l-p}\binom{m-1+p}{p}\\ &\quad\times\left(\frac{\gamma}{\gamma+\beta}\right)^{m-l}\left(\frac{1}{2}\right)^{m+p}\\ &= \sum_{l=0}^{m-1}\frac{\mathrm{e}^{-\gamma(r_j-r_k)}}{2^m l!}\sum_{p=0}^{l}\binom{l}{p}(r_k-r_j)^{l-p}\frac{(m+p-1)_p}{(2\gamma)^p}\\ &\quad\times\left(\frac{\gamma}{\gamma+\beta}\right)^{m}(\gamma+\beta)^l.\end{aligned}$$

Taking now the second case in Equation (B.2.5), we calculate

$$\mathbb{E}[1\times\mathbb{1}_{i_k>i_j}] = \mathbb{P}[i_k>i_j] = 1-\mathbb{P}[i_k<i_j].$$

This probability is the same as that calculated in the first case of Equation (B.2.5), and is given by

$$\mathbb{P}[i_k<i_j] = \sum_{l=0}^{m-1}\sum_{p=0}^{l}\mathrm{e}^{-\gamma(r_j-r_k)}\frac{\gamma^{l-p}}{(l-p)!}(r_k-r_j)^{l-p}\binom{m-1+p}{p}\left(\frac{1}{2}\right)^{m+p}.$$

Hence,

$$\begin{aligned}\mathbb{E}[1\times\mathbb{1}_{i_k>i_j}] &= 1-\sum_{l=0}^{m-1}\sum_{p=0}^{l}\mathrm{e}^{-\gamma(r_j-r_k)}\frac{\gamma^{l-p}}{(l-p)!}(r_k-r_j)^{l-p}\binom{m-1+p}{p}\left(\frac{1}{2}\right)^{m+p}\\ &= 1-\sum_{l=0}^{m-1}\frac{\mathrm{e}^{-\gamma(r_j-r_k)}}{2^m l!}\sum_{p=0}^{l}\binom{l}{p}(r_k-r_j)^{l-p}\frac{(m+p-1)_p}{(2\gamma)^p}\gamma^l.\end{aligned}$$

Adding together these two cases, we obtain that, for $r_k \geq r_j$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k\wedge i_j-i_k\wedge i_j)}\mid r_k\geq r_j\right] =& 1+\sum_{l=0}^{m-1}\frac{\mathrm{e}^{-\gamma(r_j-r_k)}}{2^m l!}\sum_{p=0}^{l}\binom{l}{p}(r_k-r_j)^{l-p}\\ &\times\frac{(m+p-1)_p}{(2\gamma)^p}\left(\left(\frac{\gamma}{\gamma+\beta}\right)^m(\gamma+\beta)^l-\gamma^l\right),\end{aligned} \tag{B.2.7}$$

as was found previously in Equation (3.4.12).

Case (ii): $r_k < r_j$

Now for $r_k < r_j$,

$$\tau_{kj} = \begin{cases} r_k - i_k & \text{if } r_k < i_j < r_j, \\ i_j - i_k & \text{if } i_k < i_j < r_k, \\ 0 & \text{if } i_j < i_k, \end{cases}$$

therefore we require,

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\beta(r_k \wedge i_j - i_k \wedge i_j)}\right] =& \mathbb{E}\left[\mathrm{e}^{-\beta(r_k - i_k)} \mathbb{1}_{r_k < i_j < r_j}\right] + \mathbb{E}\left[\mathrm{e}^{-\beta(i_j - i_k)} \mathbb{1}_{i_k < i_j < r_k}\right] \\ &+ \mathbb{E}\left[\mathrm{e}^{-\beta(0)} \mathbb{1}_{i_j < i_k}\right], \end{aligned} \tag{B.2.8}$$

where again we write $\beta_{kj} = \beta$ for ease of exposition.

For the first of these three cases,

$$\mathbb{E}\left[\mathrm{e}^{-\beta(r_k - i_k)} \mathbb{1}_{r_k < i_j < r_j}\right] = \mathbb{E}\left[\mathrm{e}^{-\beta(r_k - i_k)}\right] \mathbb{E}\left[\mathbb{1}_{r_k < i_j < r_j}\right],$$

since time period $r_k - i_k$ is independent of the event $r_k < i_j < r_j$.

The first term is equal to the probability that m stages of a Poisson process of rate γ occur before a Poisson process of rate β, given by $\left(\frac{\gamma}{\gamma+\beta}\right)^m$. The second term is equal to the probability that, working back in time from r_j, i_j occurs before r_k. This is given by $F_{m,\gamma}(r_j - r_k)$.

Combining these arguments,

$$\mathbb{E}\left[\mathrm{e}^{-\beta(r_k - i_k)} \mathbb{1}_{r_k < i_j < r_j}\right] = \left(\frac{\gamma}{\gamma+\beta}\right)^m F_{m,\gamma}(r_j - r_k). \tag{B.2.9}$$

Progressing to the second case,

$$\mathbb{E}\left[\mathrm{e}^{-\beta(i_j - i_k)} \mathbb{1}_{i_k < i_j < r_k}\right] = \mathbb{E}\left[\mathrm{e}^{-\beta(i_j - i_k)} \mid i_k < i_j < r_k\right] \mathbb{P}[i_k < i_j < r_k].$$

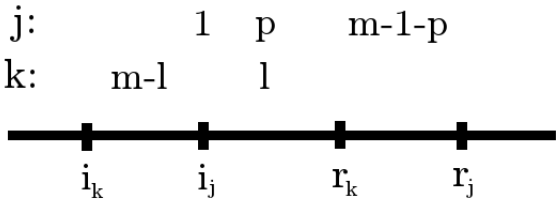


Figure B.4: For $r_k < r_j$, case 2.

Figure B.4 shows the timeline for this case, and the calculations are similar to those seen previously. We condition on there being $m-1-p$ stages for individual j in the period (r_k, r_j), with $p+1$ further stages in (i_j, r_k) for a total of m. Here $p \in (0, 1, \ldots, m)$. The probability of the $m-1-p$ stages in (r_k, r_j) is equal to $\mathrm{e}^{-\gamma(r_j-r_k)}\frac{\gamma^{m-1-p}}{(m-1-p)!}(r_j-r_k)^{m-1-p}$.

Moving back to the previous time stage (i_j, r_k), we require $p+1$ stages for individual j as well as setting l stages for individual k, for $l \in (0, 1, \ldots, m-1)$. These form a Poisson process of rate 2γ, the probability of which occurring is $(\frac{1}{2})^{p+l+1}$. The number of ways to allocate p stages for j among the l for k (since the last event must necessarily be for j at i_j) is given by $\binom{l+p}{p}$.

Lastly we consider time period (i_k, i_j). Here we have $m-l$ remaining stages for k, to occur before a Poisson process of rate β. The probability of this is $\left(\frac{\gamma}{\gamma+\beta}\right)^{m-l}$.

Combining the constituent parts,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta(i_j-i_k)}\mathbb{1}_{i_k<i_j<r_k}\right] &= \sum_{l=0}^{m-1}\sum_{p=0}^{m-1}\mathrm{e}^{-\gamma(r_j-r_k)}\frac{\gamma^{m-1-p}}{(m-1-p)!}(r_j-r_k)^{m-1-p}\left(\frac{1}{2}\right)^{p+l+1}\\ &\quad \binom{l+p}{p}\left(\frac{\gamma}{\gamma+\beta}\right)^{m-l}\\ &= \sum_{l=0}^{m-1}\frac{\gamma^{m-1}}{2^{l+1}}\frac{\mathrm{e}^{-\gamma(r_j-r_k)}}{\Gamma(m)}\sum_{p=0}^{m-1}\binom{m-1}{p}(r_j-r_k)^{m-1-p}\\ &\quad \frac{(l+p)_p}{(2\gamma)^p}\left(\frac{\gamma}{\gamma+\beta}\right)^{m-l}. \qquad \text{(B.2.10)}\end{aligned}$$

Lastly we consider the third case,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta(0)}\mathbb{1}_{i_j<i_k}\right] &= \mathbb{P}[i_j < i_k]\\ &= 1-\mathbb{P}[i_k<i_j<r_k]-\mathbb{P}[r_k<i_j<r_j]\\ &= 1-\mathbb{E}\left[\mathrm{e}^{-0}\mathbb{1}_{i_k<i_j<r_k}\right]-F_{m,\gamma}(r_j-r_k),\end{aligned}$$

where the term $\mathbb{E}\left[\mathrm{e}^{-0}\mathbb{1}_{i_k<i_j<r_k}\right]$ is simply the second case above, given in Equation (B.2.10), with $\beta = 0$.

Hence,

$$\mathbb{E}\left[\mathrm{e}^{-\beta(0)}\mathbb{1}_{i_j<i_k}\right] = 1 - F_{m,\gamma}(r_j - r_k) - \sum_{l=0}^{m-1} \frac{\gamma^{m-1}}{2^{l+1}} \frac{\mathrm{e}^{-\gamma(r_j-r_k)}}{\Gamma(m)} \sum_{p=0}^{m-1} \binom{m-1}{p} \times (r_j - r_k)^{m-1-p} \frac{(l+p)_p}{(2\gamma)^p}. \tag{B.2.11}$$

Combining all three cases from Equations (B.2.9), (B.2.10) and (B.2.11), we obtain the overall expression for $r_k < r_j$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{-\beta_{kj}(r_k \wedge i_j - i_k \wedge i_j)} \mid r_k < r_j\right] =& 1 - F_{m,\gamma}(r_j - r_k)\left(1 - \left(\frac{\gamma}{\gamma+\beta}\right)^m\right) + \sum_{l=0}^{m-1} \frac{\gamma^{m-1}}{2^{l+1}} \\ &\times \frac{\mathrm{e}^{-\gamma(r_j-r_k)}}{\Gamma(m)} \sum_{p=0}^{m-1} \binom{m-1}{p} (r_j - r_k)^{m-1-p} \\ &\times \frac{(l+p)_p}{(2\gamma)^p} \left(\left(\frac{\gamma}{\gamma+\beta}\right)^{m-l} - 1\right),\end{aligned} \tag{B.2.12}$$

which is equal to Equation (B.1.18).

Overall, combining Equations (B.2.7) and (B.2.12),

$$\mathbb{E}[\psi_j] = \prod_{\substack{k=1 \\ k\neq j}}^{n} \begin{cases} 1 + \sum_{l=0}^{m-1} \frac{\mathrm{e}^{-\gamma(r_k-r_j)}}{l!2^m} \mathbb{E}\left[((r_k - r_j + Y)^l) \mid Y \sim \Gamma(m, 2\gamma)\right] \\ \qquad \times \left(\left(\frac{\gamma}{\gamma+\beta_{kj}}\right)^m (\gamma + \beta_{kj})^l - \gamma^l\right) & \text{if } r_k \geq r_j, \\ 1 - F_{m,\gamma}(r_j - r_k)\left(1 - \left(\frac{\gamma}{\gamma+\beta_{kj}}\right)^m\right) + \sum_{l=0}^{m-1} \frac{\gamma^{m-1}\mathrm{e}^{-\gamma(r_j-r_k)}}{2^{l+1}\Gamma(m)} \\ \qquad \times \mathbb{E}\left[(r_j - r_k + Y)^{m-1} \mid Y \sim \Gamma(l+1, 2\gamma)\right] \\ \qquad \times \left(\left(\frac{\gamma}{\gamma+\beta_{kj}}\right)^m \left(\frac{\gamma+\beta_{kj}}{\gamma}\right)^l - 1\right) & \text{if } r_k < r_j. \end{cases}$$

This is in agreement with Equation (3.4.12), and so overall the whole likelihood expression agrees with that obtained via integration arguments.

Bibliography

H. Abbey. An examination of the Reed-Frost theory of epidemics. *Human Biology*, 24(3), 1952.

C.L Addy, I.M. Longini, and M. Haber. A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, 47(3):961–974, 1991.

S. Alexandersen, Z. Zhang, A.I. Donaldson, and A.J.M. Garland. The Pathogenesis and Diagnosis of Foot-and-Mouth Disease. *Journal of Comparative Pathology*, 129:1–36, 2003.

C.L. Althaus. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLoS Currents Outbreaks*, 6, 2014.

R.M. Anderson and R.M. May. Directly transmitted infections diseases: control by vaccination. *Science*, 215(4536):1053–1060, 1982.

R.M. Anderson and R.M. May. *Infectious diseases of humans*. Oxford University Press, 1991.

H. Andersson and T Britton. *Stochastic epidemic models and their statistical analysis*. Springer, 2000.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistical Computing*, 18:343–373, 2008.

N.T.J. Bailey. *The mathematical theory of infectious diseases and its applications*. Charles Griffin and Company LTD., 1975.

N.T.J. Bailey and A.S. Thomas. The estimation of parameters from population data on the general stochastic epidemic. *Theoretical Population Biology*, 2:253–270, 1971.

S. Baize, D. Pannetier, L. Oestereich, T. Rieger, L. Koivogui, N. Magassouba, B. Soropogui, M. S. Sow, S. Keïta, H. De Clerck, A. Tiffany, G. Dominguez, M. Loua, A. Traoré, M. Kolié, E. R. Malano, E. Heleze, A. Bocquin, S. Mély, H. Raoul, V. Caro, D. Cadar, M. Gabriel, M. Pahlmann, D. Tappe, J. Schmidt-Chanasit, B. Impouma, A. K. Diallo, P. Formenty, M. Van Herp, and S. Günther. Emergence of Zaire Ebola virus in Guinea. *The New England Journal of Medicine*, 371:1418–1425, 2014.

F.G. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *The Annals of Applied Probability*, pages 46–89, 1997.

F.G. Ball, T. Britton, and P.D. O'Neill. Empty confidence sets for epidemics, branching processes and Brownian motion. *Biometrika*, 89:211–224, 2002.

A.D. Barbour and G.K. Eagleson. Multiple comparisons and sums of dissociated random variables. *Advances in Applied Probability*, 17(1):147–162, 1985.

M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of Theoretical Biology*, 235:275–288, 2005.

M.S. Bartlett. Some evolutionary stochastic processes. *Journal of the Royal Statistical Society Series B*, 11:211–229, 1949.

P. Becchetti, M.R. Segal, and N.P. Jewell. Backcalculation of HIV infection rates. *Statistical Science*, 8(2):82–119, 1993.

N.G. Becker. Estimation for an epidemic model. *Biometrics*, 32:769–777, 1976.

N.G. Becker. *Analysis of infectious disease data*. Routledge, 1989.

N.G. Becker. Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases. *Statistical Methods in Medical Research*, 6(1): 24–37, 1997.

N.G. Becker. *Modeling to inform infectious disease control*. Chapman and Hall/CRC, 2015.

N.G. Becker and T. Britton. Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society Series B*, 61(2):287–307, 1999.

N.G. Becker and J. L. Hopper. Assessing the heterogeneity of disease spread through a community. *American Journal of Epidemiology*, 117(3):362–374, 1983.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. New York: Wiley, 1994.

M.G. Blum and V.C. Tran. HIV with contact tracing: a case study in approximate Bayesian computation. *Biostatistics*, 11(4):644–660, 2010.

R.J. Boys and P.R. Giles. Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates. *Journal of Mathematical Biology*, 55: 223–247, 2007.

T. Britton and N.G. Becker. Estimating the immunity coverage required to prevent epidemics in a community of households. *Biostatistics*, 1(4):389–402, 2000.

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

T. Britton, P.D. O'Neill, and T. Kypraios. Inference for Epidemics with Three Levels of Mixing: Methodology and Application to a Measles Outbreak. *Scandinavian Journal of Statistics*, 38(3):578–599, 2011.

T. Britton, T. House, A.L. Lloyd, D. Mollison, S. Riley, and P. Trapman. Five challenges for stochastic epidemic models involving global transmission. *Epidemics*, 10:54–57, 2015.

R. Brookmeyer. Reconstruction and future trends of the AIDS epidemic in the United States. *Science*, 253(5015):37–42, 1991.

S. Brooks, A. Gelman, G. Jones, and X.L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

E. Brooks-Pollock, G.O. Roberts, and M.J. Keeling. A dynamic model of bovine tuberculosis spread and control in Great Britain. *Nature*, 511:228–231, 2014.

F. Brown. New approaches to vaccination against foot-and-mouth disease. *Vaccine*, 10(14):1022–1026, 1992.

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

Centers for Disease Control and Prevention. Ebola Previous Case Counts. `https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa`, Accessed 2018-03-11.

G. Chowell, N.W. Hengartner, C. Castillo-Chavez, P.W. Fenimore, and J.M. Hyman. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology*, 229(1):119–126, 2004.

G. Chowell, P. Diaz-Duenas, J.C. Miller, A. Alcazar-Velazco, J.M. Hyman, P.W Fenimore, and C. Castillo-Chavez. Estimation of the reproduction number of dengue fever from spatial epidemic data. *Mathematical Biosciences*, 208: 571–589, 2007.

D. Clancy and P.D. O'Neill. Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis*, 3:737–758, 2008.

D.R. Cox and N. Read. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

K. Csilléry, M.G.B. Blum, O.E. Gaggiotti, and O. François. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, 25(7):410–418, 2010.

D.J. Daley and J. Gani. *Epidemic modelling: an introduction*. Cambridge University Press, 2001.

L. Danon, A.P. Ford, T. House, C.P. Jewell, M.J. Keeling, G.O. Roberts, J.V. Ross, and M.C. Vernon. Networks and the Epidemiology of Infectious Disease. *Interdisciplinary Perspectives on Infectious Diseases*, 2011.

D. De Angelis, A. M. Presanis, P. J. Birrell, G. S. Tomba, and T. House. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, 10:83–87, 2015.

N. Demiris and P.D. O'Neill. Bayesian inference for epidemicswith two levels of mixing. *Scandinavian Journal of Statistics*, 32(2):265–280, 2005.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

P.J. Diggle. Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical Methods in Medical Research*, 15(4):325–336, 2006.

M. Eichner and K. Dietz. Transmission potential of smallpox: estimates based on detailed data from an outbreak. *American Journal of Epidemiology*, 158(2): 110–117, 2003.

A. Eto, T. Saito, H. Yokote, I. Kurane, and Y. Kanatani. Recent advances in the study of live attenuated cell-cultured smallpox vaccine LC16m8. *Vaccine*, 33(45):6106–6111, 2015.

N.M. Ferguson, C.A. Donnelly, and R.M. Anderson. The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science*, 413:542–547, 2001.

N.M. Ferguson, M.J. Keeling, W.J. Edmunds, R. Gani, B.T. Grenfell, R.M. Anderson, and S. Leach. Planning for smallpox outbreaks. *Nature*, 425:681–685, 2003.

J.A.N. Filipe and G.J. Gibson. Studying and approximating spatio-temporal models for epidemic spread and control. *Philosophical Transaction of the Royal Society B*, 353:2153–2162, 1998.

C. Fronterrè, E. Giorgi, and P. Diggle. Geostatistical inference in the presence of geomasking: a composite-likelihood approach. *arXiv preprint*, arXiv:1711.00437, 2017.

R. Gani and S. Leach. Transmission potential of smallpox in contemporary publications. *Nature*, 414:748–751, 2001.

A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

A.E. Gelfand, S.K. Sahu, and B.P. Carlin. Efficient parameterizations for generalised linear models. *Bayesian Statistics*, 5:479–488, 1996.

A. Gelman, X.L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. 3rd edition, 2013.

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.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, 4:169–193, 1992.

G.J. Gibson and E. Renshaw. Estimating Parameters in Stochastic Compartmental Models using Markov Chain Methods. *IMA Journal of Mathematics Applied in Medicine and Biology*, 15:19–40, 1998.

W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in practice*. CRC Press, 1995.

A. Golightly, D.A. Henderson, and C. Sherlock. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statistics and Computing*, pages 1–17, 2014.

M.J. Grubman and B. Baxt. Foot-and-Mouth Disease. *Clinical Microbiology Reviews*, 17(2):465–493, 2004.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2014.

M.E. Halloran, I.M. Longini, A. Nizam, and Y. Yang. Containing bioterrorist smallpox. *Science*, 298:1428–1432, 2002.

J. Hammond and D.A.J. Tyrrell. A mathematical model of common-cold epidemics on Tristan Da Cunha. *Journal of Hygiene (Cambridge)*, 69(3):423–433, 1971.

W. Hastings. Monte carlo sampling using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

Y. Hayakawa, P.D. O'Neill, D. Upton, and P.S.F. Yip. Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles for several types. *Australian and New Zealand Journal of Statistics*, 45(4):491–502, 2003.

J.A.P. Heesterbeek and K. Dietz. The concept of R0 in epidemic theory. *Statistica Neerlandica*, 50(1):89–110, 1996.

D.A. Henderson and I. Arita. The Smallpox Threat: A Time to Reconsider Global Policy. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 12(3):117–121, 2014.

R.M. Huggins, P.S.F. Yip, and E.H.Y. Lau. A note on the estimation of the initial number of susceptible individuals in the general epidemic model. *Statistics and Probability Letters*, 67:321–330, 2004.

R.E. Kass, B.P. Carlin, A. Gelman, and R.M. Neal. Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.

M.J. Keeling and K.T.D. Eames. Networks and epidemic models. *Journal of the Royal Statistical Society Interface*, 2(4):295–307, 2005.

W.O. Kermack and A.G McKendrick. A contribution to the mathematical theory of epidemics. *Proceeding of the Royal Society A*, 115:700–721, 1927.

T. Kypraios. Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models. 2007.

T. Kypraios. A note on maximum likelihood estimation of the initial number of susceptibles in the general stochastic epidemic model. *Statistics and Probability Letters*, 79(18):1972–1976, 2009.

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

F. Larribe and P. Fearnhead. On composite likelihoods in statistical genetics. *Statistica Sinica*, 21:43–69, 2011.

E.H.Y. Lau and P.S.F. Yip. Estimating the basic reproductive number in the general epidemic model with an unknown initial number of susceptible individuals. *Scandinavian Journal of Statistics*, 35:650–663, 2008.

M.S.Y. Lau, G. Marion, G. Streftaris, and G.J. Gibson. New model diagnostics for spatiotemporal systems in epidemiology and ecology. *Journal of the Royal Society Interface*, 11, 2014.

P.M. Lee. *Bayesian statistics: an introduction*. John Wiley and Sons, 2012.

A.L. Lloyd. Realistic Distributions of Infectious Periods in Epidemic Models: Changing Patterns of Persistence and Dynamics. *Theoretical Population Biology*, 60(1):59–71, 2001.

A.G. McKendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1926.

T.J. McKinley, A.R. Cook, and R. Deardon. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1):Article 24, 2009.

T.J. McKinley, J.V. Ross, R. Deardon, and A.R. Cook. Simulation-based Bayesian inference for epidemic models. *Computational Statistics and Data Analysis*, 71:434–447, 2014.

M.I. Meltzer, I. Damon, J.M. LeDuc, and J.D. Millar. Modeling potential responses to smallpox as a bioterrorist weapon. *Emerging Infectious Diseases*, 7: 959–969, 2001.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

D. Mollison. *Epidemic Models: Their Structure and Relation to Data*. Cambridge University Press, 1995.

R.S. Morris, J. Wilesmith, M.W. Stern, R.L Sanson, and M.A. Stevenson. Predictive spatial modelling of alternative control strategies for the foot-and-mouth disease epidemic in Great Britain. *Veterinary Record*, 149:137–145, 2001.

P. Neal and G. Roberts. Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, 5(2):249–261, 2004.

P. Neal and G. Roberts. A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing*, 15:315–327, 2005.

M.E.J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1), 2002.

S.W.B. Newsom. Pioneers in infection control: John Snow, Henry Whitehead, the Broad Street pump, and the beginnings of geographical epidemiology. *Journal of Hospital Infection*, 64:210–216, 2006.

C. Oh. Improved estimation of the initial number of susceptible individuals in the general stochastic epidemic model using penalized likelihood. *The Scientific World Journal*, 2014, 2014.

P.D. O'Neill. Introduction and snapshot review: Relating infectious disease transmission models to data. *Statistics in Medicine*, 29(20):2069–2077, 2010.

P.D. O'Neill and N.G. Becker. Inference for an epidemic when susceptibility varies. *Biostatistics*, 2:99–108, 2001.

P.D. O'Neill and G.O. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society Series A*, 162:121–283, 1999.

O. Papaspiliopoulos, G. Roberts, and M. Sköld. Non-centered parameterisations for hierarchical models and data augmentation. *Bayesian Statistics*, 7: 307–326, 2003.

A. Rambaut, O.G. Pybus, M.I. Nelson, C. Viboud, J.K. Taubenberger, and E.C. Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619, 2008.

J. Ray and Y.M. Marzouk. A Bayesian method for inferring transmission chains in a partially observed epidemic. *Proceedings of the Joint Statistical Meetings*, pages 3–7, 2008.

G.O. Roberts and S.K. Sahu. Updating Schemes, Correlation Structure, Blocking and Parameterizationfor the Gibbs Sampler. *Journal of the Royal Statisitcal Society Series B*, 59(2):291–317, 1997.

G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120, 1997.

D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12(4):1151–1172, 1984.

R. Shanmugan. Spinned Poisson distribution with health management application. *Health Care Management Science*, 14:299–306, 2011.

M. Shibli, S. Gooch, H.E. Lewis, and D.A.J. Tyrrell. Common colds on Tristan Da Cunha. *Journal of Hygiene (Cambridge)*, 69(2):255–262, 1971.

S.A. Sisson, Y. Fan, and M.M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

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

G. Streftaris and G.J. Gibson. Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling*, 4:63–75, 2004.

M. Sunnåker, A.G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1), 2013.

M.M. Tanaka, A.R. Francis, F. Luciani, and S.A. Sisson. Using Approximate Bayesian Computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3):1511–1520, 2006.

M.A. Tanner and W.H. Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Assosiation*, 82(398): 528–540, 1987.

S. Tavaré, D. Balding, R. Griffith, and P.Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

D. Thompson and W. Foege. Faith Tabernacle smallpox epidemic. Abakaliki, Nigeria. *World Health Organization*, 3:1–9, 1968.

L. Tierney. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M.P.H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Statistical Society Interface*, 6:187–202, 2009.

C. Varin, N. Read, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

C.J. Worby, P.D. O'Neill, T. Kypraios, J.V. Robotham, D. De Angelis, E.J.P. Cartwright, S.J. Peacock, and B.S. Cooper. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Annals of Applied Statistics*, 10(1):395–417, 2016.

World Health Organisation. The Independent Advisory Group on Public Health Implications of Synthetic Biology Technology Related to Smallpox. *World Health Organization*, WHO/HSE/PED/2015, 2015.

World Health Organisation. WHO Ebola data and statistics. `http://apps.who.int/gho/data/node.ebola-sitrep`, Accessed 2017-11-27.

F. Xiang and P. Neal. Efficient MCMC for temporal epidemics via parameter reduction. *Computational Statistics and Data Analysis*, 80:240–250, 2014.

X. Xu, T. Kypraios, and P.D. O'Neill. Bayesian non-parametric inference for stochastic epidemic models using Gaussian Processes. *Theoretical Population Biology*, 17(4):619–633, 2016.

P.S.F. Yip. Estimating the initial relative infection rate for a stochastic epidemic model. *Biostatistics*, 36:202–213, 1989.