Statistical Modeling and Analysis of Partially Observed Infectious Diseases.

Chibuzor Christopher Nnanatu, B.Sc., M.Sc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University.

March, 2018.

Abstract

This thesis is concerned with the development of Bayesian inference approach for the analysis of infectious disease models. Stochastic SIS household-based epidemic models were considered with individuals allowed to be contracted locally at a given rate and there also exists a global force of infection. The study covers both when the population of interest is assumed to be constant and when the population is allowed to vary over time. It also covers when the global force of infection is constant and when it is spatially varying as a function of some unobserved Gaussian random fields realizations. In addition, we also considered diseases coinfection models allowing multiple strains transmission and recovery. For each model, Bayesian inference approach was developed and implemented via MCMC framework using extensive data augmentation schema. Throughout, we consider two most prevalent forms of endemic disease data- the individual-based data and the aggregate-based data. The models and Bayesian approach were tested with simulated data sets and successfully applied to real-life data sets of tick-borne diseases among Tanzania cattle.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chibuzor Christopher Nnanatu

Acknowledgements

First of all, I would like to show my profound gratitude to my supervisor Prof Peter Neal for believing in me and giving me the kindest opportunity to learn independently making sure that this project is successfully accomplished. Honestly, this success story would not have been told without Peter Neal's thorough supervision, guidance and patience. Thank you very much Pete!

Let me also thank some wondeful people I was so fortunate to meet during the course of my PhD studies starting with Clement, Simon, Ross and Thitiya. I am particularly indebted to Clement for his many helps throughout my PhD studies. I also would like to thank Laura, Abbie and Callum for your many helps in proofreading this thesis. Thanks to Becky, Jess, Ben, Val, Ayesha, Anna, Amy, Chantell, Okezie and the entire B18 crew for creating an amazing learning and social environment. It was indeed a pleasure having you guys as my colleagues. To my very good friend Chigozie Utazi, I say thank you very much for your many supports.

Finally and most importantly, I would like to thank my lovely family: my ever caring mother Mrs Rose C. Nnanatu for her very effective prayers, my very beautiful

wife Barr.(Mrs) Cynthia C. Nnanatu for her unalloyed support and encouragements for which I would forever remain grateful for, and my children Emerald, Fechi and Olaedo for understanding the limitations of a PhD student dad. I love you all immensely! In closing, I thank Nigerian Government through TETFund for their financial support.

Contents

List of Tables

List of Figures

List of Abbreviations

ABD	aggregate-based data
ABC	approximate Bayesian computation
ACF	autocorrelation function
CTMC	continuous-time Markov chain
DTMC	discrete-time Markov chain
EM	expectation-maximization
GRF	Gaussian Random Field
GSE	general stochastic epidemic
IBD	individual-based data
IS	Independence Sampler
MCMC	Markov Chain Monte Carlo
M-H	Metropolis-Hastings
NCP	non-centered parameterisation
PNCP	partially non-centered parameterisation
PPP	Poisson point process
RWM	Random walk Metropolis
SDE	stochastic differential equation

Chapter 1

Introduction

1.1 Overview

Outbreaks of infectious diseases (both new and re-emerging) in both human and animal hosts are a growing concern due to their attendant high morbidity and mortality rates, and the severe economic and social effects, for example, UK 2001 Foot and Mouth diseases.

Understanding the dynamics of infectious diseases transmission offers a great insight into the key drivers of the transmission processes. Mathematical models for infectious disease are normally used to capture the actual disease transmission mechanisms, gain further insight on the epidemiological, immunological and evolutionary behaviors of the infectious disease of interest. By making appropriate assumptions about the model parameters, and being able to infer the parameters of the model, an epidemic model would potentially answer several public health questions concerning the severity of an infectious disease and the final size of an epidemic. Epidemic model can also serve as means through which public health

practitioners could be kept abreast of the most effective control strategies to be implemented in the face of an outbreak, for example vaccination policies, quarantine measures, movement restrictions or other procedures, see, for example, Keeling et al. (2003).

It is well known that most infectious disease data are non-standard and are usually only partially observed, see, for example, Britton and O'Neill (2002). This partial observation of most infectious disease data, which are largely due to certain unobserved processes, presents a huge setback in the analysis of infectious disease data (for most likelihood-based inference methods requiring the evaluation of the likelihood function.) However, as more fast computing machines become available, several computer-intensive approaches are being developed. One of such methods is Markov Chain Monte Carlo (MCMC) algorithms which allows posterior inference no matter how complicated the likelihood function may be.

In this thesis, we shall be concerned with the development of efficient statistical inference approach for epidemic models. Specifically, this thesis develops such inference approach for household-based stochastic epidemic models via a Bayesian framework, and implemented using Markov Chain Monte Carlo (MCMC) algorithms. Throughout, we shall focus on the endemic SIS (susceptible $\rightarrow$ infected $\rightarrow$ susceptible) model. However, with appropriate model assumptions adjustments, the methods developed in the next three chapters can readily be applied to other class of epidemic models for both humans and animals populations. We shall use, where appropriate, a simulated data sets and/ a real-life data sets both to illustrate our approach.

The remainder of this introductory chapter is organized as follows: In Section

1.2, we give an outline of Bayes' paradigm. In Section 1.4, we give an overview of MCMC methods and discuss a few well known MCMC algorithms, especially those relevant to our purpose. In Section 1.4.10, we describe the implementation of data augmentation and explain how the idea of data augmentation allows iterative sampling. In Section 1.5, we give an overview of epidemic models, highlighting the breakthroughs and setbacks of epidemic models. Finally in Section 1.8, we outline the contributions of this thesis to literature.

1.2 Bayesian Inference

In this section, we give an overview of Bayesian inference in general and outline the theoretical framework of the Bayesian paradigm. We also give the development of Bayesian statistical inference computational methods which form the basis of the inferential methods developed in this thesis.

1.2.1 Introduction

Given data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ which is assumed to arise from a model $\mathcal{M}$ with parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, in classical (or frequentist) statistics, the parameters $\boldsymbol{\theta}$ are fixed constants. Then with the data through the likelihood function, maximum likelihood estimate (MLE) of the parameters can be calculated. Other quantities of interest such as standard error and confidence intervals for the maximum likelihood estimate can readily be obtained.

Conversely, Bayesian statistics assumes that the model parameters themselves are random variables to be estimated from the model rather than constant parame-

ters. Note that both assume parametric model and also involve the use of the likelihood function. The *likelihood* function is the conditional distribution of the data given the model parameters. In addition, Bayesian statistics allows us to place *prior* distribution on the parameters. The prior distribution represents any prior knowledge we or experts have on the parameters. The prior distribution may possess a great deal of information concerning $\boldsymbol{\theta}$ in which case we say that the prior is *informative*. On the other hand, the prior distribution may contain too little or no information about $\boldsymbol{\theta}$ in which case we say that the prior is *noninformative*. Bayesian inference then combines the likelihood function and the prior distribution using Bayes' theorem to obtain the *posterior distribution*. The posterior distribution gives the conditional distribution of the unknown parameters given the data. Therefore, the posterior distribution which depends on the parametric model, the data and the choice of prior. Once the posterior distribution has been obtained quantities of interest such as mean, mode and variance of the parameters can be estimated directly from the posterior distribution. For a more rigorous and detailed discussion of the Bayesian statistical inference framework, see, for example, Bernardo and Smith (1994).

We shall now present the basic theoretical framework of Bayesian statistical inference used in this thesis. We adopt the notations introduced above denoting the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ assumed to arise from a model $\mathcal{M}$ with d-dimensional parameters, $\boldsymbol{\theta}$. Note that throughout this section, we use $\boldsymbol{\theta}$ and θ to denote a vector of parameter values and a single parameter value, respectively. Similarly, we use $\mathbf{x}$ and x to denote a data matrix and a single data point, respectively.

1.2.2 Bayes' Theorem

Given the data $\mathbf{x}$ and the parameters of the model $\boldsymbol{\theta}$, let $\pi(\boldsymbol{\theta})$ denote the prior distribution on the parameters representing our beliefs on the parameters. Let $\pi(\mathbf{x}|\boldsymbol{\theta})$ denote the likelihood function, *i.e.*, the conditional distribution of the data given the unknown parameters. Also, let $\pi(\mathbf{x})$ denote the marginal distribution of the data $\mathbf{x}$ and let $\pi(\boldsymbol{\theta}|\mathbf{x})$ denote the posterior distribution of the parameters given the data. Then Bayes' Theorem is as follows:

$$\begin{aligned} \pi(\boldsymbol{\theta}, \mathbf{x}) &= \pi(\mathbf{x}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) \\ &= \pi(\boldsymbol{\theta}|\mathbf{x})\, \pi(\mathbf{x}). \end{aligned} \tag{1.2.1}$$

In Bayesian statistics, the primary interest is on the posterior distribution of the parameters given the data, $\pi(\boldsymbol{\theta}|\mathbf{x})$, which is given by $\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/\pi(\mathbf{x})$. That is,

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x})} \tag{1.2.2}$$

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ \pi(\theta|\mathbf{x}) &= \mathcal{K} \times \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ \pi(\theta|\mathbf{x}) &= \mathcal{K} \times \text{likelihood} \times \text{prior}. \end{aligned} \tag{1.2.3}$$

where the marginal distribution $\pi(\mathbf{x})$ does not depend on the parameters $\boldsymbol{\theta}$ and $\mathcal{K}$ is a constant of proportionality or the *normalizing constant* which should be computed. In most practical, complex situations, it is not analytically possible to calculate the constant of proportionality $\mathcal{K}$ and this has been a major hitch on the progress of Bayesian statistical inference. Thankfully, there exist modern computer intensive Bayesian statistical techniques which only requires the knowledge of the

posterior distribution up to the constant of proportionality. These techniques known as Markov chain Monte Carlos (MCMC) algorithms have been successfully applied to wide range of realistically complex models in Bayesian framework. With MCMC, the impediment that would have been encountered on with calculation of the normalizing constant is sidestepped. All that is is required is to construct a suitable MCMC algorithm whose stationary distribution is our target distribution. It is straightforward to extend the Bayesian framework to cases involving two or more independent data samples, such as data obtained sequentially over time. Suppose we have two independent data samples $\mathbf{x}_1$ and $\mathbf{x}_2$ which are assumed to arise from the model $\mathcal{M}$ and with d-dimensional parameters $\boldsymbol{\theta}$. Then

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2) &\propto \pi(\mathbf{x}_1, \mathbf{x}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \pi_1(\mathbf{x}_1|\boldsymbol{\theta})\pi_2(\mathbf{x}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto \pi_2(\mathbf{x}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x}_1). \end{aligned} \tag{1.2.4}$$

From (1.2.4) we see that we can obtain the joint posterior distribution of the parameters given the two independent data sets, $\pi(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2)$ by simply evaluating the posterior distribution of $\mathbf{x}_1$ given the parameters, $\pi(\boldsymbol{\theta}|\mathbf{x}_1) \propto \pi_1(\mathbf{x}_1|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, and then using it as the prior for the likelihood of the second data given the parameters, $\pi_2(\mathbf{x}_2|\boldsymbol{\theta})$. This framework is widely adapted for the Bayesian statistical inference approach we develop in this thesis. In Section 1.4, we give an overview of the development of MCMC algorithms and their applications in Bayesian statistical inference contexts.

We now give outline of various forms of prior distribution popular in Bayesian statistical inference framework.

1.2.3 Prior Distribution

In this section we briefly present a few commonly used prior distributions in Bayesian statistical inference. We recall that a prior distribution captures our beliefs about the parameters distribution and can be informative or non-informative. Priors can also be *proper* or *improper*. A very popular prior distribution largely used due to mathematical convenience is one which gives rise to a posterior distribution that is in the same family of parametric distribution as the prior distribution. In this case, the prior is said to be a *conjugate* prior.

Conjugate priors

As already stated above, a prior is said to be conjugate if the resulting posterior distribution belonging to the same family of parametric distribution as the prior. Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are independent and identically distributed data according to an exponentially distributed random variable X. That is, $X \sim \exp(\theta)$. Then

$$f(x|\theta) = \theta e^{-\theta x} \quad x \geq 0, \tag{1.2.5}$$

then the likelihood function is

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} f(x_i|\theta), \\ &= \prod_{i=1}^{n} \theta e^{-\theta x_i}, \\ &= \theta^n \exp(-\theta \sum_{i=1}^{n} x_i). \end{aligned} \tag{1.2.6}$$

Suppose we place gamma-distributed prior on θ, that is, $\pi(\theta) \sim Gamma(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$ are the shape and the scale parameters, respectively. Then the posterior distribution is given as follows:

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto L(\theta|\mathbf{x})\pi(\theta) \\ &\propto \theta^n \exp(-\theta \sum_{i=1}^{n} x_i) \times \theta^{\alpha-1} \exp(-\theta\beta) \\ &= \theta^{n+\alpha-1} \exp(-\theta(\beta + \sum_{i=1}^{n} x_i)), \end{aligned} \tag{1.2.7}$$

$\Rightarrow \theta|\mathbf{x} \sim Gamma(n + \alpha, \beta + \sum_{i=1}^{n} x_i)$.

Therefore, the gamma distributed prior $\pi(\theta)$ is a conjugate prior since the the posterior distribution is also gamma-distributed as the prior distribution.

Proper and Improper prior distributions

A prior distribution $\pi(\theta)$ is said to be proper if it integrates to unity, *i.e.*, $\int_{-\infty}^{\infty} \pi(\theta)d\theta = 1$. On the other hand, an improper prior distribution is one which does not integrate to unity. For example, suppose θ is assigned the prior $\pi(\theta) \propto 1$. Clearly, this is an improper prior since $\int_{-\infty}^{\infty} \pi(\theta)d\theta = \infty$. Nonetheless, placing improper prior distribution on parameters often gives rise to proper standard posterior distributions making its use unproblematic.

1.3 Markov chain Monte Carlo

In this section, we present an overview of Bayesian statistical inference via Markov chain Monte Carlo (MCMC) algorithms. In particular, we give a brief outline on the development of MCMC with a detailed description of some commonly used

MCMC algorithms. We first outline the underlying theoretical basis for MCMC beginning with Markov chains and Monte Carlo methods.

1.3.1 Markov Chains

A Markov chain $\{X_n : n \geq 0\}$ is a stochastic (random) process taking values in the state space $\mathcal{S}$, and which satisfies the following 'memoryless' property:

$$\mathbb{P}(X_{n+1} \in \mathcal{S}|X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} \in \mathcal{S}|X_n = x_n),$$

where X_n denotes the state of the chain after n steps. In order words, a Markov chain is a stochastic process in which the future state (X_{n+1}) of the process is independent of the past state ($X_{n-1}, \ldots, X_0$) given the present state (X_n). When the Markov chains do not depend upon n they are said to be *homogeneous*. That is, for $n \geq 0$ and $i, j \in \mathcal{S}$

$$\mathbb{P}(X_{n+1} = j|X_n = i) = \mathbb{P}(X_1 = j|X_0 = i). \tag{1.3.1}$$

Then there exists an $|\mathcal{S}| \times |\mathcal{S}|$ transition probability matrix $\mathbf{P} = (p_{ij})$ which describes the evolution of the chain, where

$$p_{ij} = \mathbb{P}(X_{n+1} = j|X_n = i) \tag{1.3.2}$$

is the probability of the chain being in state j from state i. The transition matrix $\mathbf{P}$ is stochastic in that its entries are non-negative values, and the row elements sum to unity. That is, $p_{ij} \geq 0$ for all i, j, and $\sum_j p_{ij} = 1$ for all i. Note that the descriptions so far given are for discrete-time Markov chain (DTMC) in which the

jump times and state space take values in the discrete set $\{0, 1, 2, \ldots\}$. We shall now give basic properties of Markov chains.

Irreducibility

A Markov chain is said to be irreducible if it is possible to get to any state from any other state. That is, $\mathbb{P}(X_n = j | X_0 = i) > 0$ for all i, j.

Aperiodicity

A Markov chain is said to be aperiodic if the greatest common divisor of $\{n; \mathbb{P}(X_n = j | X_0 = i) > 0\} = 1$, otherwise the Markov chain is said to be periodic.

Recurrent

A Markov chain is said to be recurrent if the probability the chain would return to state i having started from state i is unity for all i. That is, $\mathbb{P}(X_n = i | X_0 = i) = 1$ for all i. A positive recurrent Markov chain is one in which the mean recurrent time is finite.

Stationary distribution

If a Markov chain is ergodic, *i.e.*, irreducible, aperiodic and recurrent, then regardless of what the value of the initial state (X_0) is, the distribution of X_n will converge to a distribution, π. We call π the *stationary* distribution of the chain. That is, for all i

$$\lim_{n\to\infty} \mathbb{P}(X_n = j | X_0 = i) = \pi_j, \tag{1.3.3}$$

where $\sum_i \pi_i p_{ij} = \pi_j$. We shall write this in matrix notation as $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ and these are used extensively in the models we analyzed in this thesis. Note that the descriptions given are largely on *discrete-time* Markov chains (DTMC) with both discrete state space and discrete jump times.

1.3.2 Monte Carlo methods

In this section, we briefly describe the usefulness and limitations of Monte Carlo methods. Monte Carlo methods are primarily employed for the evaluation of integrals of random variables whose integrals do not have analytical solutions.

Suppose that we have a multidimensional random variable X with probability density function, $\pi(x)$, and a function of interest, $\phi(x)$ say, interest may be in the calculation of the expected value of $\phi(x)$, $\mathbb{E}_\pi[\phi(X)]$. This requires the evaluation of the integral

$$\int \phi(x)\pi(x)dx, \tag{1.3.4}$$

for which the analytical solution is not feasible. We can estimate $\mathbb{E}_\pi[\phi(X)]$ by drawing a sequence of values of size n, $X_1, X_2, \ldots, X_n$, such that $\{X_i\}$ are independent and identically distributed according to $\pi(x)$. Then by the *strong law of large numbers* (SLLN), as $n \to \infty$, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) \to \mathbb{E}_\pi[\phi(X)]. \tag{1.3.5}$$

The Monte Carlo estimate in (1.3.5) is unbiased. Also, suppose there exits a finite second central moment, $\sigma_\phi^2(<\infty)$, then by Central Limit Theorem (CLT), we have that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \mathbb{E}_\pi[\phi(X)]\right) \xrightarrow{D} N(0,\sigma_\phi^2), \tag{1.3.6}$$

where $\xrightarrow{D}$ means convergence in distribution. The application on the Monte Carlo methods is possible when we are able to simulate sample from $\pi(x)$. In most realistically complex situations obtaining sample from $\pi(x)$ for Monte Carlo integration is not possible. Markov chain Monte Carlo algorithms allow us to obtain samples from such complex models by constructing a Markov chain whose stationary distribution is our target distribution. In what follows, we now give details on the development of MCMC and its application in Bayesian statistics framework.

1.4 Overview of MCMC algorithms

The idea behind MCMC is to construct a Markov chain whose stationary distribution is the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$, of interest. Unlike the conventional Monte Carlo simulation methods where independent samples are obtained, MCMC algorithm enables us to simulates dependent and auto-correlated samples, $\{\boldsymbol{\theta}^{(t)}\}$, iteratively from the posterior distribution of interest, $\pi(\boldsymbol{\theta}|\mathbf{x})$. Asymptotic results enable us to obtain ergodic average, such as that in (1.3.5), from the dependent

samples realized from the MCMC. Certain mild conditions, see, for example Robert and Casella (1999), ensure that the limiting distribution of the Markov chain $\{\boldsymbol{\theta}^t\}$ is our target posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$, regardless of what the value of the initial state of the chain, $\pi(\boldsymbol{\theta}|\mathbf{x})$, may be.

The use of MCMC dates back to Metropolis (Metropolis et al., 1953) who used it in the context of physics, and it was generalized in statistical context by Hastings (Hastings, 1970). However, the introduction of MCMC into mainstream statistics was by Gelfand and Smith (1990). Since then, there has been a tremendous development in the application of MCMC in all aspects of Bayesian modelling, especially in realistically complex Bayesian models. Gilks et al. (1996), Robert and Casella (1999), Gamerman and Lopes (2006) and Brooks et al. (2011), provide a comprehensive account on the advances of MCMC in statistical methodology. We shall now present MCMC algorithms most relevant to the models considered in this thesis.

1.4.1 Metropolis-Hastings

Suppose we have a d-dimensional posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}) = f(\boldsymbol{\theta})$, and interest is to simulate samples $\{\boldsymbol{\theta}^{(t)}\}$ for inference purposes. A typical Metropolis-Hastings algorithm is given in Algorithm 1 below.

The algorithm defined above can be modified in several ways to enhance its efficiency. Note that $q(.,.)$ denotes the proposal distribution of a given component which needs scaling for improvement. The choice of the proposal distribution is crucial to the success of the MCMC algorithm. We shall see various ways through

Algorithm 1 Metropolis-within-Gibbs algorithm

1. For $t \geq 0$, set the current values of the parameters $\boldsymbol{\theta}^{curr} = \boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_d^{(t)})$.

2. For $i = 1, 2, \ldots, d$, propose θ^{prop} from the proposal density $q_i(\theta_i^{prop} | \theta^{(t)})$.

3. Set $\boldsymbol{\theta}^{prop} = (\theta_i^{prop}, \boldsymbol{\theta}_{-\theta}^{(t)})$, where $\boldsymbol{\theta}_{-\theta}^{(t)}$ is the vector $\boldsymbol{\theta}^{(t)}$ without component i.

4. Compute the probability $\alpha(\theta^{(t)}, \theta^{prop})$, where

5. $$\alpha = \begin{cases} \min\left\{1, \frac{\pi(\boldsymbol{\theta}^{prop}) q(\theta^{prop}, \theta^{(t)})}{\pi(\boldsymbol{\theta}^{curr}) q(\theta^{(t)}, \theta^{prop})}\right\} & \text{if } \pi(\boldsymbol{\theta}^{prop}) q_i(\theta^{(t)}, \theta^{prop}) > 0, \\ 1 & \text{if } \pi(\boldsymbol{\theta}^{curr}) q(\theta^{(t)}, \theta^{prop}) = 0. \end{cases}$$

6. Set $\boldsymbol{\theta}^{curr} = \boldsymbol{\theta}^{prop}$ and $\theta^{(t+1)} = \theta^{prop}$ with probability $\alpha(\theta^{(t)}, \theta^{prop})$. Otherwise set $\theta^{(t+1)} = \theta^{(t)}$.

7. Store the required value.

8. Repeat the steps until sample of the desired size are obtained.

which the proposal density, $q(.,.)$ can be specified for optimal performance of the algorithm.

1.4.2 Choice of proposal distribution

In this section, we shall show how the proposal density $q(.,.)$ can be chosen effectively. Being able to choose a suitable proposal density for a Metropolis-Hastings algorithm enhances the efficiency and the convergence of the algorithm. On the other hand, a bad choice of the proposal density might lead to reducible Markov chains. There are many possible choices for the proposal distribution, but we shall focus on the two most relevant for our purposes.

1.4.3 Independent Sampler

In Independence sampler, the proposal density, $q(\theta^{curr}, \theta^{prop})$, is chosen such that it is independent of the current value,say θ^{curr}. That is;

$$q(\theta^{curr}, \theta^{prop}) = q(\theta^{prop}). \tag{1.4.1}$$

Then the acceptance probability α, is given by;

$$\begin{aligned} \alpha(\theta^{curr}, \theta^{prop}) &= \min\left\{1, \frac{\pi(\theta^{prop})q(\theta^{prop}, \theta^{curr})}{\pi(\theta^{curr})q(\theta^{curr}, \theta^{prop})}\right\} \\ &= \min\left\{1, \frac{\pi(\theta^{prop})q(\theta^{curr})}{\pi(\theta^{curr})q(\theta^{prop})}\right\}. \end{aligned} \tag{1.4.2}$$

The implementation of Independence sampler is straightforward. However, one of the requirements of the Independence sampler to be efficient is that proposal distribution be a good approximation of the posterior distribution.

Random Walk Metropolis

Random Walk Metropolis (RWM) is perhaps the most widely used choice of proposal. The popularity of RMW might be because it is easy and straightforward to implement, it is generally efficient even for high dimensions, it can easily be improved for efficiency and optimality. Suppose we have the proposal density $q(\theta^{curr}, \theta^{prop})$, RMW ensures that the candidate value θ^{prop} is centered on the current value. That is

$$\theta^{prop} = \theta^{curr} + \zeta, \tag{1.4.3}$$

where the random variable ζ has mean of zero and is usually symmetric. A popular case when ζ is Gaussian distributed with zero mean and variance σ_ζ^2 has attracted a lot of attention in recent years. This is because the performance of the RWM algorithm is influenced by the size of σ_ζ^2. To achieve optimality, studies have suggested different ways of scaling and tuning the proposal σ_ζ^2 for the best result, see, for example, Roberts et al. (1997) and Sherlock et al. (2010). In particular, Roberts et al. (1997) suggested that for a muti-parameter density, the σ_ζ^2 value which gives rise to an acceptance rate close to 23.4% is optimal. We employ RWM algorithms extensively in this thesis as it works well for cases where Gibbs sampling can not be used easily. RWM algorithm is as follows.

$$q(\theta^{curr}, \theta^{prop}) \;=\; q(|\theta^{curr} - \theta^{prop}|)$$

The acceptance probability is given by;

$$\begin{aligned} \alpha(\theta^{curr}, \theta^{prop}) &= \min\left\{1, \frac{\pi(\theta^{prop})q(\theta^{prop}, \theta^{curr})}{\pi(\theta^{curr})q(\theta^{curr}, \theta^{prop})}\right\} \\ &= \min\left\{1, \frac{\pi(\theta^{prop})}{\pi(\theta^{curr})}\right\}. \end{aligned} \tag{1.4.4}$$

Observe that the term in $q(.,.)$ canceled out due to symmetry.

1.4.4 Gibbs Sampler

Gibbs sampler, example, Geman and Geman (1984) and Gelfand and Smith (1990) is a special form of Metropolis-Hastings algorithm. It is simple and easy to implement in that it only requires the full conditional distribution of each parameter of interest. In Gibbs sampler, the acceptance probability of the general M-H algorithm is equal to 1. Suppose we have a posterior distribution $\pi(\boldsymbol{\theta}|.)$, with d-dimensional parameters, $\boldsymbol{\theta}$ and interest is on carrying drawing samples for posterior inference. For $i = 1, 2, \ldots, d$, suppose the full conditional density of each of the parameters given everything else, $\pi(\theta_i|.)$, is available in a closed form. Gibbs sampler is used to successively and repeatedly simulating from the conditional distributions of each component of the distribution given the other components as given in Algorithm 2 below.

Algorithm 2 Gibbs Sampler

1. Initialise the parameters, *i.e.*, set $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$
2. For $t = 1, 2, \ldots, n$
   - For $i = 1, 2, \ldots, d$,
   - Draw $\theta_i^{(t)}$ from $\pi(\theta_i|\boldsymbol{\theta}_{-i}^{(t-1)})$, where $\boldsymbol{\theta}_{-i}^{(t-1)}$ is a vector of $\boldsymbol{\theta}^{(t-1)}$ without $\theta_i^{(t-1)}$.
3. Store $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})$
4. Repeat from step 2 down until sample of the desired size are obtained.

The convergence of the Markov chain to the posterior distribution is guaranteed under mild regularity conditions. Therefore, the samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(n)}$, so obtained are believed to come from $\pi(\boldsymbol{\theta}, .)$. Then posterior inference can be carried out to obtain estimates of the quantities of interest.

1.4.5 Hybrid MCMC algorithms

In most practical situations especially in missing data problems, it is most unlikely that each component of the chain will have a conditional distribution in a closed form. In this case, Gibbs sampler can not be implemented. However, we can employ a Metropolis-Hastings algorithm, for example RWM, to obtain sample from such distribution, while we use Gibbs sampler to sample the components standard conditional distribution. In some cases however, Gibbs sampler might not be feasible, then either Independence sampler or RWM could be used in such cases. For example, Neal and Roberts (2006) considered RWM-within-Gibbs algorithm with Gaussian proposal density. In this thesis, we used hybrid MCMC algorithms extensively in for the models analyzed in Chapters 2, 3 and 4.

1.4.6 Burn-in

The samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(n)}$ obtained from MCMC runs are usually highly correlated. In most cases, except in *perfect simulations* the starting value of the Markov chain $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$ is not from the stationary distribution. However as we take longer runs, or for large B, the chain increasingly *forgets* its starting values $\boldsymbol{\theta}^{(0)}$, such that $\boldsymbol{\theta}^{(B)}$ is approximately from the stationary distribution. The idea behind *burn-in* is to discard the first B iterations and then carry out out poste-

rior inference based upon the last $n - B$ samples. That is, burn-in discards the first $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(B)}$ samples and considers the samples, $\boldsymbol{\theta}^{(B+1)}, \boldsymbol{\theta}^{(B+2)}, \ldots, \boldsymbol{\theta}^{(n)}$ as approximately drawn from the stationary distribution.

1.4.7 MCMC Convergence diagnostics

A d-dimensional Markov chain $\{\boldsymbol{\theta}^t\}$ is said to have converged when the stationary distribution of the chain well approximates the target distribution. As mentioned earlier, MCMC draws are correlated and it is a good practice to use the part of the posterior sample obtained after burn-in for convergence examination. Suppose interest is on estimating $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$, at *stationarity*, we have that

$$\hat{f}_n \approx \mathbb{E}_\pi[f(\boldsymbol{\theta})], \tag{1.4.5}$$

where $\hat{f}_n$ is the Monte Carlo estimate

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}_i). \tag{1.4.6}$$

If σ_f^2 exists, the convergence of the chain is guaranteed by the Central Limit Theorem (CLT)

$$\sqrt{n}\left(\hat{f}_n - \mathbb{E}_\pi[f(\boldsymbol{\theta})]\right) \xrightarrow{D} N(0, \sigma_f^2), \tag{1.4.7}$$

where $\xrightarrow{D}$ denotes convergence in distribution. Furthermore, a practical way of assessing the *mixing* of an MCMC algorithm is by calculating the effective number of dependent sample that is equivalent to a single independent sample. This is also

known as the *integrated autocorrelation function* (IAF), see, for example, Neal and Roberts (2005). Let $(\ldots, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots)$ denote sample of $\boldsymbol{\theta}$ obtained from the stationary distribution $\pi(.)$. Then for $t \geq 0$ and for $k > 0$,

$$\begin{aligned} \rho_k &= Corr(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+k}), \\ &= Corr(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k), \\ &= \frac{Cov(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k)}{Var(\boldsymbol{\theta}_k)}, \end{aligned} \tag{1.4.8}$$

is the chain's *autocorrelation function* at lag k. Then

$$C_{int} = 1 + 2\sum_{k=1}^{\infty} \rho_k, \tag{1.4.9}$$

is the integrated autocorrelation function. We can then estimate C_{int} directly from the sample $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n)$ using

$$\hat{C}_{int} = 1 + 2\sum_{k=1}^{T} \hat{\rho}_k, \tag{1.4.10}$$

where

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k} (\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_{t+k} - \hat{\boldsymbol{\theta}})}{\sum_{t=1}^{n} (\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}})^2}. \tag{1.4.11}$$

Notice that

$$\hat{\boldsymbol{\theta}} = \frac{1}{n^*} \sum_{t=1}^{n^*} \boldsymbol{\theta}_t.$$

A challenge in 1.4.10 is how to choose T optimally. Choosing T too little makes the estimate obtained with $\hat{C}_{int}$ unreliable as crucial correlation terms ρ_k might

be left out. On the other hand, a too large T makes hard the distinction between the actual correlation and Monte carlo error.

While monitoring the performance of an MCMC algorithm may be relatively easy, the construction of an efficient MCMC algorithm can be hard. In the construction of MCMC algorithms, the key point is being able to construct a chain whose stationary distribution is the posterior distribution $\pi(.)$ of interest.

1.4.8 Traceplot

Traceplots are the plots of the posterior sample which contain the historical evolution of the chain as it explores the posterior model parameters space. Figure 1.4.1 shows traceplots (up) from the posterior samples of two different MCMC runs each for 1×10^4 iterations. The information made available by Figure 1.4.1 indicates that the MCMC algorithm for the first sample (left) is by far better in performance that the second (right). While the traceplot on the left seems to have converged to the target distribution, the second traceplot is seen to wandering and will potentially continue this way after 1×10^4 iterations for a long time. In some cases, the chain will never converge. This then necessitates some remedial actions such as checking the codes for errors, reparameterisation or using certain optimal approaches.

Autocorrelation function plot

In Bayesian framework, the autocorrelation function (ACF), $Cov(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k)$, defined in (1.4.8) measures the lag k ($k \geq 1$)autocorrelation in a given MCMC sample. The ACF plot therefore shows how correlated (or dependent) a given posterior

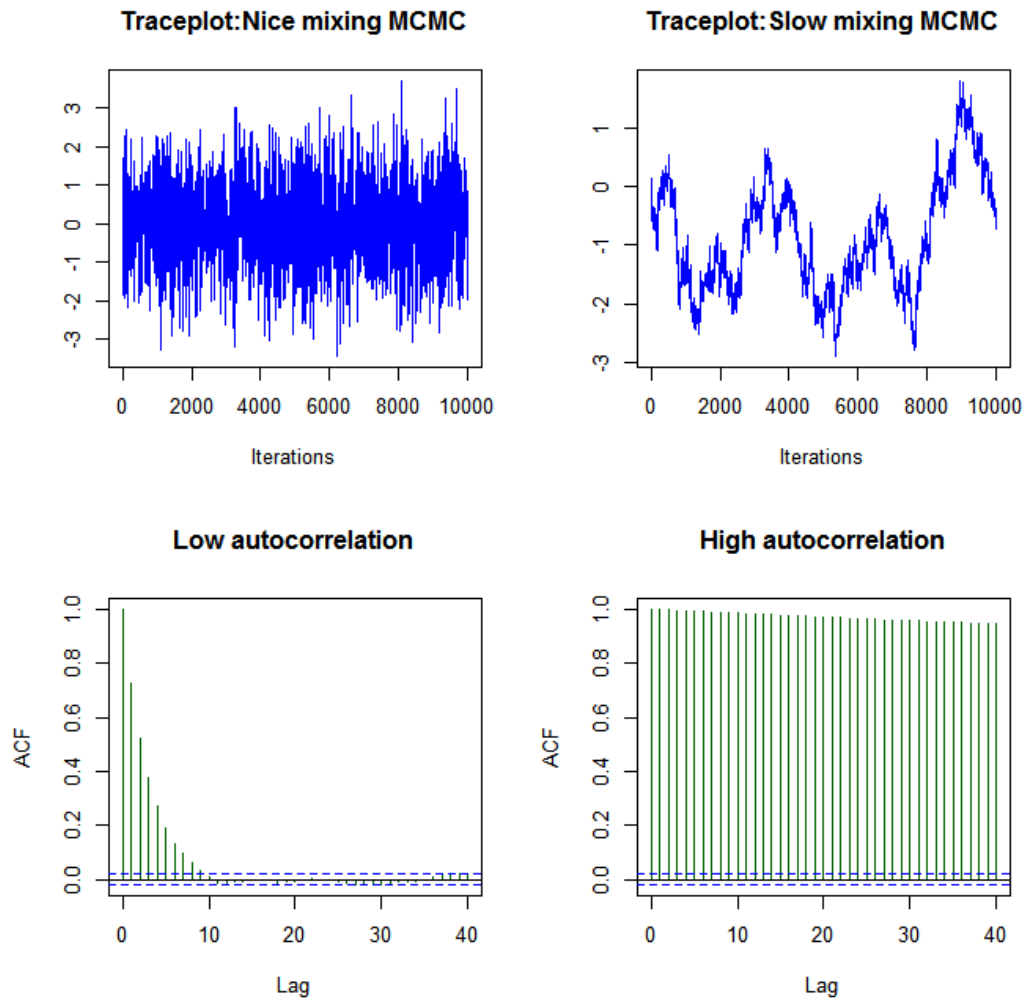

Figure 1.4.1: Traceplots and autocorrelation function (ACF) plots of two different MCMC samples and for two different parameters showing a good mixing chain (left) and a slow mixing chain (right). The green lines on the traceplots are the estimates of their respective posterior means.

samples are. The bottom section of Figure 1.4.1 shows the ACF plots from two different MCMC chains for two different parameters. The first ACF plot (left) shows a fairly small autocorrelation among the elements of the posterior sample. This indicates a good mixing MCMC algorithm. On the other hand, the second ACF plot (right) indicates very high dependence between the MCMC samples. Note that in most cases, the lag-1 autocorrelation approximates lag-k, *i.e.*, $Cov(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) \approx Cov(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. Asymptotically, we expect $Cov(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) \to 0$ as $k \to \infty$. We shall discuss both optimal scaling, shaping and reparameterisation in the section that follows in a bid for an increased efficiency of MCMC algorithms.

1.4.9 MCMC efficiency

As stated earlier, the traceplots and the ACF plots inform us about the mixing the MCMC chains. From both plots, it is easy to see if the MCMC algorithm requires some sort of adjustments for an improved mixing. In this section, we shall outline posterior variance scaling and posterior distribution shaping strategies for RWM algorithm. Later we shall give how to use both strategies adaptively.

Optimal scaling

For optimum performance of the Random Walk Metropolis algorithm, Roberts et al. (1997) suggested using a posterior standard deviation, σ_ζ say, such that about a quarter (or 23.4%) of the proposed moves are accepted. This is the asymptotically optimal acceptance rate for a RWM with Gaussian proposal density. That is, for a d-dimensional chain, propose

$$\boldsymbol{\zeta} \sim N(\mathbf{0}, c\mathbf{I}_d), \tag{1.4.12}$$

where $\mathbf{I}_d$ is a $d \times d$ identity matrix. The key point in 1.4.13 is how to specify the scaling parameter c for optimality. When we choose a too low σ_ζ, the algorithm only explores the posterior distribution inefficiently, moves slowly and accepts most moves. Similarly, if a too large σ_ζ is used, then only a very few proposed jumps are accepted and the chain appears to be stuck at a point for a long time. There is therefore need to have a systemic approach of specifying the scaling parameter c for optimal performance of the chain. For a d-dimensional chain, Roberts and Rosenthal (2001) suggests setting $c = 2.38/d^{1/2}$, and this is found to work well by giving rise to acceptance rate close to the optimal acceptance rate of 0.234, see, for example, Neal and Roberts (2006).

Optimal shaping

Optimal shaping ensures that the algorithm quickly learns the shape of the target distribution thereby enhancing the mixing of the algorithm. A good way to enhance the RWM algorithm so that the chain can learn the shape of the posterior is to propose jumps for the n^{th} iteration from

$$\boldsymbol{\zeta} \sim N(\mathbf{0}, c\Sigma_{n-1}), \tag{1.4.13}$$

where Σ_{n-1} is the posterior variance-covariance matrix from the $n-1$ MCMC run. This allows the algorithm to adjust itself to the shape of the posterior. However starting the chain from points far away from the main posterior mass will force the

chain to learn the shape of unimportant regions of the posterior. Discarding the initial K iterations, say, as burn-in would help to adjust for poor starting points.

Adaptive RWM

From the foregoing, we see that while the scaling parameter ensures optimal acceptance, optimum scaling enables good mixing of the chain. However, we note that the scaling parameter has to be tuned for a number of times before optimality is achieved. This might be burdensome and time consuming. On the other we see that having a wrong start of the chain could lead to inefficient exploration of the posterior. *Adaptive* RWM (see, for example, Haario et al. (1999), Haario et al. (2005), Roberts and Rosenthal (2007) and Roberts and Rosenthal (2010)) steps ensure that both the optimal scaling and optimal shaping are incorporated into the algorithm, which also enables automatic tuning of the algorithm. We summarize this section with the following adaptive RWM steps:

1. Start the chain with a sensible value of σ_ζ which gives an acceptance rate close to 23.4% and run a fairly small number of iterations, long enough for a reasonable set of estimates to be obtained.

2. Then run a longer chain. This allows the chain to drop any errors inherited from the first early iterations. It is expected that changes in the transition kernel diminishes with the number of iterations, see, Roberts and Rosenthal (2007).

3. Obtain the posterior variance-covariance matrix, Σ.

4. For the n^{th} run, propose jumps using the proposal variance-covariance matrix $\Sigma_n = c\Sigma$, where $c = 2.38^2/d$.

5. When a proposed jump is accepted, increase the size of the move by setting Σ_n equal to

$$\left(1 + \frac{0.03}{\sqrt{J}}\right)\Sigma_n \ , \tag{1.4.14}$$

6. When a proposed move is rejected, decrease the jump size and set Σ_n equal to

$$\left(1 - \frac{0.01}{\sqrt{J}}\right)\Sigma_n \ , \tag{1.4.15}$$

7. Repeat steps 3 and 4 and use the Σ^* obtained for the main MCMC runs,

for $J = 1, 2, \ldots, B$, where B is the size of the chain discarded as burn-in, for example, Xiang and Neal (2014).

1.4.10 Data Augmentation

In this section, we give a brief overview of data augmentation with focus on its application in Bayesian statistics via Markov Chain Monte Carlo (MCMC) algorithms.

The name *data augmentation* (DA) originates with Tanner and Wong (1987) who exploited the underlying techniques to obtain samples in a straightforward manner from the posterior distribution of a stochastic model. However, this technique was first applied to deterministic problems by Dempster, Laird and Rubin, (Dempster et al., 1977), for likelihood function maximization using the well known E-M (Expectation-Maximization) algorithm, and has also been applied in the context of Physics to improve the speed of iterative samplings, see, for example, Swendsen

and Wang (1987). Data augmentation offers a natural way of simulating iteratively from very complex models and this makes it a powerful tool in the area of Bayesian Statistics, especially in missing data problems. In a Bayesian framework, data augmentation is usually implemented via Markov Chain Monte Carlo (MCMC) algorithms. There are several natural and social processes which give rise to incomplete data. In an infectious disease process, for example, we only observe times at which symptoms begin to manifest (for symptomatic diseases) and the time when an individual recovers, the actual times of infection are not usually observed. A DA scheme allows the unobserved infection times to be imputed as extra information using appropriate MCMC algorithms, see, for example, Neal and Roberts (2005). See also, Demiris and O'Neill (2005a) and Cauchemez et al. (2004) for applications of data augmentation in the analysis of infectious disease data.

Let $\mathbf{y}$ denote the partially observed data and $\boldsymbol{\theta}$ the parameters of interest. Note that only a subset of the data $\mathbf{y}$ are observed and in most practical situations, it is impossible to evaluate the likelihood function of the observed data $\mathbf{y}$ given the parameters $\boldsymbol{\theta}$, $\pi(\mathbf{y}|\boldsymbol{\theta})$. However, given an additional information $\mathbf{z}$, the augmented likelihood function, $\pi(\mathbf{x} = (\mathbf{z}, \mathbf{y})|\boldsymbol{\theta}, \mathbf{y})$ becomes tractable, where $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ are the full (or the augmented data). Then the probability of observing the augmented data given the observed data $\mathbf{y}$ and the parameters satisfies

$$\pi(\mathbf{x} = (\mathbf{z}, \mathbf{y})|\boldsymbol{\theta}, \mathbf{y}) \quad \propto \quad \pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\theta}).$$

We need to construct a Markov Chain whose stationary distribution is our target distribution, the joint posterior $\pi(\boldsymbol{\theta}, \mathrm{x}|\mathbf{y})$, by drawing the following samples according to the steps described in Algorithm 3

Algorithm 3 Data Augmentation

1. Initialize the parameter $\boldsymbol{\theta}^{(0)} \in \Theta$,
2. Initialize the augmented data $\mathbf{x}^{(0)}$,
3. Construct the Markov chain $\left\{(\boldsymbol{\theta}^{(r)}, \mathbf{x}^{(r)}); t \geq 1\right\}$ as follows,
   - draw $\mathbf{z}^{(r+1)} \sim \pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(r)})$,
   - draw $\boldsymbol{\theta}^{(r+1)} \sim \pi(\boldsymbol{\theta}|\mathbf{x}^{(r+1)} = (\mathbf{y}, \mathbf{z}^{(r+1)})$.

The sampled values of the parameters $\boldsymbol{\theta}^r; r \geq 1$ are normally stored while $\mathbf{z}^r$ may be stored for further use or discarded as nuisance parameters not needed.

Here, the key challenge is how to efficiently choose the auxiliary variable $\mathbf{z}$ so as to be consistent with the observed data $\mathbf{y}$. This thesis utilizes the concept of data augmentation extensively as we shall see in the next chapter.

1.5 Epidemic models

In this section, we give an overview of epidemic modelling and a brief history of the development of mathematical models for infectious diseases.

1.5.1 Historical Background

Epidemic models are developed to capture the dynamics of infectious diseases. An excellent summary of the early history of infectious disease modelling is given in Bailey (1975) . Good reviews of more recent developments on epidemic models are provided in Isham (2005) and Greenwood and Gordillo (2009). The first attempt

on mathematical modelling of infectious diseases is taken to be a paper presented by Daniel Bernoulli (Bernoulli, 1760) account of which is found in Daley and Gani (1999). Other works in infectious disease that follow after over 100 years of the work by Bernoulli are due to Hamer (1906), Ross (e.g, Ross (1911)), McKendrick (1926), Kermack and McKendrick (1927) amongst others. The paper by (Bernoulli, 1760) discusses the use of inoculation to prevent smallpox. Hamer (1906) proposed a discrete time model which assumes that the probability of an infection in the next time period of time was proportional to the product of the number of infectives (infected and infectious individuals) and susceptibles. Among the early epidemic models studied are stochastic epidemic models (McKendrick, 1926), deterministic general epidemic model(Kermack and McKendrick, 1927). Reed-Frost model, a discrete-time stochastic epidemic model was presented in lectures by Reed and Frost in 1928 (Daley and Gani, 1999). Continuous-time stochastic epidemic model of the SIR type was studied by Bartlett (1949) and there has been an increased volume of literature ever since.

1.6 The General Stochastic Epidemic Model

In this section we present the model which forms the basis of the models studied in this thesis. The *general stochastic epidemic* (GSE) model is the most well studied stochastic epidemic model. We use the SIR (susceptible $\rightarrow$ infected $\rightarrow$ removed) compartmental model (Figure 1.6.1) to describe the GSE. The model assumptions are as follows. A closed population of N individuals, with a initial infectives and $N - a$ initial susceptibles. The population is said to be closed in

that no births, deaths, emigration or immigration are allowed during the course of the epidemic. The closed population assumption makes sense for infectious disease which spread so fast within a very short period of time, for example chickenpox, in that the population is not likely to witness major demographic changes during this time. At every given point in time, an individual is classified as a susceptible if the individual is susceptible to the disease (can be infected) and therefore is said to be in the susceptible state (or state S). A non-susceptible individual is either infected and infectious or has recovered. Infected individual is said to be in the infected state (or state I), while a recovered individual is said to be in the removed state (or state R). Recovery from such diseases confers immunity so that a recovered individual becomes immune and cannot be re-infected with the disease again. Therefore, once recovered, the individual plays no further role in the infectious process. We call I the infectious period which is the difference between the time of the individual has been confirmed to have recovered from the disease and the time of actual infection. Infectious periods of different individuals are assumed to be independent and identically distributed according to some random variable $\mathcal{I}$, where the distribution of $\mathcal{I}$ can be arbitrary but specified. Several distributions of $\mathcal{I}$ have been studied. For example, O'Neill and Becker (2001) assumes a gamma distributed infectious period, while Streftaris and Gibson (2004) assumes a Weibull distributed infectious period. However, the general stochastic epidemic model assumes an exponentially distributed infectious period (Bailey, 1975). The exponential distribution is not necessarily biologically plausible for many diseases, but is mathematically attractive in that it makes the epidemic process to be Markov with the memoryless property that given the present state, the

future state is independent of the previous states. This memoryless property is widely exploited in this thesis.

Individuals in the population are assumed to mix homogeneously so that whilst infectious, an infective makes infectious contacts with an individual chosen uniformly at random from the entire population at the points of independent Poisson processes at rate $\beta > 0$. Only such contacts between an infective and a susceptible results in an infection. Infected individual immediately becomes infectious and can pass on the contagion to other susceptibles. At the end of its infectious period $I \sim \exp(\gamma)$ with parameter $\gamma > 0$. The epidemic goes on until there are no more infectives in the population. The model described here dates back to (Kermack and McKendrick, 1927) who studied the deterministic aspect of the model, while Bartlett (1949) studied continuous time stochastic SIR epidemic model. Throughout this thesis, the focus is on stochastic epidemic models. In Section 1.7 we shall describe a class of model in which the homogeneously mixing population assumption is relaxed to allow some heterogeneity.

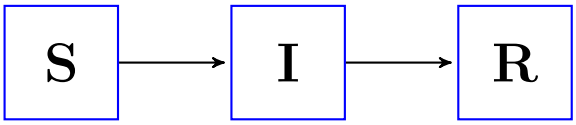

Figure 1.6.1: Transition states of an individual in SIR model. At time t the population size $N = \mathrm{S}(t) + \mathrm{I}(t) + \mathrm{R}(t)$, where $\mathrm{S}(t)$, $\mathrm{I}(t)$ and $\mathrm{R}(t)$ are the number of susceptibles, infectives and removed individuals at time t.

1.6.1 SIS Stochastic Epidemic Model

In this section, we describe the advances and features of the SIS (susceptible $\rightarrow$ infective $\rightarrow$ susceptible) stochastic epidemic model.

We consider the SIS stochastic epidemic model based upon the assumptions of the general stochastic epidemic described above. We note that the SIS model does not confer immunity after recovery so that the individual immediately returns to the susceptible state and can be reinfected. Therefore, only two possible state transitions are allowed: from susceptible state to infective state (or S $\rightarrow$ I) and from the infective state to susceptible state (or I $\rightarrow$ S), see Figure 1.6.2. As the individual does not become immune after recovery , there is no removed state. This means that such an SIS disease will establish itself within a finite population and become endemic for a long period of time before eventually going extinct. The SIS epidemic model has been described as the simplest epidemic model that exhibits endemic behavior, see, for example, Ball (1999), Neal (2006) and Neal (2014). Examples of diseases that follow the SIS model are *gonorrhea*, *pneumococcus* and *tuberculosis* in humans, and most tick-borne diseases in animals, in that an individual who recovers from such diseases does not become immune and therefore can be reinfected. Stochastic SIS epidemic model can potentially provide answers to some important public health questions such as; Will the disease ever become endemic? If the disease ever becomes endemic, what level of endemicity will it attain? For how long will the disease remain endemic before going extinct? There has been a considerable efforts in developing SIS epidemic models. Neal (2014) studies endemic behaviour of SIS epidemics in a finite population with general infectious period distribution using branching process with immigration approximation. SIS epidemics among a community of households have also been studied, example, Britton and Neal (2010) studies stochastic SIS epidemic, while Ball (1999) and Neal (2006) studied both deterministic and stochastic household

SIS epidemics. Economou et al. (2015) studies a stochastic SIS epidemic model with heterogeneous contacts. Hethcote and van den Driessche (1995) and Greenhalgh et al. (2016) studied SIS epidemic within a population with variable size. Neal and Huang (2015) studies stochastic SIS epidemic for interacting strains of Human Papillomavirus (HPV) amongst an MSM (men who have sex with men) community. Gao et al. (2016) studies SIS epidemic for coinfection and cotransmission of two diseases spreading through a single host population. In this thesis, our main focus is on the development of Bayesian inference methods for the stochastic SIS epidemics. In Section 1.6.4 we discuss the development of inference methods for epidemic models including those developed for the stochastic SIS epidemic model.

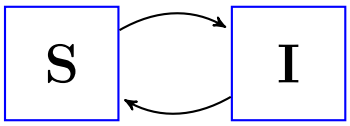

Figure 1.6.2: Transition states of an individual in SIS model. At time t the population size $N = \mathrm{S}(t) + \mathrm{I}(t)$, where $\mathrm{S}(t)$ and $\mathrm{I}(t)$ are the number of susceptibles and infectives at time t.

1.6.2 Basic Reproduction number

In epidemic modelling a key epidemiological quantity of interest is the basic reproduction number $R_0 = \beta/\gamma$, which is defined as the number of new secondary cases of infection from an initially infectious individual in a completely susceptible population. When $R_0 > 1$ there is high probability that a major outbreak will occur.

On the other hand, when $R_0 < 1$, major outbreak does not occur. In addition, the fraction of the population necessary to be vaccinated to be sure that no major outbreak would occur, also known as the *critical vaccination coverage,* v_c, can be obtained with the knowledge of the basic reproduction number as follows

$$v_c = 1 - \frac{1}{R_0}\ , \tag{1.6.1}$$

for example, Britton (2009).

1.6.3 Model setup

Poisson processes

Given the counting process $X = \{X(t) : t \geq 0\}$ taking values from the discrete state space $\mathcal{S} = \{0, 1, 2, \ldots\}$, then X is a Poisson process with intensity $\beta > 0$ if

1. $X(0) = 0$.

2. $$\mathbb{P}(X(t + \Delta t) = u + v | X(t) = u) = \begin{cases} \beta\Delta t + o(\Delta t) & \text{if } v = 1, \\ o(\Delta t) & \text{if } v > 1. \\ 1 - \beta\Delta t + o(\Delta t) & \text{if } v = 0. \end{cases}$$

3. If $s < t$, then

   - $X(s) \leq X(t)$

   - $X(t) - X(s)$ is independent of the events that occurred within $[0, s]$.

There is a wide range of applications of the Poisson processes in sciences. Given an interval $(0, t]$, for $t \geq 0$, let $X(t)$ denote the number of events that occurred

within the interval $(0, t]$ with intensity $\beta > 0$. Then the random variable $X(t)$ is Poisson distributed with parameter βt. That is

$$\mathbb{P}(X(t) = x) = \frac{(\beta t)^x \ \exp(-\beta t)}{x!}. \tag{1.6.2}$$

Therefore, the probability of no occurrence of an event in $(0, t]$ is given by

$$\mathbb{P}(X(t) = 0) = \exp(-\beta t), \tag{1.6.3}$$

with the complementary probability

$$\mathbb{P}(X(t) > 0) = 1 - \exp\left(-\beta t\right). \tag{1.6.4}$$

Let $X(t)$ denote the number of infectives at time t. Given an $|\mathcal{S}| \times |\mathcal{S}|$ transition rate or *generator* matrix $\mathbf{G} = (g_{u,v})$, we describe the epidemic process $X = \{X(t) : t \geq 0\}$ in terms of the continuous -time Markov chain using the following transition rates:

- $u \to u + 1$ if an event is infection.
- $u \to u - 1$ if an event is recovery.
- $u \to u$ if nothing happens.

where in this case $u \in \mathcal{S}$ is the number of infectives at a point in time, the row entries of the generator matrix sum to zero, *i.e.*, $\sum_v g_{uv} = 0$ for all u or $\mathbf{G1}' = \mathbf{0}'$, where 1 and $\mathbf{0}$ are row vectors of ones and zeros. Then following Grimmett and Stirzaker (2001), pages $256 - 260$, for $t \geq 0$, we define the infinitesimal transition probability matrix $\mathbf{Q}_t = (q_{u,v}(t))$ according to

$$\mathbf{Q}_t = \exp(t\mathbf{G}), \tag{1.6.5}$$

where

$$q_{u,v} = \mathbb{P}\big(X(t+\Delta t) = v | X(t) = u\big). \tag{1.6.6}$$

Here, interest is particularly on what happens within the interval $[t, t+\Delta t)$. Δt is chosen to be sufficiently small so that only one event is allowed to happen within the interval $[t, t+\Delta t)$, *i.e*, either infection, recovery or nothing happens. The Q-matrix is stochastic in that its entries are non-negative and its row entries sum to unity. In addition, at $t = 0$, the transition matrix returns an $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix, *i.e.*, $\mathbf{Q}_0 = \mathbf{I}$. For more rigorous details on Markov chains, their properties and applications, see Chapter 6 of Grimmett and Stirzaker (2001). In Chapter 2 of this thesis, we shall discuss in details the construction of the generator (rate) matrix and the calculation of the corresponding transition probability matrices for the models considered.

1.6.4 Inference on epidemic models

In this section, we give an overview on inference methods for epidemic models. First, we present some recent developments on methods of inference for epidemic models with focus on Bayesian inference methods. We shall also outline the procedures for inferring model parameters in a Bayesian inference framework.

A number of approaches has been developed for inferring the parameters of infectious disease models. In classical statistical inference approach, maximum-likelihood (ML) estimation via an Expectation-Maximization (EM) algorithm has been used to analyze infectious disease data by treating the missing data as parameters to be estimated (Becker, 1993). However, in most realistic situations, the evaluation of the E-step becomes very difficult. In recent years, following the availability on fast computing machines and the huge advances of Markov Chain Monte Carlo (MCMC) methods, Bayesian inference method using MCMC has been successfully used for Bayesian inference on infectious disease models even for very complex models, see, for example, Gibson (1997),Gibson and Renshaw (1998) and O'Neill and Roberts (1999) for some preliminary works on this,Marion et al. (2003), Streftaris and Gibson (2004), Neal and Roberts (2005) and Neal and Xiang (2017).

1.7 Household-based epidemic models

One of the earliest household models developed was due to Longini and Koopman (1982). The model by Longini and Koopman (1982) considers individuals in a population partitioned into independent households. Individuals in a given household can be infected by their family members as well as by other members of the population. The model also assumes that the transmission processes within a given household does not depend on the transmission dynamics of the entire community. Addy et al. (1991) studied generalized stochastic models involving a population partitioned into households which was applied on serologic data from two influenza epidemics. Becker and Dietz (1995) consider a household model for

highly infectious diseases, such as smallpox. They assume that once a member of a given household contracts infection, then every member of the household gets infected. The models were used to evaluate various vaccination strategies taking the household structures into account. Findings from Becker and Dietz (1995) suggested the use of different vaccination coverage for different household structures. For example, it was observed that it is better to immunize randomly selected individuals when the households are of equal size, while it is better to immunize all members of large households when the sizes of the households are unequal. Examples of other works on the development of household epidemic models are due to House and Keeling (2008) and Goldstein et al. (2009).

Many of the results obtained for household disease models are asymptotic as $n \to \infty$.

1.7.1 Household-based epidemics with two-levels mixing

Ball et al. (1997) introduced *two levels mixing* in household-based epidemic model and since then household models for the spread of infectious diseases have received a considerable attention, see, for example Ball and Neal (2004), Ball and Lyne (2001), Neal (2006), Britton and Neal (2010) and Longini et al. (2005).

Suppose we have n mutually exclusive households each of size h so that that the population size is $N = nh$. The *two levels mixing* household-based epidemic model of Ball et al. (1997) assumes that an infectious individual makes *global* contact and *local* contacts. A global contact is made with an individual chosen uniformly at random from the $N(= nh)$ population at rate $\lambda > 0$. Similarly, a local contact is

made with an individual chosen uniformly at random from the n individuals in the infective's household at a typically larger rate $h\beta > 0$. Therefore, the individual to individual global and local infection rates are λ/N and β. Contacts are made at the points of mutually independent Poisson processes. The infectious period of different infectives are independent and identically distributed according to a random variable I with an arbitrary, but specified distribution. This model can also be generalized for various household structures including when the household sizes are unequal. Let $h = 1, 2, \ldots,$ denote the possible sizes of the households in the population. Let n_h denote the number of households of size h so that $n = \sum_{n=1} h_n$ and $N = \sum_{n=1}^{\infty} nh_n$ are the total number of households and the population size respectively. The models we analyse in this thesis are based on the concept of two levels mixing epidemics.

1.7.2 Need for Household based epidemic models

There are a number of reasons why models are developed. Developing models which capture the basic household structures of a population is needed to effectively analyse infectious disease data emanating from household level. According to Ball et al. (2015), household is the most crucial aspect of human society that can affect disease transmission. Contacts among individuals of a given household are longer and more frequent than with members of another household. Individuals become ill after being infected by the members of their household or by other members of the community often stay at home making regular contact with their household members. Several control strategies are implemented and monitored on

household levels. A plausible household-based epidemic model will potentially provide appropriate answers to certain public health questions, such as, 'who infects whom?'. Household models take into account heterogeneity in population behavior, which is a key determinant among the factors that determine the occurrence of major epidemic outbreak, how fast it spreads if it does occur and the number of individuals ultimately infected during the course of the epidemic. Moreover, household-based infectious disease models can suggest applicable control strategies given the household structures, for example in the administration of vaccines (Becker and Dietz, 1995). Other measures such as contact tracing and isolation of infected individuals (if necessary) are readily applicable through households. Therefore, there is need to develop epidemic model which accurately captures the key transmission mechanisms of infectious diseases at household levels. In this thesis, we develop Bayesian inference methods on such models which capture the inherent structure in both human and animal populations, where in this case a household could be childcare facilities, workplaces, dwelling places for humans or animal holdings (farms). Our main focus is the so-called two levels mixing model of Ball et al. (1997).

1.7.3 Inference on household models

Despite the advances so far recorded by household based endemic models, only a few works are channeled on inference. Drawing inference from household based model is usually very complicated due to the high computational complexity involved. Household epidemic data are often very highly dependent especially for

temporal data where information is obtained from the same group of individuals over time. Also, as with most epidemic data, some processes are unobserved, for example, actual infection time, thereby giving rise to outbreak data that are only partially observed. Although the problem of high dependence among household outbreak data can be minimized by using simplified assumptions. For example assuming that the households are independent households (Addy et al., 1991) makes it possible for the likelihood function $\pi(\mathbf{x}|\boldsymbol{\theta})$ to be expressed as the product of likelihood function of all the n households. That is, the dependence between households is broken by independence households assumption then the likelihood function of the data given the parameters is

$$\begin{aligned}
\pi(\mathbf{x}|\boldsymbol{\theta}) &= \pi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n|\boldsymbol{\theta}) \\
&= \pi(\mathbf{x}_1|\boldsymbol{\theta}) \times \pi(\mathbf{x}_2|\boldsymbol{\theta}) \times \ldots \times \pi(\mathbf{x}_n|\boldsymbol{\theta}) \quad (\textit{independence}) \\
&= \prod_{i=1}^{n} \pi(\mathbf{x}_i|\boldsymbol{\theta})
\end{aligned}$$

On the other hand, the problem of incomplete data is minimized by designing appropriate data imputation strategies, see,for example, O'Neill (2009) and Neal and Kypraios (2015). Most available literature on household based epidemic use the Markov chain Monte Carlo (MCMC) algorithms to sample from the target distribution, example, Britton and O'Neill (2002), Cauchemez et al. (2004), O'Neill et al. (2000) and O'Neill (2009). In most practical situations, the likelihood function in (1.7.1) is very complicated that the posterior distribution can never be available in a closed form no matter what the choice of the prior distribution may

be. Other likelihood-free methods of inference on household epidemics have been developed, see, for example, Neal (2012).

In this thesis, we shall focus on developing inference methods on household-based SIS epidemics with respect to two different data forms which we shall introduce in Chapter 2.

1.8 Contributions of the Thesis

In this section, we outline the major contributions of this thesis to literature.

Despite the growing popularity of household epidemic models (see, for example, Longini and Koopman (1982), Ball and Neal (2004) and Neal (2006)), there still exist several challenges which need to be overcome to allow widespread use such models. For example, Ball et al. (2015) outlines the following seven challenges affecting the progress of household epidemic model:

1. The need to clarify the usefulness and limitations of systems of weakly coupled large sub-populations in infectious disease modelling.

2. Development of theory for household-based endemic models.

3. Generalization of the framework of household models to more complex human social structures.

4. Incorporation of spatial element into household epidemic models.

5. The need to develop methods of drawing inference for household data on emerging phase of epidemics.

6. Development of computationally efficient methods for calculating principal epidemiological quantities.

7. The need to integrate within-host and between-host dynamics into household models.

Consequently, in this thesis we seek to address challenges number 3, 4 and 7 directly, while challenges 2 and 6 are indirectly addressed. First, the model developed in Chapter 2 extends the GSE model to cases involving structured populations. Specifically, we developed Bayesian inference methods for stochastic household-based models of the SIS type, where the individuals are allowed to mix heterogeneously making both local and global contacts. Two major data forms were separately considered- individual based data (IBD) and the aggregate-based data (ABD). We successfully developed and applied MCMC algorithms for both IBD- and ABD- type infectious disease data. This approach was applicable for both when data are fully observed and when the proportion of the missing data is up to 90%. This is the first attempt in developing Bayesian inference method using MCMC for the SIS-type household-based epidemic model using the IBD and ABD data forms. The methods developed can be extended and applied to a wide range of problems.

Second, the model considered in Chapter 2 assumes a constant population size. This assumption makes sense for infections that spread very fast during its course and it is very unlikely that the population would experience any significant demographic changes. However, the constant population assumption may be unrealistic for most endemic models. There is therefore need to develop inference methods

taking into account the varying population size over time. Consequently, in Chapter 3 of this thesis, we develop Bayesian inference methods using MCMC for the analysis of SIS-type open-population household-based endemic model. Also, here the MCMC algorithms were developed based upon the two data forms considered - IBD and ABD. The models were successfully analysed and applied using the inference methods developed. Again, this is the first attempt to develop this class of inference methods for open-population household based models of the SIS-type in Bayesian framework.

Furthermore, we extend both the closed population model and the open-population model to allow for the incorporation of the spatial element. The assumption here is that whilst local contacts are made locally at a constant rate $\beta > 0$, the global contact rate $\lambda > 0$ depends on the spatial locations of the individual households. The incorporation of the spatial introduced more complexity to the model. We successfully developed Bayesian inference method using MCMC for the analysis of such household-based models with spatially varying global contact rates. This is also novel at least for the models we considered.

Finally, this thesis seeks to develop inference approach for household-based endemic models with interacting diseases. This model is considered in Chapter 4 of this thesis. The MCMC algorithms developed considered both the IBD and the ABD data forms. Model complexity grew with household size. Using appropriate parameterisations, we develop MCMC algorithms for the analysis of models of this type. In both cases, we considered when data are fully observed and when a given

proportion of the data is missing. However, our main focus is on missing data cases which is usually the form of most infectious disease data and especially for most temporal data associated with endemic diseases. The methods developed were successfully applied to both simulated data set and a real life infectious disease data on Tanzania cattle. This is the first attempt to model and analyse infectious disease data in household setting for the class of models we considered.

1.9 Structure of the Thesis

In this section, we give the outline of this thesis. This thesis contains five chapters with three main chapters- 2, 3 and 4. Chapter 1 is the introductory chapter where we discussed all relevant historical and theoretical basis of our models. In Chapter 2, we introduce a two levels mixing stochastic SIS epidemic models among a community divided into non-overlapping households. Two different data forms are considered: The aggregate-based data (which holds information only on the infectiousness of a given household at a point in time), and the individual-based data, which is more informative. Later we shall see how we discovered the strength and weakness of each data type using rigorous sensitivity analyses. In Chapter 3, we introduce stochastic SIS household model with individuals making both within group and between groups contacts. Here the population is open in that we allow changing population sizes over time. Also, in Chapter 3, we incorporate the spatial elements into the models taking into their separation distances. In Chapter 4, we consider the effects of getting infected with multiple diseases. The models developed are illustrated using simulated and real life data sets. Finally in Chapter 5, we give a review of what we have done in this thesis, outline the

limitations of the study and draw conclusions on our findings.

Chapter 2

Closed Population SIS Household Model

This chapter is mainly concerned with the development of Bayesian inference methods for stochastic SIS epidemic models in a closed population partitioned into households. We consider two major outbreak data for endemic diseases, namely, the individual-based data and the aggregate-based data. We develop easy-to-implement MCMC algorithms which are found to work well using an extensive simulation study. The model and the Bayesian inference methods developed in this chapter form the basis of the model and Bayesian inference framework developed in Chapter 3, which allows for changing population size over time as well as spatially varying risk of infection. In Chapter 4, this would be extended to allow for interacting infectious diseases.

2.1 Introduction

Mathematical models for the spread of epidemics in a population divided into smaller groups have received increased attention in recent years, see, for example, Addy et al. (1991), Becker and Dietz (1995), Ball et al. (1997), Ball (1999), O'Neill et al. (2000), Ball and Lyne (2001), Ball and Neal (2004), and Neal (2006), Demiris and O'Neill (2005a), Cauchemez et al. (2004), Blake et al. (2009), Demiris and O'Neill (2005b), O'Neill (2009), Neal (2012). An example of such models is the famous household model in which the population of interest is partitioned into households or farmsteads. Ball et al. (1997) developed a two level mixing household epidemic model which allows individuals to mix both locally and globally at given rates. We note that the majority of the works on household epidemic models have focused on the SIR (susceptible $\rightarrow$ Infected $\rightarrow$ Removed) epidemic model in which an infective acquires immunity to further infection upon recovery and therefore can not be reinfected following recovery. The individual is said to be removed and therefore no longer takes part in the infectious process. On the contrary, in this chapter, interest is on endemic models in which a given individual can be infected and reinfected following recovery, a multiple number of times. This is the well known SIS (susceptible $\rightarrow$ infected $\rightarrow$ susceptible) epidemic model in which an individual is either a susceptible or an infective at any given point in time. In addition, in order to take into account the randomness which characterizes the dynamics of the transmission of infectious diseases, a *stochastic* SIS household epidemic model is considered.

The SIS epidemic model which dates back to Ross (Ross, 1915) has been adjudged the simplest epidemic model with endemic behavior. Of all the examples given

above, only Ball (1999), Blake et al. (2009) and Neal (2006) considered SIS epidemics within a closed population partitioned into households. Ball (1999) and Neal (2006) study the asymptotic behavior of both the deterministic and stochastic SIS household epidemic model as the total number of households tends to infinity, *i.e.*, as $N \to \infty$. Neal (2006) proved a law of large numbers for the convergence to deterministic limits of the mean number of infectives in a given household and also derived a Gaussian limit process for the fluctuations of the stochastic process. Ball (1999) used branching process approximations to show that for an SIS epidemic model, there is a non-zero probability that the epidemic will take off if and only if the threshold parameter R_*, is greater than unity, *i.e.* $R_* > 1$.

Other works such as those by O'Neill et al. (2000), Demiris and O'Neill (2005a), Demiris and O'Neill (2005b), O'Neill (2009) and Neal (2012) focus on the development of Bayesian inference methods for structured population models. Neal (2012) develops an efficient likelihood-free Bayesian computation approach for household epidemics, namely, coupled Approximate Bayesian computation (ABC). O'Neill et al. (2000), Demiris and O'Neill (2005a), Demiris and O'Neill (2005b) and O'Neill (2009) considered the use of Markov chain Monte Carlo (MCMC) algorithms for the analysis of household epidemic data. Demiris and O'Neill (2005a) develops Bayesian inference methods for two levels mixing SIR household model, while Demiris and O'Neill (2005b) focuses on multitype SIR epidemic model (Ball and Lyne, 2001), in which the individuals of a given household are classified into k types, say. Also, Blake et al. (2009) used methods of maximum likelihood to estimate the model parameters of household and community transmission of Ocular *Chlamydia trachomatis*. Several infectious diseases such as sexually transmitted

infections (STIs) follow the SIS epidemic model and there is need to accurately define the transmission processes through which data from such infectious processes are generated. The ultimate aim is to be able to analyse such a model. Noting the crucial role played by household structures in the transmission of infectious diseases, there is also need to develop SIS epidemic model which takes household structures into account.

As far as we are aware, there has not been any work on the development of Bayesian inference methods implemented via MCMC framework for stochastic SIS household epidemic models with respect to the two commonest forms of endemic disease data we consider. Therefore, we seek to fill this gap in literature by developing Bayesian inference methods which allows straightforward estimation of the principal parameters of stochastic SIS household epidemic models primarily using Markov chain Monte Carlo (MCMC) algorithms.

We consider two most prevalent forms of household endemic data (individual-based and aggregate-based data) which keep track of the infectious activities going on in a given household over a set of observation time points. Therefore, the data are longitudinal, temporal data. We note that Cauchemez et al. (2004) also used a longitudinal data to study the transmission of influenza among a community of households implemented via MCMC framework. However, whilst our model is an endemic SIS model, Cauchemez et al. (2004) studies the SIR type in which recovery from the disease confers immunity to further infections. The two data forms provide different amounts of information on the spread of the disease of interest. While one contains only information on the number of infectives in a given household at a point in time, the second data form is more informative and keeps

track of the infectious activities of a particular individual over the set of observation time points. This implies that with the latter we are able to know which individual infected at a point in time as well as the total number of infectives in a given household at that point in time. The full description of these data forms including when they are assumed to be only partially observed is given in Section 2.2.

We follow Demiris and O'Neill (2005a) and assume that the households are independent. This assumption makes sense in that we consider household-to-household infection to be negligible, and also allows simple specification of the likelihood function. In this chapter, the key focus is on the development of Bayesian inference methods for the estimation of the two infection rates, namely, β and λ. Given the observed data $\mathbf{x}$, we develop easy-to-use MCMC algorithms that allow the samples to be drawn from the posterior distribution of the parameters given data, $\pi(\boldsymbol{\theta}|\mathbf{x})$, where $\boldsymbol{\theta} = (\lambda, \beta, \gamma)$ and $\gamma > 0$ is the rate at which an infective recovers and immediately becomes susceptible again and can be reinfected. We also develop flexible data augmentation schema in MCMC framework which eases the analysis of the infectious disease data when the data are only partially observed.

The rest of this chapter is structured as follows. In Section 2.2, we describe the two basic forms of household SIS epidemic data: Individual-based data (IBD) and Aggregate-based data (ABD). In Section 2.3, we outline the model setup and describe the construction of the infinitesimal rate matrix (G-matrix), and the calculation of the transition probability matrix (Q-matrix). We start by assuming that the data are fully observed at a set of discrete time points. We then relax

this to allow the data to be only partially observed at the observation time points.

In Section 2.4, we develop Bayesian inference approach for household-based SIS epidemics with respect to ABD and IBD using MCMC algorithms. This is divided into two parts. The first part develops MCMC algorithms for a fully observed household SIS epidemic data. In the second part, we develop *data augmentation*, see, Tanner and Wong (1987) schema using MCMC algorithms for the analysis of partially observed epidemics.

In Section 2.5, a comprehensive simulation study is carried out. The aim of the simulation study is to assess the performances of the MCMC algorithms based upon IBD and ABD with respect to the run time and accuracy of posterior parameter estimates. The study covers both the fully observed case (with no missing data) and the partially observed case (with a missing proportion of the data). For the partially observed data, the key point is to assess how robust the algorithms are at various forms of missingness and as the proportion of missing data increases. In particular, how does a given proportion of missing data affect estimation of parameters?

Finally, in Section 2.6, we give some concluding remarks with reference to the results from the simulation study.

2.2 Data Description

Household-based SIS data may be available in a form such that it is possible to obtain information on the infectious status of all individuals in a household at each observation time. We call this Individual-based data (IBD). Sometimes, the only information that may be available is the number of infected individuals in the household at each observation time. We call this Aggregate-based data (ABD). From now on, we will use *susceptible* to refer to an individual who is prone to being infected and *infective* to refer to an infected and infectious individual.

2.2.1 Individual-based Data (IBD)

The Individual-based data (IBD) provides us with information on the infection statuses of each individual of a given household at a point in time. This implies that the IBD also tells us the number of infectives in a given household at a point in time. Therefore, in a household of size $h \geq 1$, the IBD contains information on the infectious status of the h individuals of the household. Given that there are two possible states of an individual at a point in time, we encode susceptible 0 and infective 1. Since each individual is in one of the two (2) possible sates, there are 2^h possible states which the household can belong to at a given point in time. For $j = 1, 2, \ldots, h$, let $x_j(t) \in \{0, 1\}$ denote the infectious status of the j^{th} individual of the household at time t. Then the data $\mathbf{x}(t) = \big(x_1(t), x_2(t), \ldots, x_h(t)\big) \in \{0, 1\}^h$ is the state of the household at time t.

For $i = 1, 2, \ldots, N$, where N is the number of susceptible households, let h_i (≥ 1) and n_i denote the size and the number of observation time points of household i, respectively. Also, for $k = 1, 2, \ldots, n_i$, let t_{ik} denote the k^{th} observation time of

household i. Then, for $j = 1, 2, \ldots h_i$, it follows that

- $\mathbf{t}_i = (t_{i0}, t_{i1}, t_{i2}, \ldots, t_{in_i})$ are the observation times for household i, where t_{i0} is the initial time point of observation,
- $x_{ij}(t_{ik}) \in \{0, 1\}$ is the infectious status of individual j in household i at time point t_{ik},
- $\mathbf{x}_i(t_{ik}) = (x_{i1}(t_{ik}), x_{i2}(t_{ik}), \ldots, x_{ih_i}(t_{ik})) \in \{0, 1\}^{h_i}$ is the infection status of household i at time t_{ik},
- $\mathbf{x}_i(\mathbf{t}_i) = (\mathbf{x}_i(t_{i1}), \mathbf{x}_i(t_{i2}), \ldots, \mathbf{x}_i(t_{in_i}))$ is the infectious status of household i at time points $\mathbf{t}_i$.

Therefore, $\mathbf{x}(\mathbf{t}) = (\mathbf{x}_1(\mathbf{t}_1), \mathbf{x}_2(\mathbf{t}_2), \ldots, \mathbf{x}_N(\mathbf{t}_N))$ is the full data for the N households with a population of $M = \sum_{i=1}^{N} h_i$ individuals, where $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N)$ are the observation times.

When we observe every individual of a household at every given observation time point, we say that the data is *completely observed*. Often time, the data may only be *partially observed* in that only a subset of the household is observed in which case we define

$$y_{ij}(t_{ik}) = \begin{cases} 1 & \text{if } \textit{infectious}, \\ 0 & \text{if } \textit{susceptible}, \\ 2 & \text{if } \textit{status unknown}, \end{cases} \tag{2.2.1}$$

where $y_{ij}(t_{ik})$ is the infectious status of the j^{th} individual of household i at time point t_{ik}. Note that we assume that $P\big(y_{ij}(t_{ik}) = 2 | x_{ij}(t_{ik}) = 0\big) = P\big(y_{ij}(t_{ik}) = 2 | x_{ij}(t_{ik}) = 1\big)$, that is, the probability of the missingness depends neither on the

observed infectious status nor on the missing values themselves. In other words, the data are missing completely at random (MCAR), see, Rubin (1987). It follows that

- $\mathbf{y}_i(t_{ik}) = (y_{i1}(t_{ik}), y_{i2}(t_{ik}), \ldots, y_{ih_i}(t_{ik}))$ is the partially observed infectious state of household i at the k^{th} observation time point,
- $\mathbf{y}_i(\mathbf{t}_i) = (\mathbf{y}_i(t_{i1}), \mathbf{y}_i(t_{i2}), \ldots, \mathbf{y}_i(t_{in_i}))$ are the partially observed infectious states of household i over the observation time points $\mathbf{t}_i$,

Therefore, $\mathbf{y}(\mathbf{t}) = (\mathbf{y}_1(\mathbf{t}_1), \mathbf{y}_2(\mathbf{t}_2), \ldots, \mathbf{y}_N(\mathbf{t}_N))$ is the partially observed data for N households at time points $\mathbf{t}$. Note that $\mathbf{y}_i(t_{ik}) = \mathbf{x}_i(t_{ik})$ when there are no missing data.

In Section 2.4.3, we develop extensive data augmentation algorithms for the analysis of a partially observed IBD, $\mathbf{y}(\mathbf{t})$.

2.2.2 Aggregate-based Data (ABD)

SIS household epidemic data may only contain information on the total number of infectives in a household at a given point in time. As already mentioned, we call this Aggregate-based data (ABD) which unlike the IBD does not provide us with any information about a given individual's infectious status at a point in time.

For a given household of size $h \geq 1$, there are $h + 1$ possible states which the household can belong to at a point in time. Let

For $i = 1, 2, \ldots, N$, where N is the number of households considered, let $h_i \geq 1$ and n_i denote the size and number of observation time points for household i, respectively. Then, for $k = 1, 2, \ldots, n_i$, we have that

- $\tilde{x}_i(t_{ik}) \in \{0, 1, 2, \ldots, h_i\}$ is the infectious state of household i at the k^{th} time point,

- $\tilde{\mathbf{x}}_i(\mathbf{t}_i) = (\tilde{x}_i(t_{i1}), \tilde{x}_i(t_{i2}), \ldots, \tilde{x}_i(t_{in_i}))$ are the infectious states of household i at time points $\mathbf{t}_i$.

Therefore, $\tilde{\mathbf{x}}(\mathbf{t}) = (\tilde{\mathbf{x}}_1(\mathbf{t}_1), \tilde{\mathbf{x}}_2(\mathbf{t}_2), \ldots, \tilde{\mathbf{x}}_N(\mathbf{t}_N))$ is the full data for the N households with a total population of $M = \sum_{i=1}^{N} h_i$ individuals.

The aggregate-based data (ABD) may be completely or partially observed. When the data are completely observed, we observe $\tilde{x}_i(t_{ik})$, infectives in household i out of the h_i individuals of the household at time t_{ik}. When the ABD are only partially observed, we only observe a subset of the household out of which we observe a number of infectives. Let $\tilde{s}_i(t_{ik})$ denote the number of individuals observed in household i (of size $h_i \geq 1$) at the k^{th} time point, where $\tilde{s}_i(t_{ik}) \leq h_i$. Let $\tilde{y}_i(t_{ik})$ denote the number of infectives observed in a sample of $\tilde{s}_i(t_{ik})$ individuals in household i of size h_i at time t_{ik}, where $\tilde{y}_i(t_{ik}) \leq \tilde{x}_i(t_{ik})$. Observe that we have used *tilde* for the description of ABD in order to make a distinction from IBD. For $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, h_i$ and $k = 1, 2, \ldots, n_i$, we note the following equations

$$\tilde{x}_i(t_{ik}) = \sum_{j=1}^{h_i} x_{ij}(t_{ik}), \tag{2.2.2}$$

$$\tilde{y}_i(t_{ik}) = \sum_{j=1}^{h_i} \mathbb{1}_{\{y_{ij}(t_{ik})=1\}} \tag{2.2.3}$$

and

$$\tilde{s}_i(t_{ik}) \;=\; \sum_{j=1}^{h_i} \mathbb{1}_{\{y_{ij}(t_{ik})<2\}}, \tag{2.2.4}$$

where $y_{ij}(t_{ik})$ is as defined in (2.2.1) and where $\mathbb{1}_{\mathbf{A}} : \mathbf{X} \to \{0,1\}$ is an indicator function with

$$\mathbb{1}_{\mathbf{A}}(x) := \begin{cases} 1 & \text{if } x \in \mathbf{A}, \\ 0 & \text{if } x \notin \mathbf{A}. \end{cases} \tag{2.2.5}$$

(2.2.6)

2.3 Model Setup

In this section, we describe the formulation of our model by briefly recalling the main features of our model, and outline the construction of the rate matrices (G-Matrices) for the two aforementioned data forms (IBD & ABD) as well as the calculation of the corresponding transition probability matrices (Q-matrices).

Consider a closed population partitioned into non-overlapping households. First, we consider equal size households so that the population contains N households each of size h. Then the population size is $M = Nh$. Given that our model is the SIS model, at any given point in time, an individual is either in the susceptible state or in the infected state at any given point in time. Then the only transitions allowed are from S$\to$ I and from I $\to$ S. Once successfully contacted by an infective, a susceptible becomes infected and infectious and remains in the infective state for an exponentially distributed time I with rate parameter $\gamma > 0$, *i.e.*, $I \sim \exp(\gamma)$ with mean γ^{-1}.

A susceptible can contract infection from an infectious member of his household and from outside the household. Infection enters a given household with a global force of infection λ. While infectious, an infective makes a *local* infectious contact at points of a Poisson process with rate β. A successful local contact is made with a susceptible chosen uniformly at random from its own household. By successful contact we mean contacts which ultimately results to the infection of the susceptible so contacted. Therefore, only contacts made globally or locally with a susceptible can result to an infection, and nothing happens when contact is with another infective. We assume that a newly infected individual does not undergo any latent period and thus immediately becomes infectious. The population is closed in that there are no births, no deaths and no migrations. All the Poisson processes, which in this case include those associated with the same individual, are assumed to be mutually independent. The infectious periods of different individuals are also assumed to be mutually independent. An infective recovers at rate γ and at the end of its infectious period immediately returns to the susceptible state and can be reinfected. Therefore there is no removed state and recovery from the disease does not confer immunity. The model described here can easily be extended to various household structures. In particular, we can extend the model to a population made up of unequal sized households with N_h households of size h, so that $N(=\sum_{n=1}^{h_{max}} N_h)$ is the total number of households with $M(=\sum_{h=1}^{h_{max}} hN_h)$ individuals, where h_{max} is the maximum household size.

2.3.1 The Infinitesimal Rate Matrix and Transition Probability Matrix

In this section we give a full description of the construction of the rate matrix (or G-matrix) and the calculation of the appropriate transition probability matrix (or Q-matrix). We shall first recall the construction of the single population stochastic SIS epidemic and then extend this to the household stochastic SIS models we consider here.

We now consider the model for when the population consists of N households each of size h, where $h \geq 1$. We begin by describing the construction of the rate matrices for both the individual-based data (IBD) and the aggregate-based data (ABD).

G-matrix for individual-based data

For a household of size h, label the individuals $1, 2, \ldots, h$, and let $\mathbf{u} = \big(u_1, u_2, \ldots, u_h\big)' \in \{0,1\}^h$ denote the state of the household at a given point in time. For $1 \leq j \leq h$, we define a *standard basis vector*

$$\mathbf{e}_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^h \ ,$$

in which the j^{th} element is equal to 1 and all other elements equal to 0. Then provided that $u_j = 0$, the infection of individual j corresponds to the transition $\mathbf{u} \to (\mathbf{u} + \mathbf{e}_j)$. Similarly provided that $u_j = 1$, the transition $\mathbf{u} \to (\mathbf{u} - \mathbf{e}_j)$ corresponds to the recovery of individual j.

For example, if $h = 3$ and $j = 2$, then $\mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, and for $\mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, say, the transition $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \to \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ (or $\mathbf{u} \to \mathbf{u} + \mathbf{e}_2$) is the infection of the second individual.

Given the global force of infection λ and the rate of local infection β, the total rate of infection of an individual is $\lambda + \beta \sum_{i=1}^{h} u_i$. Further, the recovery rate γ, is constant irrespective of whether infection is contracted locally or globally.

We define the $2^h \times 2^h$ infinitesimal rate matrix $\mathbf{G}^{(h)} = \left(g_{\mathbf{uv}}^{(h)}\right)$ for the individual based data (IBD) as

$$g_{\mathbf{uv}}^{(h)} = \begin{cases} \lambda + \beta \sum_{i=1}^{h} u_i & \text{if } u_j = 0 \text{ and } \mathbf{v} = \mathbf{u} + \mathbf{e}_j, \\ \gamma & \text{if } u_j = 1 \text{ and } \mathbf{v} = \mathbf{u} - \mathbf{e}_j \\ -\sum\limits_{\mathbf{w} \neq \mathbf{u}} g_{\mathbf{uw}}^{(h)} & \text{if } \mathbf{v} = \mathbf{u} \\ 0 & \text{Otherwise.} \end{cases} \tag{2.3.1}$$

for $\lambda, \beta, \gamma > 0$ and $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w} \in \mathcal{S}$. For $h = 3$, for example, the G-matrix for IBD is given as

$$\mathbf{G}^{(3)} = \begin{pmatrix} (0,0,0) & (0,0,1) & (0,1,0) & (0,1,1) & (1,0,0) & (1,0,1) & (1,1,0) & (1,1,1) \\ -3\lambda & \lambda & \lambda & 0 & \lambda & 0 & 0 & 0 \\ \gamma & -(\gamma + 2(\lambda+\beta)) & 0 & \lambda+\beta & 0 & \lambda+\beta & 0 & 0 \\ \gamma & 0 & -(\gamma + 2(\lambda+\beta)) & \lambda+\beta & 0 & 0 & \lambda+\beta & 0 \\ 0 & \gamma & \gamma & -(2\gamma+\lambda+2\beta) & 0 & 0 & 0 & \lambda+2\beta \\ \gamma & 0 & 0 & 0 & -(\gamma + 2(\lambda+\beta)) & \lambda+\beta & \lambda+\beta & 0 \\ 0 & \gamma & 0 & 0 & \gamma & -(2\gamma+\lambda+2\beta) & 0 & \lambda+2\beta \\ 0 & 0 & \gamma & 0 & \gamma & 0 & -(\gamma+\lambda+2\beta) & \lambda+2\beta \\ 0 & 0 & 0 & \gamma & 0 & \gamma & \gamma & -3\gamma \end{pmatrix}$$

G-matrix for aggregate-based data

We construct the infinitesimal rate matrix, G-matrix, for the aggregate-based data as follows: Given a household of size h, let $m \in \{0, 1, 2, \ldots, h\}$ denote the state (the total number of infectives) of the household at a given point in time. Also,

let $s = h - m$ denote the number of susceptibles in the household. We adapt the methods described above for the construction of the G-matrix for the IBD here. Given that we consider the infectiousness of the entire household rather than that of a single individual at a point in time, then in the event of an infection, the total rate of infection is given by

$$\beta m(h-m) + \lambda(h-m). \tag{2.3.2}$$

On the other hand, the total rate of recovery is γm. Therefore, for $0 \leq m, n \leq h$, we define a $(h+1) \times (h+1)$ infinitesimal rate matrix $\mathbf{G}^{(h)} = (g_{m,n}^{(h)})$for the ABD as

$$g_{m,n}^{(h)} = \begin{cases} \beta m(h-m) + \lambda(h-m) & \text{if } n = m+1 \\ \gamma m & \text{if } n = m-1 \\ 0 & \text{if } |n-m| > 1 \\ -\sum\limits_{k \neq m} g_{m,k}^{(h)} & \text{if } n = m \end{cases} \tag{2.3.3}$$

for $\lambda, \beta, \gamma > 0$ and m, n, $k \in \mathcal{S}$. For example, the G-matrix for a household of size $(h =)$ 3 for ABD is given by

$$
\mathrm{G}^{(3)} = \begin{pmatrix}
(0) & (1) & (2) & (3) \\
-3\lambda & 3\lambda & 0 & 0 \\
\gamma & -(\gamma + 2(\lambda+\beta)) & 2(\lambda+\beta) & 0 \\
0 & 2\gamma & -(2\gamma+\lambda+2\beta) & \lambda+2\beta \\
0 & 0 & 3\gamma & -3\gamma
\end{pmatrix}.
$$

Given that the G-matrix depends only on the household size, we only need to compute $\mathbf{G}^{(1)}$, $\mathbf{G}^{(2)}$, $\ldots, \mathbf{G}^{(h_{max})}$, where h_{max} denotes the maximum household size,. We note that the computational cost for the G-matrix for the IBD grows exponentially with h. For example, when $h = 10$, there are $1,048,576$ entries in the G-matrix which places a huge burden on computer memory. Therefore, we seek to calculate the G-matrix in as much efficient manner as obtainable.

Note that for both G-matrices (IBD and ABD) and at least for the Markov chains used for our purposes, $\sum_j g_{ij} = 0$ or $\mathbf{G1}' = \mathbf{0}'$ (with $g_{ii} < 0$ and $g_{ij} \geq 0$), where $\mathbf{1}$ and $\mathbf{0}$ are row vectors of ones and zeros.

Transition Probability Matrix (Q-Matrix)

We define the transition probability matrices for the two prevalent household SIS epidemic data considered here as follows. As before, for $t \geq 0$, let $X_h(t)$ denote the infectious state of a given household of size h at time t. Then for the IBD, $\mathbf{Q}_t^{(h)} = (q_{\mathbf{u},\mathbf{v}}^{(h)}(t))$ is the $2^h \times 2^h$ transition probability matrix of the continuous time Markov chain $\{X_h(t); t \geq 0\}$, with

$$q^{(h)}_{\mathbf{u},\mathbf{v}}(t) = P\big(X_h(t+\Delta t)) = \mathbf{v} | X_h(t) = \mathbf{u}\big), \tag{2.3.4}$$

and

$$P\big(X_h(t+\Delta t)) = \mathbf{v} | X_h(t) = \mathbf{u}\big) = \begin{cases} (\lambda + \beta \sum_{i=1}^{h} u_i)\Delta t + o(\Delta t) & \text{if } \mathbf{v} = \mathbf{u}_j^+, \\ \gamma \Delta t + o(\Delta t) & \text{if } \mathbf{v} = \mathbf{u}_j^- \\ 1 - \big[\lambda + \beta \sum_{i=1}^{h} u_i + \gamma\big]\Delta t + o(\Delta t) & \text{if } \mathbf{v} = \mathbf{u} \\ o(\Delta t) & \text{Otherwise,} \end{cases} \tag{2.3.5}$$

where $\mathbf{u}_j^+ = \mathbf{u} + \mathbf{e}_j$ and $\mathbf{u}_j^- = \mathbf{u} - \mathbf{e}_j$.

Similarly, for the ABD, $\mathbf{Q}_t^{(h)} = (q^{(h)}_{m,n}(t))$ is a $(h+1) \times (h+1)$ transition probability matrix, where $P\big(X_h(t+\Delta t)) = n | X_h(t) = m\big)$ is given by

$$q^{(h)}_{m,n}(t) = \begin{cases} \big[\beta m(h-m) + \lambda(h-m)\big]\Delta t + o(\Delta t) & \text{if } n = m+1 \\ \gamma m \Delta t + o(\Delta t) & \text{if } n = m-1 \\ o(\Delta t) & \text{if } |n-m| > 1 \\ 1 - \big[\beta m(h-m) + \lambda(h-m) + \gamma m\big]\Delta t + o(\Delta t) & \text{if } n = m. \end{cases} \tag{2.3.6}$$

Note that the Q-matrix is a stochastic matrix in that $\sum_j q^{(h)}_{i,j}(t) = 1$ for all i and $q^{(h)}_{i,j}(t) \geq 0$. The Q-matrix is calculated by taking the matrix exponential of the product of the transition rate matrix and t as below:

$$\begin{aligned} \mathbf{Q}_t^{(h)} &= e^{tG^{(h)}} \\ &= \sum_{n=0}^{\infty} \frac{(tG^{(h)})^n}{n!} \\ &= \mathbf{I} + \sum_{n=1}^{\infty} \frac{t^n (G^{(h)})^n}{n!}, \end{aligned} \tag{2.3.7}$$

where $\mathbf{I}$ is a $2^h \times 2^h$ identity matrix if IBD or $(h+1) \times (h+1)$ identity matrix if ABD.

The Q-matrix depends upon both t and h and is thus calculated for each household size at every given time point. We seek to minimize computational cost and increase computational speed and so in all cases where t is integer ($t \in \mathbb{N}$), we can compute $\mathbf{Q}_1^{(h)}$ and then raise it to the t^{th} power to obtain $\mathbf{Q}_t^{(h)}$.

2.4 Bayesian Inference on Household-based SIS Epidemic

In this section, we outline the procedures for performing Bayesian inference on both the two data forms- the IBD and ABD. Specifically, we outline the implementation of Markov Chain Monte Carlo (MCMC) algorithms for Bayesian posterior inference with respect to the data.

In Section 2.4.1 we give Bayesian inference procedure for the analysis of a household-based SIS data when the data are completely observed. We later extend this to partially observed data in Section 2.4.3 where we give a step by step approach for the data imputation strategy used via data augmentation strategies for partially observed epidemics.

2.4.1 Bayesian Inference on Completely Observed Household SIS Data

Setup

Let $\mathbf{x} = \{\mathbf{x}(\mathbf{t})\}$ denote the observed data from the SIS epidemic model with model parameters $\boldsymbol{\theta} = (\lambda, \beta, \gamma)'$. We assume that there are no missing components of $\mathbf{x}$, *i.e.*, the data are completely observed. In a Bayesian framework, we seek to explore the target distribution in order to obtain some vital information such as $\mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x}]$ and $var(h(\boldsymbol{\theta})|\mathbf{x})$. As already stated in Chapter 1, Bayesian inference, requires that we calculate the posterior distribution of the parameters given data, $\pi(\boldsymbol{\theta}|\mathbf{x})$ $\big($ $= \pi(\mathbf{x}|\boldsymbol{\theta})$ $\pi(\boldsymbol{\theta})/\pi(\mathbf{x})\big)$, where $\pi(\mathbf{x}|\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathbf{x})$ is the likelihood function of the parameters given the observed data, $\pi(\boldsymbol{\theta})$ is the prior distribution of the parameters, and $\pi(\mathbf{x}) = \int \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal likelihood of the data.

The calculation of the marginal likelihood, $\pi(\mathbf{x})$ is usually problematic, but Markov Chain Monte Carlo (MCMC) algorithms allow us to draw samples directly from $\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, hence no need to calculate $\pi(\mathbf{x})$.

Priors

Given that our model parameters are positively defined, *i.e.*, λ, β, $\gamma > 0$, we constrain the state-space of the Markov Chain by assigning appropriate prior distributions. In particular, for our purposes we assign independent Gamma distributed priors to the parameters, $\pi(\boldsymbol{\theta}) = \pi(\lambda)\pi(\beta)\pi(\gamma)$, such that

$$\lambda \sim Gamma(A_\lambda, B_\lambda)$$

$$\beta \sim Gamma(A_\beta, B_\beta)$$

$$\gamma \sim Gamma(A_\gamma, B_\gamma) \quad . \tag{2.4.1}$$

That is,

$$\pi(\lambda) = \frac{B_\lambda^{A_\lambda}}{\Gamma(A_\lambda)} \lambda^{A_\lambda - 1} exp(-B_\lambda \lambda) \quad \lambda > 0 \tag{2.4.2}$$

$$\pi(\beta) = \frac{B_\beta^{A_\beta}}{\Gamma(A_\beta)} \beta^{A_\beta - 1} exp(-B_\beta \beta) \quad \beta > 0 \tag{2.4.3}$$

$$\pi(\gamma) = \frac{B_\gamma^{A_\gamma}}{\Gamma(A_\gamma)} \gamma^{A_\gamma - 1} exp(-B_\gamma \gamma) \quad \gamma > 0 \quad , \tag{2.4.4}$$

where $A_\lambda > 0$, $B_\lambda > 0$, $A_\beta > 0$, $B_\beta > 0$, $A_\gamma > 0$ and $B_\gamma > 0$ are hyper-parameters.

Likelihood

Evaluation of the likelihood function is required in order to obtain the posterior distribution of interest. Let $X(t_k)$ denote the state of a given household at time t_k. Also let $\mathbf{x}(t_k)$ denote the realizations of $X(t_k)$ at time t_k.

The likelihood function of the household over the n observation time points, $\pi\big(\mathbf{x}(\mathbf{t})|\boldsymbol{\theta}\big)$ satisfies

$$\pi\big(\mathbf{x}(\mathbf{t})|\boldsymbol{\theta}\big) = \prod_{k=2}^{n} \Big\{ \pi\big(\mathbf{x}(t_k)|\mathbf{x}(t_{k-1}), \boldsymbol{\theta}\big) \Big\}. \tag{2.4.5}$$

It follows that the full likelihood function is given by

$$L(\boldsymbol{\theta}; \mathbf{x}) := \pi\big(\mathbf{x}|\boldsymbol{\theta}\big) = \prod_{i=1}^{N} \prod_{k=2}^{n_i} \Big\{ \pi\big(\mathbf{x}_i(t_{i,k})|\mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta}\big) \Big\} . \tag{2.4.6}$$

where $\pi(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta})$ is the probability of of household i of being in state $\mathbf{x}_i(t_{ik})$ at time point t_{ik} from state $\mathbf{x}_i(t_{i,k-1})$ at time point $t_{i,k-1}$.

Posterior Distribution

The posterior distribution of the parameters given the data, $\pi(\boldsymbol{\theta}|\mathbf{x})$, up to proportionality is then given by

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\mathbf{x}) &\propto \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \qquad (2.4.7)\\
&\propto \left\{\prod_{i=1}^{N}\prod_{k=2}^{n_i}\left\{\mathbb{P}\big(X_i(t_{ik}) = \mathbf{x}_i(t_{ik})|X_i(t_{i,k-1}) = \mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta}\big)\right\}\right\}\\
&\times \pi(\lambda) \times \pi(\beta) \times \pi(\gamma)\\
&\propto \left\{\prod_{i=1}^{N}\prod_{k=2}^{n_i}\left\{\pi\big(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta}\big)\right\}\right\}\\
&\times \lambda^{A_\lambda - 1}e^{-B_\lambda \lambda} \times \beta^{A_\beta - 1}e^{-B_\beta \beta} \times \gamma^{A_\gamma - 1}e^{-B_\gamma \gamma}. \qquad (2.4.8)
\end{aligned}$$

Note that the marginal distribution $\pi(\mathbf{x})$ and the constants $B_\lambda^{A_\lambda}/\Gamma(A_\lambda)$, $B_\lambda^{A_\lambda}/\Gamma(A_\lambda)$ and $B_\lambda^{A_\lambda}/\Gamma(A_\lambda)$ are independent of $\boldsymbol{\theta}\{= (\lambda, \beta, \gamma)'\}$ and are not necessary for drawing samples from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$ using MCMC.

2.4.2 MCMC

Having obtained the posterior distribution, we are now in position to draw samples from the posterior distribution of interest for posterior inference. Markov chain Monte Carlo (MCMC) algorithms allow us to construct a Markov chain $\{\boldsymbol{\theta}^r; r \geq 1\}$ whose stationary distribution is our target density, $\pi(\boldsymbol{\theta}|\mathbf{x})$.

Given that the Q-matrix is a complicated function of the parameters resulting to a non-standard posterior distribution in (2.4.7), Gibbs sampler is therefore

not an applicable updating scheme here. Therefore, we update the parameters, $\boldsymbol{\theta} = (\lambda, \beta, \gamma)'$, using Random Walk Metropolis algorithms with multivariate Gaussian proposal and block updating scheme which has been shown to outperform univariate variable-at-a-time scheme. Throughout, we use the adaptive MCMC steps outlined in Section 1.4.9 of Chapter 1 to improve the efficiency of the MCMC algorithms. For example, we exploit the optimal shaping approach which enables the algorithm to quickly learn the posterior shape of the model.First we obtain a posterior variance-covariance matrix, Σ, from a pilot run with an acceptance rate that is very close to the optimal rate of 23.4%, and at least which yields a sensible estimates of the parameters. Then, we run the main MCMC using the Σ so obtained. We give a generic MCMC algorithm used for sampling from the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$ according to Algorithm 4.

2.4.3 Bayesian Inference on Partially Observed Household Epidemic (Data Augmentation)

Setup

In this section, we detail the procedures for Bayesian inference on partially observed household-based SIS epidemic.

Let $\mathbf{y}$ denote the observed data from the model with parameters $\boldsymbol{\theta} = (\lambda, \beta, \gamma)$. As already noted in Section 2.4.1, the likelihood function $\pi(\mathbf{y}|\boldsymbol{\theta})$ is not tractable, but given some extra information, $\mathbf{z}$, which is consistent with the observed data, the likelihood $\pi(\mathbf{x} = (\mathbf{z}, \mathbf{y})|\boldsymbol{\theta})$ becomes tractable. Then Bayesian inference on the posterior distribution of the parameters given the complete data $\pi(\boldsymbol{\theta}|\mathbf{x} = (\mathbf{z}, \mathbf{y}))$ becomes possible. To do this we need to construct an efficient MCMC algorithm

Algorithm 4 Random Walk Metropolis (RWM) algorithm

1. Start the Chain with initial values of $\boldsymbol{\theta}^0 = (\lambda^0,\ \beta^0,\ \gamma^0)'$ and set $\boldsymbol{\theta}^{curr} = \boldsymbol{\theta}^0$.
2. Set $c = \frac{2.38^2}{d}$.
3. Set $\Sigma = c\Sigma_n$, where Σ_n is the posterior covariance matrix from the previous n runs.
4. For $r = 1, \ldots, N$ (where N is the desired number of iterations),
5. Propose $\boldsymbol{\theta}^{prop} \sim N_d(\boldsymbol{\theta}^{curr}, \Sigma)$
6. Accept $\boldsymbol{\theta}^{prop}$ with probability:
7. $\alpha = \left\{ \frac{\pi(\boldsymbol{\theta}^{prop}|\mathbf{X})}{\pi(\boldsymbol{\theta}^{curr}|\mathbf{x})} \wedge 1 \right\} = min\left\{ 1, \frac{\pi(\mathbf{x}|\boldsymbol{\theta}^{prop})\pi(\boldsymbol{\theta}^{prop})}{\pi(\mathbf{x}|\boldsymbol{\theta}^{curr})\pi(\boldsymbol{\theta}^{curr})} \right\}$
8. Draw $u \sim U[0, 1]$.
9. If $u \leq \alpha$,
   - Accept $\boldsymbol{\theta}^{prop}$
   - Set $\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{prop}$,

   else,
   - $\boldsymbol{\theta}^{prop}$ not accepted
   - Set $\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{r} (r \geq 0)$.
10. Repeat 5 to 8 until a sample of the desired size is obtained.

to obtain samples from the joint posterior density $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ by alternating between updating $\pi(\boldsymbol{\theta}|\mathbf{x} = (\mathbf{z}, \mathbf{y}))$ and $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$. This is called *data augmentation.*

To exploit the data augmentation scheme, we consider the joint posterior distribution $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ and given that

$$\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \frac{\pi(\mathbf{z}, \mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})} \tag{2.4.9}$$

$$\propto \pi(\mathbf{z}, \mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{2.4.10}$$

then we have that

$$\pi(\boldsymbol{\theta}|\mathbf{x} = (\mathbf{z}, \mathbf{y})) \propto \pi(\mathbf{z}, \mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2.4.11}$$

and

$$\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}) \propto \pi(\mathbf{z}, \mathbf{y}|\boldsymbol{\theta}) \ . \tag{2.4.12}$$

Therefore, under mild conditions, MCMC samples drawn iteratively from $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{x} = (\mathbf{z}, \mathbf{y}))$ will give us samples from $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$. These two steps are summarized in Algorithm 5.

Algorithm 5 Data Augmentation steps

1. Draw $\mathbf{z}^r \sim \pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{r-1})$ using Independence Sampler or Gibbs sampler (as appropriate).
2. Draw $\boldsymbol{\theta}^r \sim \pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}^r)$ using Random Walk Metropolis algorithm as outlined in Algorithm 4.

The sampled values of the parameters $\boldsymbol{\theta}^r; r \geq 1$ are normally stored while $\mathbf{z}^r$ may be stored for further use or discarded as nuisance parameters not needed.

2.4.4 Bayesian Inference on partially observed Individual-based SIS data (IBD)

In this Section, we shall now focus on developing data augmentation schemes with respect to the individual-based data (IBD). Interest here is on using an efficient data imputation approach for the updating of $\pi(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$. As before, both $\mathbf{z}$ and $\mathbf{y}$ denote the imputed data and the partially observed data, respectively. Then, $z_{ij}(t_{ik})$ is the imputed infectious status of the j^{th} individual in household i at time point t_{ik} $\{i = 1, \ldots, N; j = 1, \ldots, h_i; k = 1, \ldots, n_i\}$. To initialize the data, we assume that an individual with unknown infectious status at the first observation time point is in the susceptible state. This assumption can be relaxed to allow for different infectious statuses of individuals at the first observation time point. On the other hand, if the infectious status of individual (i, j) is unknown at time point t_{ik}, we assume that the infectious status of the individual at time point $t_{i,k-1}$ did not change at time point t_{ik}.

Therefore, if $y_{ij}(t_{ik}) < 2$, set $x_{ij}(t_{ik}) = y_{ij}(t_{ik})$. Otherwise, if $y_{ij}(t_{i1}) = 2$, we set $x_{ij}(t_{i1}) = 0$, and for $2 \leq k \leq n_i$ set $x_{ij}(t_{ik}) = x_{ij}(t_{ik-1})$, if $y_{ij}(t_{ik}) = 2$.

Then, from (2.4.12) we have,

$$\begin{aligned} \pi(\mathbf{x}|\boldsymbol{\theta}) &\propto \prod_{i=1}^{N}\prod_{k=2}^{n_i}\Big\{\pi\big(\mathbf{y}_i(t_{ik})|\mathbf{x}_i(t_{ik})\big) \\ &\times \pi\big(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{ik-1}), \boldsymbol{\theta}\big)\pi\big(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), \boldsymbol{\theta}\big)\Big\}. \end{aligned} \tag{2.4.13}$$

Therefore updating $\pi(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$ involves drawing $z_{ij}(t_{ik})$ from

$$\begin{aligned} \pi(z_{ij}(t_{ik})|\boldsymbol{\theta}, \mathbf{x}_{-ij}(t_{ik})) &\propto \pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}), \boldsymbol{\theta}) \\ &\times \pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik}), \boldsymbol{\theta}) \end{aligned} \tag{2.4.14}$$

where $\mathbf{x}_{-ij}(t_{ik})$ is the complete data vector for the infectious state of household i without $z_{ij}(t_{ik})$ and $\mathbf{z}_i(t_{ik})$ is the state of household i at time t_{ik}.

Independence Sampler

We propose $z_{ij}^{prop}(t_{ik}) = 1 - z_{ij}(t_{ik})$, if $y_{ij}(t_{ik}) = 2$ (switching the states). The proposed values are accepted using a quick and simple acceptance probability, $\alpha(z_{ij}(t_{ik}), z_{ij}^{prop}(t_{ik}))$, which depends on a given household at the three time points $(t_{ik-1},\, t_{ik},\, t_{ik+1})$, and is given by

$$\min\left\{1, \frac{\pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}^{prop}(t_{ik})|\mathbf{x}_i(t_{ik-1}), \boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}^{prop}(t_{ik}), \boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}), \boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik}), \boldsymbol{\theta})}\right\} \tag{2.4.15}$$

Algorithm 6 Independence Sampler (IS) algorithm

1. With fixed values of $\boldsymbol{\theta} = (\lambda,\, \beta,\, \gamma)'$,
2. Propose to switch states by setting $z_{ij}^{prop}(t_{ik}) = 1 - z_{ij}(t_{ik})$, if $y_{ij}(t_{ik}) = 2$.
3. Calculate the acceptance probability α.
4. Draw $u \sim U[0, 1]$.
5. If $u \leq \alpha$,
   - accept $z_{ij}^{prop}(t_{ik})$

- Set $\mathrm{z}_{ij(t_{ik})} = z_{ij(t_{ik})}^{prop}$ else,
  - $z_{ij}^{prop}(t_{ik})$ is rejected

2.4.5 Bayesian Inference on partially observed Aggregate-based SIS data (ABD)

In this section, we describe a data augmentation schema for the aggregate-based data (ABD).

Given a household of size h, suppose we are not able to observe every individual in the household, that is, we only observed a subset of the household. Let $\tilde{\mathbf{y}}$ denote the partially observed data and as before, $\boldsymbol{\theta} = (\lambda, \beta, \gamma)'$ denotes the parameters of the model. Let $\tilde{s}(t_k)$ denote the number of individuals observed in the household at time t_k out of which we observe $\tilde{y}(t_k)$ infectives when in fact there actually $\tilde{x}(t_k)$ infectives in the household at time t_k. Therefore, when the observed number of individuals in the household at time t_k is equal to the size of the household, then the observed number of infectives is equal to the unobserved actual number of infectives in the household at that point in time. In other words, when $\tilde{s}(t_k) = h$, then $\tilde{y}(t_k) = \tilde{x}(t_k)$.

On the other hand, when the observed number of individuals in the household at time t_k is less than the size of the household, then the observed number of infectives is at most the unobserved actual number of infectives in the household at that point in time. In other words, when $\tilde{s}(t_k) < h$, then $\tilde{y}(t_k) \leq \tilde{x}(t_k)$. The latter statement implies that when we do not observe all the individuals of the household, then the number of infectives observed may be less than or same as the actual number of infectives. Furthermore, given that only a subset of the household is observed, it is possible that the unobserved actual number of infectives in the household at time t_k is less than or equal to the size of the household, that is, $\tilde{x}(t_k) \leq h$ when $\tilde{s}(t_k) < h$. Therefore, when $\tilde{s}(t_k) < h$, then $\tilde{y}(t_k) \leq \tilde{x}(t_k) \leq h$.

Now let $\tilde{z} \in \{0, 1, \ldots, h\}$ denote the vector of all possible number of infectives in the

household at a point in time. The probability of observing $\tilde{y}(t_k)$ infectives in the household at time t_k given the number of individuals observed $\tilde{s}(t_k)$, the household size h and $\tilde{\mathbf{z}}$, $\pi(\tilde{y}|\tilde{s}, h, \tilde{\mathbf{z}})$ the hypergeometric distribution

$$\pi(\tilde{y}|\tilde{s}, h, \tilde{\mathbf{z}}) = \frac{\binom{\tilde{\mathbf{z}}}{\tilde{y}}\binom{h-\tilde{\mathbf{z}}}{\tilde{s}-\tilde{y}}}{\binom{h}{\tilde{s}}}, \tag{2.4.16}$$

where $\tilde{y} = \tilde{y}(t_k)$ and $\tilde{s} = \tilde{s}(t_k)$. We compute (up to the constant of proportionality), $\tilde{P}_{\tilde{x}} = \pi(\tilde{x}(t_k)|.)$, the probability of observing $\tilde{x}(t_k)$ infectives at time t_k given the number of infectives in the household at other time points according to

$$\begin{aligned} \tilde{P}_x \quad \propto \quad & \prod_{k=2}^{n} \Big\{ \pi(\tilde{y}(t_k)|\tilde{s}(t_k), h, \tilde{x}(t_k)) \\ \times \quad & \pi\big(\tilde{x}(t_k)|\tilde{x}(t_{k-1}), \boldsymbol{\theta}\big)\pi\big(\tilde{x}(t_{k+1})|\tilde{x}(t_k), \boldsymbol{\theta}\big) \Big\}. \end{aligned} \tag{2.4.17}$$

Finally, choose $\tilde{x}(t_k)$ from $\{0, 1, \ldots, h\}$ with probability $\tilde{P}_x$ and accept $\tilde{x}(t_k)$ with probability one. Note that $\tilde{P}_x = \pi(\tilde{x}(t_k)|\tilde{y}(t_k), \tilde{s}(t_k), h)$, the conditional distribution of $\tilde{x}$ given everything else. The steps described above are applied to all the $i = 1, 2, \ldots, N$, independent households whilst updating $\tilde{x}_{ij}(t_{ik})$ for every unobserved $\tilde{y}_{ij}(t_{ik})$.

We summarize the MCMC algorithms in Algorithm 7.

Algorithm 7 Data Augmentation

1. Update $\tilde{\mathbf{x}}|\tilde{\mathbf{y}}, \boldsymbol{\theta}$ using Gibbs sampling steps.
2. Update $\boldsymbol{\theta}|\tilde{\mathbf{x}}, \tilde{\mathbf{y}}$ using Algorithm 4.

2.5 Simulation Study

In this section, we carry out a simulation study on our proposed household-based SIS epidemic models. There are three overarching aims of this study. First, we want to assess

Parameters	True λ	True β	True γ
	2.00	1.50	2.50

Table 2.5.1: The parameter values used for simulating household-based SIS epidemics. Each set of parameter values was used to simulate samples size $N = 500$ households.

the efficiency of the MCMC algorithms outlined in Section 2.4. Specifically, we want to see how fast it takes the MCMC algorithms to run in a given computing machine and how accurate they estimate desired parameters with minimum auto-correlation. Second, we want to see how using aggregate-based data (ABD) compares with using the more informative individual-based data (IBD). In particular, we want to know if estimation with the ABD-based MCMC algorithm leads to a significant loss of information compared with the IBD-based MCMC. The third aim is to assess the robustness of our algorithms at different types of missingness and at different proportions (P) of missing data. That is, we want to see how random missingness (A), individual missingness (B) and time point missingness (C) as well as a given proportion of missingness affect the ability of the MCMC algorithms to estimate the parameters of the model.

2.5.1 Method

To address the first aim of this study, we choose $\boldsymbol{\theta} = (\lambda,\ \beta,\ \gamma)' = (2, 1.5, 2.5)'$ and simulate individual-based SIS epidemic data (IBD) for $N = 500$ independent households where the size of a given household, h_i $(i = 1, 2, \ldots, N)$, is chosen uniformly, at random from $\{1, 2, \ldots, 5\}$. Also, the initial state, $\mathbf{x}_i(t_{i0})$, of household i is chosen uniformly, at random, from $\{0, 1\}^{h_i}$ possible states to which household i can belong at a given point in time.

We choose the observation time points $(t_1, t_2, \ldots, t_{n_i})$ of each household such that there would be moderate changes between time points. Note that when t, the time difference, is too small, the transition probability matrix, $\mathbf{Q}_t^{(h)}$ tends to identity matrix, $\mathbf{I}$ and we observe only very little or no changes. On the other hand, when the time points are far apart (t is large), each row of the transition probability matrix, $\mathbf{Q}_t^{(h)}$, gets very close to the stationary distribution $\boldsymbol{\pi}$, making the infectious states of a given household at different timepoints appear independent. Similarly, for any $c > 0$, the transition probability matrix, $\mathbf{Q}_t^{(h)}$ tends to $\mathbf{I}^{(h)}(\boldsymbol{\pi})$ as $c\boldsymbol{\theta}$ gets smaller (larger).

Furthermore, the number of time points of observation of a given household, n_i, is chosen uniformly from $(2, 3, \ldots, 5)$. Next, we compute the two matrices, the rate matrix (G-matrix) using (2.3.1) and the transition probability matrix ($\mathbf{Q}$-matrix) using (2.3.7). When t is an integer, we only compute $\mathbf{Q}_1^{(h)}$ and obtain $\mathbf{Q}_t^{(h)}$ by raising it to the t^{th} power. This is found to greatly reduce the cpu time involved in the computation of $\mathbf{Q}_t^{(h)}$ by up to 70% of the time.

Finally, we sample $\mathbf{x}_i(t_{ik})$, the infectious state of household i, from row $\mathbf{x}_i(t_{ik-1})$ of $\mathbf{Q}_t^{(h)}$ and then record the data each time. We repeat the procedure for each sample using the same time points data for the various data until the desired sample size if observed.

To address the second aim which compares the performance of ABD-based and IBD-based MCMC algorithms, we obtain the aggregate-based data from the individual-based data by simply using (2.2.2). Note that the IBD and ABD simulated in this instance are assumed to be completely observed. The MCMC algorithms used including the choice of prior distributions for both ABD and IBD are outlined in Section 2.5.3.

The third aim of this simulation study seeks to assess how a given proportion of missing data affects the ability of the MCMC algorithms to accurately estimate the parameter

values. To address this, using same set of parameter values or by simply using the completely observed IBD, $\mathbf{x}$, already simulated in the first instance, we further simulate partially observed household-based SIS epidemic data $\mathbf{y}$ for various proportions, P, of missing data. First, we sample a uniform random variable u from $U(0,1)$ and set $\mathbf{y} = \mathbf{x}$. Three different forms of missing data are considered:

(A) Where the missingness is due to the infection status of an individual of a given household missing at random at a particular time point with a given probability, P. Here, we assume that $y_{ij}(t_{ik})$, the infectious status of individual j of household i at time t_{ik}, is missing whenever $u < P$. That is, $y_{ij}(t_{ik}) = 2$ if $u < P$ and $y_{ij}(t_{ik}) = x_{ij}(t_{ik})$ if $u \geq P$.

(B) Where the missingness is due to the infectious status of a randomly selected individual of a given household missing completely in all time points. In other words, a randomly selected individual with probability, P, is not observed at every time point. Here, each element of $\mathbf{y}_{ij}(\mathbf{t}_i) = \big(y_{ij}(t_{i1}), y_{ij}(t_{i2}), \ldots, y_{ij}(t_{in_i})\big)$ is unknown or equal to 2 when $u < P$ and $\mathbf{y}_{ij}(\mathbf{t}_i) = \mathbf{x}_{ij}(\mathbf{t}_i)$ if otherwise, where $\mathbf{y}_{ij}(\mathbf{t}_i)$ is the vector of infection status of individual j of household i across the n_i time points.

(C) Where the missingness is due to a randomly selected observation time point of a given household missing with probability, P, for all individuals in the household. That is to say that a given state of household i at a particular time point, $\mathbf{y}_i(t_{ik}) = \big(y_{i1}(t_{ik}), y_{i2}(t_{ik}), \ldots, y_{ih_i}(t_{ik})\big)'$, is completely missing or each element is equal to 2 when $u < P$, otherwise $\mathbf{y}_i(t_{ik}) = \mathbf{x}_i(t_{ik})$. For the purpose of analysis, we further classify this into

- $\mathrm{C}^{(1)}$ in which each unobserved column is deleted and the data treated as completely observed.

- $C^{(2)}$ in which case we employ the data augmentation algorithms outlined in Section 2.4.3 to analyse the incomplete data.

In all cases, the procedure is repeated for $P = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ (up to 0.9 or 90% missingness). As before, it is straightforward to obtain the corresponding partially observed aggregate-based data (ABD) using (2.2.3) and (2.2.4). Also, note that in all cases the data are missing at random: individuals' infectious status missing at random (A), rows missing at random (B) and time points or columns missing at random (C) from an array of data $\mathbf{X}$.

The main reason for considering the three afore-mentioned missing data forms is to assess how each data form affects parameters estimation and identify the most stable missing pattern.

In all cases, to assess the effects of sample sizes on the accuracy of the algorithms, we further obtain randomly selected sub-samples of sizes $N = 100, 200$ from the original sample ($N = 500$).

2.5.2 Sensitivity Analysis

We carry out a sensitivity analysis on the models to see how the MCMC algorithms are sensitive to changes in parameter values. We scale the initial parameters of interest $\boldsymbol{\theta}$ by the constant c taking values in $\{1/10, 1/5, 1/2, 2, 5, 10\}$ and obtain six more parameter sets, see, Table 2.5.2. Observe that the first row of Table 2.5.2 contains the original parameters, $\boldsymbol{\theta}$, used for the initial analysis.

In every case we simulate SIS household epidemic data for $N = 500$ and then randomly choose subsets of the households of sizes 100 and 200, respectively. We shall also consider the performance of the MCMC algorithms with respect to IBD and ABD. For each

Parameters	c	True λ	True β	True γ
I	1.00	2.00	1.50	2.50
II	0.10	0.20	0.15	0.25
III	0.20	0.40	0.30	0.50
IV	0.50	1.00	0.75	1.25
VI	2.00	4.00	3.00	5.00
VII	5.00	10.0	7.50	12.5
VIII	10.0	20.0	15.0	25.0

Table 2.5.2: The parameter values used for simulating household-based SIS epidemics. Each set of parameter values was used to simulate samples size $N = 500$ households.

sample, we analyze both the complete data case. We use the data augmentation approach described in Section 2.4.3 for the analysis of incomplete data with various proportion of missing data.

2.5.3 MCMC

Complete Data case

For every $c = \{1, 2, 5, 10, 1/2, 1/5, 1/10\}$, we used $(1/c)$ prior with mean equal to c. In Sections 2.4.1 and 2.4.3 to obtain the required samples for Bayesian posterior inferences. As already stated, the choice of the gamma-distributed priors is to ensure that our parameter values remain positive. Using Random Walk Metropolis (RWM) algorithms, we use block updating and propose $\boldsymbol{\theta}' \sim N_3\big(\boldsymbol{\theta}, \Sigma\big)$ (multivariate Random Walk proposal). From a pilot study, we found that a good initial value of the proposal variance-covariance

matrix, Σ, is

$$\begin{pmatrix} 0.01 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.01 \end{pmatrix},$$

and by exploiting the posterior variance tuning strategy outlined in Section 2.4.1, we ensured acceptance rates for the main MCMC runs are close to the suggested 23.4% for optimality according to

$$\alpha = min\left\{ \frac{L(\boldsymbol{\theta}') \times \pi(\boldsymbol{\theta}')}{L(\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}, 1 \right\}.$$

Incomplete Data Case

As outlined in Section 2.4.3 we alternate between updating $\pi(\tilde{z}|\tilde{\mathbf{y}}, \boldsymbol{\theta})$ and $\pi\big(\boldsymbol{\theta}|\tilde{\mathbf{x}} = (\tilde{\mathbf{z}}, \tilde{\mathbf{y}})\big)$. To update $\pi(\tilde{\mathbf{z}}|\tilde{\mathbf{y}}, \boldsymbol{\theta})$, we employed the methods outlined in Section 2.4.3 for the three forms of missing data considered. The only difference is in choosing the initial guess for the infection status, $x_{ij}(t_{ik})$, whenever $y_{ij}(t_{ik}) = 2$ (unknown) $(i = 1, \ldots, N; j = 1, \ldots, h_i; k = 1, \ldots, n_i)$. For the missing data form A, whenever $y_{ij}(t_{ik}) = 2$, we set the initial guess to be $x_{ij}(t_{ik}) = x_{ij}(t_{ik-1})$ (the infection status at the immediate preceding time point) and when $y_{ij}(t_{i1}) = 2$, we set $x_{ij}(t_{i1}) = 0$ (assuming initially susceptible). For missing data form B, we sample $x_{ij}(t_{ik})$ from $(0, 1)$ whenever $y_{ij}(t_{ik}) = 2$. The initial guess for $x_{ij}(t_{ik})$ for missing data form $\mathrm{C}^{(2)}$ is handled in the same way as in missing data form A in which case each affected column (time point) takes the values in the preceding column when $y_{ij}(t_{i1}) = 2$ or are all 0's (susceptibles) whenever there were no observations at the initial time point. Alternatively, we can also initialize $\mathrm{C}^{(2)}$ by setting $x_{ij}(t_{i1}) = x_{ij}(t_{i\kappa})$, where κ is the next observed time point. As already mentioned, the method we used in handling missing data form $\mathrm{C}^{(1)}$ is to delete the entire unobserved columns (or time points) and analyse the data as though were completely observed.

In all cases with IBD and for $P = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 0.9)$, we propose $x'_{ij}(t_{ik}) \rightarrow 1 - x_{ij}(t_{ik})$ (switching state) using an Independence Sampler step. The proposed value $x'_{ij}(t_{ik})$ is accepted (or rejected) according to (4.4.5). For the corresponding aggregate-based data, we update $\tilde{\mathbf{z}}|\tilde{\mathbf{y}}, \boldsymbol{\theta}$ using the Gibbs sampling steps also introduced in Section 2.4.3.

2.5.4 Results and Discussions

We have carried out a simulation study on our proposed models addressing three major aims- the accuracy of parameter estimates, comparison between the IBD-based and ABD-based algorithms in terms of speed of runs as well as the ability of the algorithms to yield posterior estimates (means) with low auto-correlations, and the stability of the algorithms as the percentage of missing data increases. We note that the second aim is addressed simultaneously with the first and the third aims since the IBD and ABD can always be compared for both complete data case and incomplete (partially observed) data case.

In all cases, we discuss the results obtained from 1×10^5 iterations after discarding 2×10^4 iterations as burn-in for $c = (0.1, 0.2, 0.5, 1, 2, 5, 10)$ and the corresponding parameter values (see, Table 2.5.2). Convergence diagnostic tools employed include trace plots, ACF plots, and paired density plots. These diagnostics were used throughout. Figure 2.5.1 shows the trace and density plots for the individual-based data (IBD) when up to 50% of the data are missing. The traceplot (left) which contains the history of the sojourn of the Markov Chain for the last $80,000$ iterations show that the MCMC algorithms are mixing well and convergence is deemed to have been achieved.

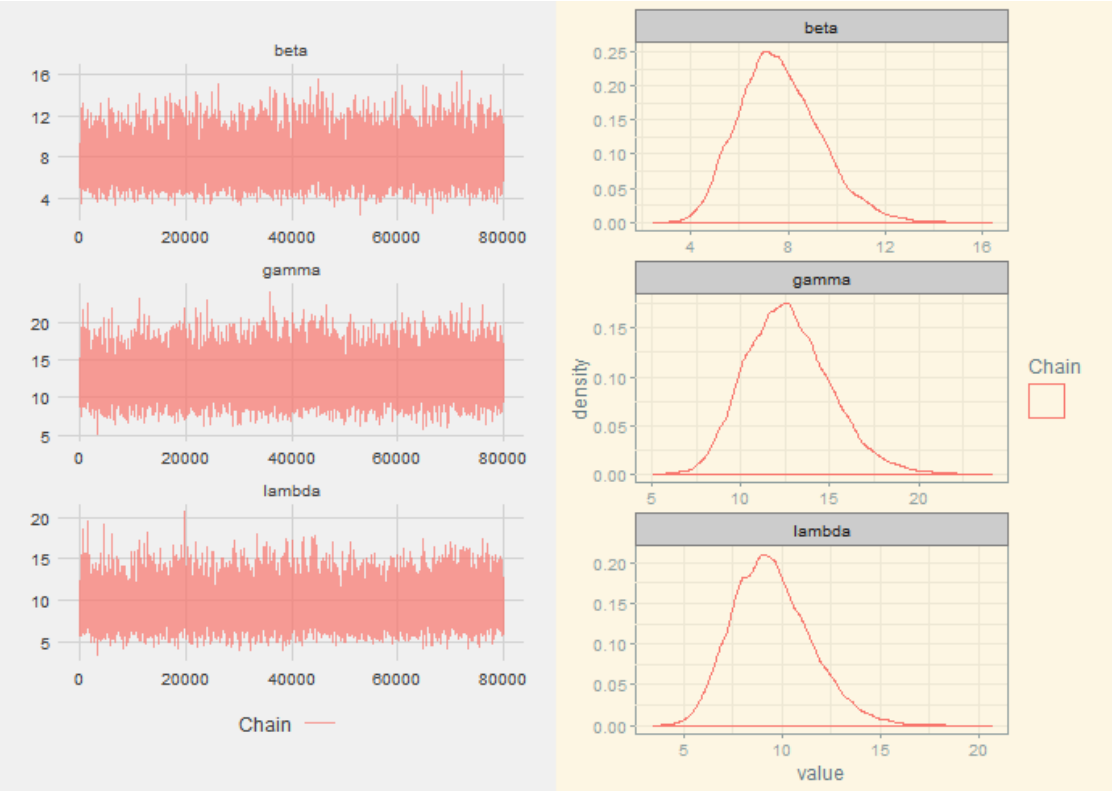

Figure 2.5.1: Right: trace plots obtained from 1×10^5 iteration after discarding the first 2×10^4 iterations as burn in. This plot if for the aggregate-based data for $c = 5$ and $P = 50\%$. Left: density plot. The plots here show that the mixing of the chains are good. The posterior estimates for the means of the three parameter are 7.90, 12.90 and 9.60 for β, γ and λ, respectively.

To check for autocorrelation, we used the autocorrelation function (ACF) plot to have a quick glance and ascertain the degree of the dependence between the sampled parameter values. Figure 2.5.2 shows that the serial autocorrelation between the parameter values is minimal.

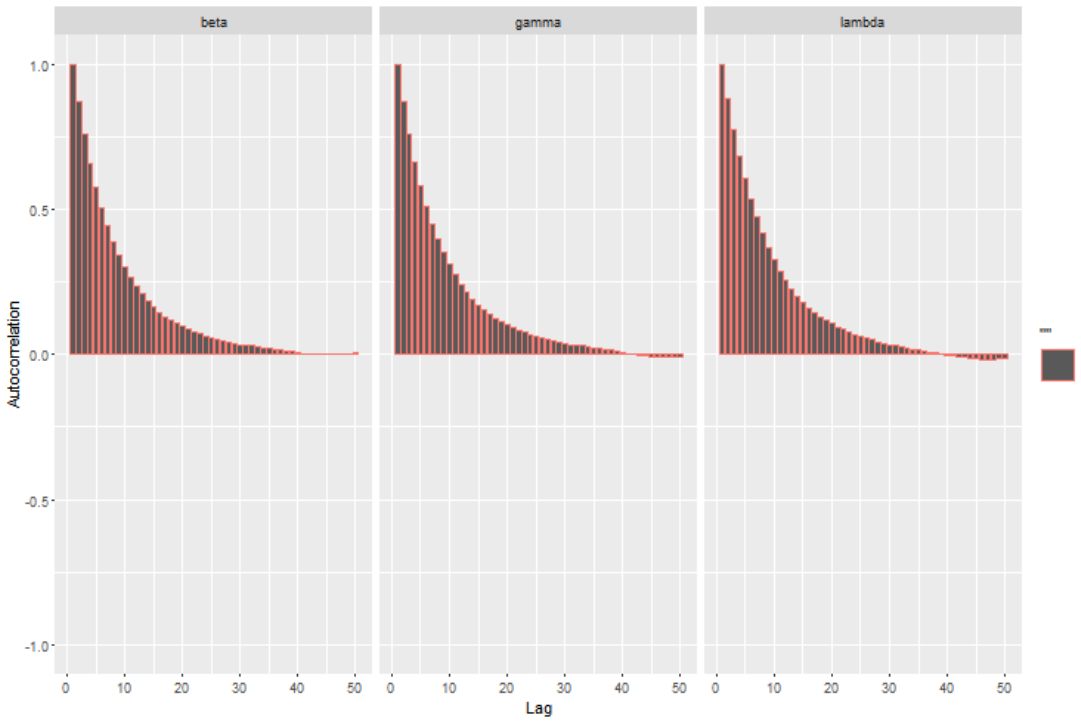

Figure 2.5.2: Autocorrelation function plot (ACF) for the individual-based data (IBD) for $c = 5$ or $(\beta,\ \gamma,\ \lambda) = (7.90,\ 12.90,\ 9.60)$ for 50% missing data.

We shall now proceed and report the results of the analysis. In Table 2.5.4, we compare the results obtained from the MCMC outputs when the data are completely observed (or no missing data) for both the individual-based data and the aggregate-based data. This comparison is done across the three sample sizes of $N = 100, 200, 500$ for $(\lambda, \beta, \gamma) = (2, 1.5, 2.5)$ or $c = 1$. It happened that the performance of both the IBD-based and ABD-based MCMC algorithms are similar in terms of posterior means, standard deviation (SD) and effective sample size (ESS).

Data	Size (N)	Parameter	Mean	SD	ESS
		$\lambda = 2.0$			
ABD	100		1.17	0.41	4768
IBD			1.15	0.40	5234
ABD	200		1.85	0.54	4489
IBD			1.89	0.53	5014
ABD	500		1.75	0.48	4039
IBD			1.76	0.48	4296
		$\beta = 1.5$			
ABD	100		1.68	0.60	5098
IBD			1.66	0.59	5322
ABD	200		1.34	0.42	4802
IBD			1.37	0.41	5596
ABD	500		1.21	0.35	4615
IBD			1.22	0.35	4314
		$\gamma = 2.5$			
ABD	100		1.90	0.65	4673
IBD			1.88	0.63	4932
ABD	200		2.30	0.68	4558
IBD			2.35	0.65	4966
ABD	500		2.05	0.57	4228
IBD			2.07	0.57	4258

Table 2.5.3: Posterior Means, Standard Deviations and Effective Sample Sizes for **completely observed** Household-based SIS epidemic for parameter sets $\boldsymbol{\theta} = (\lambda, \beta, \gamma)' = (2, 1.5, 2.5)'$ from 1×10^5 iterations after 2×10^4 burn-in.

Though we observe that the algorithms performed better for when the sample sizes are 200 and 500 than when the sample size was 100, though only slightly. The effect of sample size is not well established. We note that the ABD-based algorithms were about 1.5 times faster that the IBD-based MCMC algorithms. Next we consider when the data are only partially observed. Four different forms of missingnesses were considered according to Section 2.5.1: individual's infectious status missing completely at random at random observation time points or A; randomly selected individual missing at random across or the time points or B; randomly selected observation time point missing completely or $C^{(1)}$ (for this, we simply deleted the entire unobserved column and carried on with the available data as if it were completely observed ab initio) or $C^{(2)}$ (for this, we imputed the missing time points and data via data augmentation). Table 2.5.4 compares the posterior estimates of mean, standard deviation (SD) and effective sample size (ESS) obtained from the MCMC outputs of the last $80,000$ iterations for (λ, β, γ)$'$= $(0.20, 0.15, 0.25)'$ or $c = 0.1$, and for 50% and 90% missingness across the various types of missingnesses and across the various sample sizes ($N = 100, 200, 500$) of the IBD.

Missingness	Size (N)	% Missing	Mean	SD	ESS
			$\lambda(\beta)\gamma$	$\lambda(\beta)\gamma$	$\lambda(\beta)\gamma$
		$P = 50\%$			
A	100		0.21 (0.19) 0.23	0.03 (0.04) 0.04	7445 (7270) 6054
B			0.23 (0.09) 0.22	0.04 (0.03) 0.03	575 (622) 1462
$C^{(1)}$			0.35 (0.31) 0.45	0.05 (0.10) 0.13	4754 (6295) 6155
$C^{(2)}$			0.24 (0.23) 0.29	0.04 (0.07) 0.09	2411 (1691) 1586
A	200		0.20 (0.21) 0.31	0.02 (0.03) 0.04	6883 (6946) 6578
B			0.21 (0.08) 0.21	0.03 (0.02) 0.03	1913 (2305) 2357
$C^{(1)}$			0.36 (0.32) 0.47	0.04 (0.07) 0.08	3384 (4555) 6028
$C^{(2)}$			0.20 (0.20) 0.30	0.02 (0.04) 0.05	2215 (1153) 1777
A	500		0.20 (0.17) 0.26	0.01 (0.02) 0.02	7310 (6420) 6777
B			0.21 (0.08) 0.21	0.02 (0.02) 0.02	423 (297) 600
$C^{(1)}$			0.35 (0.28) 0.46	0.03 (0.04) 0.05	4919 (3571) 4728
$C^{(2)}$			0.21 (0.18) 0.29	0.02 (0.03) 0.03	1783 (1191) 1015
		$P = 90\%$			
A	100		0.22 (0.25) 0.30	0.03 (0.07) 0.08	4721 (2645) 3004
B			0.32 (0.29) 0.51	0.13 (0.33) 0.40	187 (230) 150
$C^{(1)}$			0.39 (0.40) 0.43	0.08 (0.15) 0.21	6220 (7665) 6286
$C^{(2)}$			0.24 (0.20) 0.24	0.14 (0.18) 0.23	179 (280) 187
A	200		0.18 (0.20) 0.27	0.02 (0.04) 0.04	5059 (3576) 3729
B			0.22 (0.09) 0.36	0.05 (0.09) 0.17	335 (190) 135
$C^{(1)}$			0.45 (0.50) 0.66	0.08 (0.14) 0.20	7594 (4407) 5900
$C^{(2)}$			0.21 (0.16) 0.28	0.04 (0.05) 0.08	820 (977) 756
A	500		0.18 (0.17) 0.25	0.01 (0.02) 0.03	4673 (3326) 3152
B			0.21 (0.10) 0.31	0.04 (0.05) 0.07	149 (313) 165
$C^{(1)}$			0.48 (0.41) 0.62	0.06 (0.08) 0.13	7646 (6619) 7011
$C^{(2)}$			0.19 (0.17) 0.25	0.02 (0.03) 0.04	318 (551) 401

Table 2.5.4: Posterior Means, Standard Deviations and Effective Sample Sizes for **partially observed** Household-based SIS epidemic for parameter $\boldsymbol{\theta} = (\lambda,\ \beta,\ \gamma)' = (0.20, 0.15, 0.25)'$ or $c = 0.1$ from 1×10^5 iterations after 2×10^4 burn-in.

It happened that when 50% data are missing, the missingness of type A appeared to yield more accurate estimates across the three sample sizes as well as having the highest amount of ESS across the samples. Immediately behind A in terms of performance is $C^{(1)}$, while the missingness of type B showed the worst performance. On the other hand, when the data are up to 90% missing, the missingness of forms A and $C^{(2)}$ appear to perform better than the rest in terms of parameters estimates (mean). However, the missingness of type $C^{(1)}$ has the highest ESS, albeit with the least accurate estimate. However, the standard deviation of A is the least throughout. The results show that the missingness of form A may be said to have the overall best performance. A further assessment of the effects of missing data on the performance of our algorithms is presented on Table 2.5.5, which compares the posterior estimates for across the four types of missingness considered for $c = 1$ and for $N = 500$ using the individual-based data with $(P\% = 10\%, 30\%, 70\%, 90\%)$. The results on Table 2.5.5 show that when only up to 30% of the data are missing, the algorithms show fairly similar performance across the various types of missingness. However as the proportion of missing data increases to 90%, $C^{(2)}$ rapidly deteriorates, while $C^{(1)}$ appeared to outperform the rest followed by A indicating that the effect of sample size might be significant in this case.

Missingness	% missing	**Mean**	**SD**	**ESS**
		$\lambda(\beta)\gamma$	$\lambda(\beta)\gamma$	$\lambda(\beta)\gamma$
A	10%	1.828 (1.181) 2.045	0.506 (0.350) 0.575	4701 (4977) 4763
B		1.700 (1.109) 1.918	0.488 (0.345) 0.558	3528 (4128) 3432
$C^{(1)}$		1.828 (1.449) 2.277	0.484 (0.404) 0.610	5011 (5214) 4998
$C^{(2)}$		1.646 (1.120) 1.905	0.472 (0.348) 0.558	3540 (3789) 3576
A	30%	1.586 (1.257) 1.976	0.446 (0.381) 0.568	3604 (3782) 3780
B		1.863 (0.929) 1.940	0.594 (0.332) 0.609	2791 (2735) 2976
$C^{(1)}$		1.369 (1.162) 1.744	0.353 (0.318) 0.467	5258 (5176) 5131
$C^{(2)}$		1.536 (0.933) 1.692	0.533 (0.350) 0.595	2529 (2599) 2602
A	70%	1.455 (1.137) 1.784	0.442 (0.369) 0.549	2546 (2596) 2469
B		1.964 (0.512) 1.234	1.207 (0.500) 0.615	408 (1089) 752
$C^{(1)}$		2.026 (1.261) 2.210	0.630 (0.480) 0.688	5280 (5620) 5762
$C^{(2)}$		1.059 (0.848) 1.275	0.460 (0.393) 0.567	717 (538) 541
A	90%	1.207 (0.916) 1.450	0.465 (0.370) 0.561	1307 (1379) 1224
B		1.237 (0.968) 1.586	0.789 (0.570) 0.663	315 (192) 423
$C^{(1)}$		1.910 (1.338) 2.144	0.546 (0.443) 0.637	5645 (5423) 5530
$C^{(2)}$		0.922 (0.797) 1.159	0.450 (0.416) 0.582	520 (529) 491

Table 2.5.5: Posterior Means, Standard Deviations and Effective Sample Sizes for **partially observed** individual-based data (IBD) ($P\% = 10\%, 30\%, 70\%, 90\%$) for parameter $\boldsymbol{\theta} = (\lambda,\ \beta,\ \gamma)' = (2.0, 1.5, 2.5)'$ or $c = 1$ from 1×10^5 iterations after 2×10^4 burn-in. $N = 500$.

Furthermore, we explored the relationships between different parameter values. Figure 2.5.4 shows the paired contour plots, density and correlation plots for $(\lambda, \beta, \gamma) =$

$(10, 7.5, 12.5)$ or $c = 5$ for the IBD at 30% missing. The paired plot show that there is high correlation between the infection and recovery rates, while the correlation between the two rates of infection is low at 0.467.

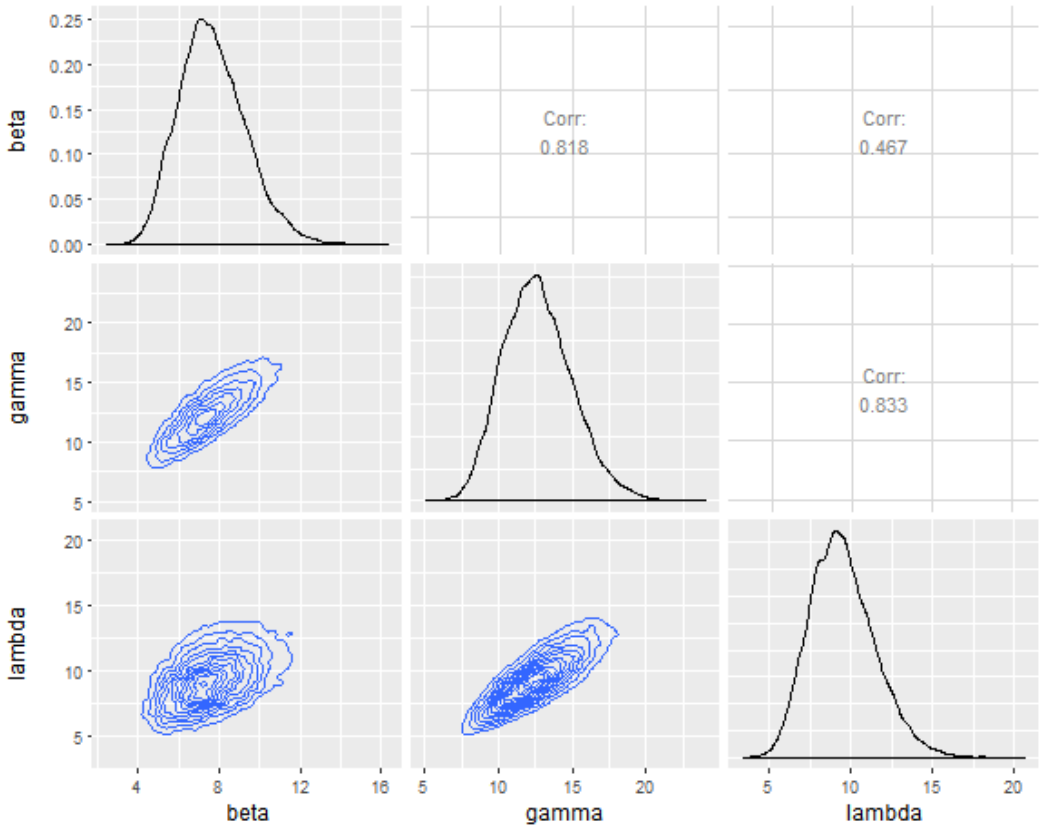

Figure 2.5.3: Paired plots. The contour plots (blues) show that there is a strong correlation between β vs γ and a weak correlation exits between γ vs β

.

As the values of c further increases $(c > 1)$ or as the parameter , $\boldsymbol{\theta}$ values get larger (e.g., Figure 2.5.4), we see that only a few changes can be observed as the transition probability matrix, $\mathbf{Q}_t$, quickly approaches the stationary distribution, $\boldsymbol{\pi}$, in that each row of $\mathbf{Q}_t$ contains the same set of elements. In other words, if t, s are quite big, then $\mathbf{Q}_t \approx \mathbf{Q}_s$.

Finally, the findings from the simulation study show that the proposed models are robust and especially with the data form A (partially observed data) even when data are up to 90% missing. We note that in all cases, the MCMC algorithms based upon the IBD and ABD peformed nearly equally well, although the IBD-based algorithms only slightly outperformed the ABD-based algorithms. However, the ABD-based algorithms are somewhat faster to run and is about 1.5 times faster than the more informative IBD. For example, the cpu time required to run a 1×10^5 iterations of ABD on a Dell computer Intel (R) Core (TM) with 64-bit Operating System is approximately 2 hours, while same number of iterations requires nearly 3 hours cpu time to run the MCMC for IBD on the same computer.

2.6 Conclusions

In this chapter, we introduced stochastic household-based SIS epidemic models in a closed population. Two main data forms were considered- the individual-based data (IBD) and the aggregate-based data (ABD). We outlined the procedures for the analysis of the models in Bayesian framework and developed robust and easy-to-use MCMC algorithms for the analysis of such infectious diseases data. Two main scenarios were considered- the completely observed data case and the partially observed data case. Analysis involving the completely observed data is straightforward as outlined in Section 2.4.1 in that only $\pi(\boldsymbol{\theta}|\mathbf{x})$ is updated. For the partially observed household-based SIS data case, we developed robust and flexible data augmentation algorithms as outline in Section 2.4.3. The simulation study carried out in Section 2.5 shows that our models are robust for both the completely observed data case and the partially observed data case even when the $100P\%$ is up to 90% (especially for data form A and for moderate t

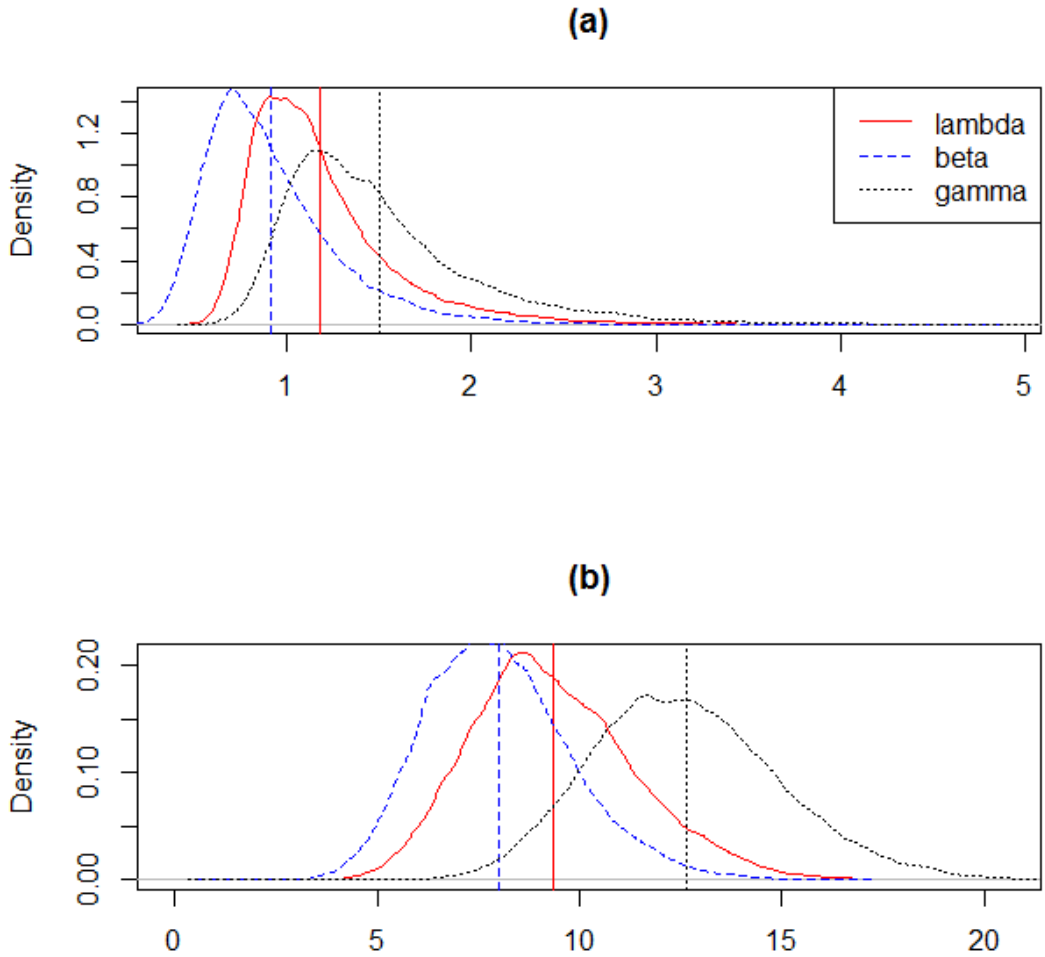

Figure 2.5.4: Posterior density plots of IBD, $c = 5$ with $Gamma(1, 1)$ priors (a) and $Gamma(10, 1)$ priors (b) for $N = 200$ at 10% missing of data form A. The vertical lines are the posterior means.

and parameter values).

We will extend the models introduced here to allow for changing population sizes over time as well as the incorporation of spatial element in Chapter 3 and later allow for interacting infectious diseases (or co-infection) in Chapter 4.

Chapter 3

Open Population, Spatial SIS Model

3.1 Introduction

This chapter is divided into two parts. The first part of this chapter seeks to develop Bayesian inference methods for the analysis of endemic diseases within a population partitioned into households, such that the household sizes are allowed to vary over time. The main focus is on the estimation of the major parameters which are the main drivers of the epidemic, for example, the global and local rates of infection (or λ and β). To do this, we develop efficient MCMC algorithms for sampling from the posterior distribution of the parameters given the data, $\pi(\boldsymbol{\theta}|\mathbf{x})$, say. The second part of this chapter is concerned with the development of a spatial epidemic model which allows the global force of infection λ to depend on the spatial locations of the households. This is modelled using Gaussian process, such that the global force of infection is a function of Gaussian random field (GRF) realizations. We used a distance dependent correlation function to account for spatial variations in the data. As in the first part, the main focus is also on

the development of Bayesian inference framework for the estimation of the varying risks of infection as well as other major parameters of the model including the parameters of the correlation function. The models and the implementation of the MCMC algorithms developed in both the first and second parts of this chapter are illustrated using simulated data sets as well as real life data set on the spread of tick-borne diseases among Tanzania cattle.

The rest of this chapter is organized as follows: In Section 3.2 we give an overview of the open population models and in Section 3.3 we describe the two most prevalent forms of endemic disease data from the open population model. In Section 3.4, we briefly describe the model setup which is very similar to that described in Section 2.3. MCMC algorithms developed for the analysis of an open population SIS epidemic data, with details on the development of the data augmentation schema employed are given in Section 3.5. In Section 3.6, we give an overview of the development of the spatial disease model. MCMC algorithms for the estimation of the parameters of the spatial epidemic model are developed in Section 3.8. Furthermore, in Section 3.9, we illustrate the implementation of the MCMC algorithms developed using simulated data sets. An application of the MCMC algorithms to the analyses of a real life data set on tick-borne diseases among Tanzania cattle is shown in Section 3.10. Finally, in Section 3.11, we give concluding remarks with focus on the findings from analyses of the simulated data set and the real life data.

3.2 Open Population SIS epidemic

In Chapter 2 we studied SIS epidemics within households in which the population is assumed to be closed, *i.e.*, no births, no deaths, no immigration and no emigration. The size of a given household h is independent of time. Closed population assumption

are reasonable if there is no significant change in the population over the study period. Closed population SIS epidemic model provides a first order approximation and might be used in modelling the spread of *pneumococcus* amongst school children and in modelling diseases such as meningitis, streptococcal sore throat and tuberculosis (Hethcote, 1976). Closed population assumption can also be suitable in modelling epidemics that last longer, but in which disease-related deaths are so insignificant and natural deaths are immediately balanced by births (Hethcote and van den Driessche, 1995). For most endemic diseases such as measles, it is highly unlikely that the epidemic would not make a major impact on the population demography. In some kind of endemic diseases, the number of deaths (both natural and disease related) might be lower than births into the population. In such situations, where there are significant demographic changes in the population, the closed population assumption becomes unrealistic. Therefore, there is need to develop epidemic models which incorporate the more realistic assumption that the population sizes vary over time.

So far, only a few studies have explored open or varying population epidemic model, see, for example, Hethcote and van den Driessche (1995), O'Neill (1996), Clancy et al. (2001) and Greenhalgh et al. (2016). O'Neill (1996) considers an open population SIR epidemic model and incorporates immigration ($\mu_1 > 0$) and emigration ($\mu_2 > 0$) parameters into the susceptible class of the model and then used coupling argument to illustrate the strong convergence of sequence of infectives to the birth-and-death process. Clancy et al. (2001) studies long term behavior of an open population stochastic epidemic model of the SIR type incorporating birth parameter (μ) into the susceptible class of the model, and using diffusion approximations to describe the temporal behavior of the epidemic. Hethcote and van den Driessche (1995) studies the asymptotic behavior of varying population SIS epidemic model incorporating births ($b > 0$) and deaths ($d > 0$) into the modelling. Greenhalgh et al. (2016) studies a two-dimensional stochastic differential

equation (SDE) SIS epidemic model incorporating births and deaths as stochastic processes. In this chapter, interest is on the development of Bayesian inference methods for such SIS epidemic models whose population sizes vary over time rather than modelling the infectious process itself. We note that none of the studies mentioned above focuses on developing inference methods for estimation of the parameters of an open population stochastic SIS epidemic model.

Given the foregoing, in this chapter, we develop inference methods for open population stochastic SIS epidemic models in Bayesian framework primarily using Markov chain Monte Carlo (MCMC) algorithms. We shall focus on household based stochastic SIS epidemic models with a setup similar to the closed population model studied in Chapter 2. The major differences between the model studied in Chapter 2 and the model we consider here is that here we allow individuals to enter and leave a given household at a point in time. This makes the household size to vary at different observation time points with the possibility of having new individuals not previously present in the household at time t being present at time $t+1$ or individuals present at time t leaving the household at $t+1$. For as far as we are aware of, there has never been any work on the development of Bayesian inference methods using MCMC for open population stochastic SIS household epidemic models. Therefore, we seek to fill this gap in literature by developing novel MCMC algorithms via data augmentation for the analysis of open population stochastic SIS household epidemic model. In what follows, we shall give a detailed description of the two most prevalent endemic disease data we consider here.

3.3 Data Description

In this section, we describe the endemic disease data we consider as follows. For a given household, let $h(t)$ denote the size of the household at time t, then the household size

is said to vary whenever $h(t+1) \neq h(t)$, *i.e.*, we can have that $h(t+1) < h(t)$ or that $h(t+1) > h(t)$. For ease of exposition we shall use *arrival*(*departure*) to mean an individual or individuals joining (leaving) a given household at a point in time. Therefore, an event of arrival is assumed to occur approximately halfway between time t and time $t+1$ whenever $h(t+1) > h(t)$. Similarly, the event of departure is assumed to have occur approximately halfway between t and $t+1$, whenever $h(t+1) < h(t)$. Individuals (animals) join a given household (farm) through birth or immigration (birth or acquisition), while individuals (animals) leave a given household (farm) following deaths or emigration (deaths or animals being sold).

We shall now describe the open population epidemic data for the two most prevalent endemic disease data forms, namely, individual-based data (IBD) and aggregate-based data (ABD).

3.3.1 Individual-based data (IBD)

For $k = 1, 2, \ldots, n$, let $h(t_k)$ denote the size of a given household at the k^{th} observation time point. Let H denote the number of distinct individuals ever in the household across the n observation time points. For $j = 1, 2, \ldots, H$, let $x_j^*(t_k)$ denote the infection status of individual j at time t_k. We encode $x_j^*(t_k) = 5$ if individual j is not in the household at time t_k, otherwise $x_j^*(t_k) = x_j(t_k) \in \{0, 1\}$, where 0 denotes susceptible and 1 denotes infective. Then, $\mathbf{x}(t_k) = \big(x_1^*(t_k), x_2^*(t_k), \ldots, x_H^*(t_k)\big)$ is the state of the household at the k^{th} observation time point. Therefore, $\mathbf{x(t)} = \big(\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_n)\big)$ is the open population individual-based endemic disease data for a given household over n observation time points. Table 3.3.1 shows an example of an open population household SIS individual-based epidemic data with $H(= 5)$ distinct individuals across the $n(= 9)$ observation time points. Note that we encode $x_j^*(t_k) = 2$, when we are not

able to ascertain the j^{th} individual's infectious status even though he is a member of the household at time t_k. Reasons why the infectious status of an individual might not be known include refusal or unavailability to submit to a clinical tests. Note also that we assume that the interval $[t, t + \Delta t)$ is small enough to allow only the occurence of one event at a time, *that is*, event of arrival or departure.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
I_1	0	2	0	0	1	0	0	0	1
I_2	0	0	0	0	1	0	0	5	5
I_3	5	0	0	0	0	0	1	0	0
I_4	5	5	5	5	5	0	0	2	0
I_5	5	5	5	5	5	5	5	5	0

Table 3.3.1: Individual-based data (IBD) for an open population SIS epidemic.

From Table 3.3.1, we see that individuals $3, 4$ and 5 or (I_3, I_4 and I_5) were observed in the household at the 2^{nd}, 6^{th} and 9^{th} observation time points, respectively. Also, observe that individual 2 departed the household before the 8^{th} observation time point (t_8). As noted above, arrival or departure occurs at an unobserved time point $\tilde{t} \approx \frac{1}{2}(t_k + t_{k+1})$, where $\tilde{t} \in \mathbb{N}$, for $k = 1, 2, \ldots, n$. This concept is further illustrated in Section 3.5.1.

3.3.2 Aggregate-based data (ABD)

Let $\tilde{x}(t_k) \in \{0, 1, \ldots, h(t_k)\}$ denote the number of infectives in a given household of size $h(t_k)$ at time t_k. Therefore, $\tilde{\mathbf{x}}(\mathbf{t}) = \big(\tilde{x}(t_1), \tilde{x}(t_2), \ldots, \tilde{x}(t_n)\big)$ denotes the aggregated open population SIS epidemic data of a given household over n set of observation time points. Table 3.3.2 shows an example of an open population endemic disease aggregated data. Observe that Table 3.3.2 is the row sum of Table 3.3.1 for values of $1s$ and $0s$, *i.e.*,

$\tilde{x}(t_k) = \sum \mathbf{x}(t_k) \mathbb{1}_{\left\{x_j^*(t_k) \in \{0,1\}\right\}}$.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
Number of Infectives	0	0	0	0	2	0	1	0	1
Household size	2	3	3	3	3	4	4	3	4

Table 3.3.2: Aggregate-based Data (ABD) for open population models.

3.4 Model setup (Open population)

In this section, we describe the construction of the open population model. Again, we note that the model construction is very similar to that described in Section 2.3, except that here the household size varies over time. First, let us briefly recall that our model assumes that there is a global force of infection $\lambda > 0$ and a local rate of infection $\beta > 0$. Infections occur when an infective makes a successful infectious contact with a susceptible chosen uniformly at random from a given household. All the contacts are made at points of mutually independent Poisson processes. At the end of its infectious period, which is distributed exponentially with mean γ^{-1}, an individual recovers at rate $\gamma > 0$ and immediately returns to the susceptible state and can be reinfected. Therefore, there is no removed state as recovery from the disease does not confer immunity. Only infectious contacts with susceptibles confer infection and there is no latent period so that the infected individual becomes infectious immediately. Therefore, the only transitions allowed at a point in time are from susceptible to infective (S $\rightarrow$ I) or from infective to susceptible (I $\rightarrow$ S), hence SIS epidemic model.

We shall now give details on how the infinitesimal transition rate matrix (G-matrix) is constructed and how the corresponding transition probability matrix (Q-matrix) is

calculated.

3.4.1 Infinitesimal Transition rate Matrix (G-matrix)

First of all, we note that the infinitesimal transition rate matrix is calculated as described in Section 2.3.1. Recall that the G-matrix for the individual-based data (IBD) is calculated according to (2.3.1) as follows

$$g_{\mathbf{uv}}^{(h)} = \begin{cases} \lambda + \beta \sum_{i=1}^{h} u_i & \text{if } u_j = 0 \text{ and } \mathbf{v} = \mathbf{u} + \mathbf{e}_j, \\ \gamma & \text{if } u_j = 1 \text{ and } \mathbf{v} = \mathbf{u} - \mathbf{e}_j \\ -\sum\limits_{\mathbf{w} \neq \mathbf{u}} g_{\mathbf{uw}}^{(h)} & \text{if } \mathbf{v} = \mathbf{u} \\ 0 & \text{Otherwise} \end{cases}$$

for $\lambda, \beta, \gamma > 0$ and $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w} \in \mathcal{S}$, where $u_j \in \{0, 1\}$ is the infection status of the j^{th} individual of a household of size h.

Similarly, the G-matrix for the aggregate-based data (ABD) is calculated according to (2.3.3) as

$$g_{m,n}^{(h)} = \begin{cases} \beta m(h-m) + \lambda(h-m) & \text{if } n = m+1 \\ \gamma m & \text{if } n = m-1 \\ 0 & \text{if } |n-m| > 1 \\ -\sum\limits_{k \neq m} g_{m,k}^{(h)} & \text{if } n = m \end{cases}$$

for $\lambda, \beta, \gamma > 0$ and m, n, $k \in \mathcal{S}$, where m is the number of infectives in a household of size h at a point in time.

Now, let $\{h_1, h_2, \ldots, h_p\}$ denote an ordered set of p distinct sizes of the household across the n observation time points. For both the IBD and ABD, we calculate the household size dependent G-matrix for each of the p distinct sizes, so that for every given household

observed n times, we obtain the matrices $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \ldots, \mathbf{G}^{(h_p)}$ for the p distinct sizes of the household across the observation time points. This helps to save computer memory and minimize computational costs that would have been incurred should we calculate the G-matrix at every time point whenever the household size changes.

3.4.2 Infinitesimal Transition probability Matrix (Q-matrix)

We shall now describe how the transition probability matrix or the Q-matrix is calculated. The Q-matrix is calculated as described in Section 2.3.1. Observe that the Q-matrix depends on both the household size h and the time difference t. Therefore, for the open-population model in which household sizes vary at time points, we calculate the Q-matrix for every given t and for every given distinct h according to (2.3.7) by taking matrix exponent of the product of $t(= t_{k+1} - t_k)$ and the corresponding G-matrix for the household at time t_k, *i.e.*, $\mathbf{Q}_t^{(h(t_k))} = \exp(t\mathbf{G}^{(h(t_k))})$.

In what follows, we shall outline Bayesian inference procedure for an open population SIS epidemic data.

3.5 Bayesian Inference for Open population SIS epidemic model

In this section, we outline Bayesian inference framework for the estimation of the key parameters of the epidemic, namely, the global force of infection λ, the local rate of infection β and the recovery rate γ. Given that the size of a given household varies over time, it implies individuals leave and join the household at points in time. We actually do not observe the actual point in time when an individual leaves or joins the household. Therefore, the open population endemic disease data is only partially observed and this makes inference on such data more complicated. Progress can however be made

using appropriate data imputation strategies. In this section, we develop a class of data augmentation schema in MCMC framework for efficient analysis of an open population endemic disease data. We shall begin with the individual-based data (IBD).

3.5.1 Generic setup

Let $\mathbf{y}$ denote the partially observed data generated from the open population SIS model with parameters $\boldsymbol{\theta} = (\lambda, \beta, \gamma)$. It is well known that the likelihood function $L(\boldsymbol{\theta}|\mathbf{y}) = \pi(\mathbf{y}|\boldsymbol{\theta})$ is rarely tractable. We augment $\mathbf{y}$ with $\mathbf{z}$ and set $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ as the complete data, then the likelihood $L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = \pi(\mathbf{x} = (\mathbf{z}, \mathbf{y})|\boldsymbol{\theta})$ becomes tractable. Here, the extra information $\mathbf{z}$ are the imputed observations and time points at which individuals join or leave a given household. We proceed as follows.

Individual-based data (IBD)

For $k = 1, 2, \ldots, n$, let $h(t_k)$ denote the number of individuals in the household at time t_k. Let $\mathbf{y}(t_k) = \big(y_1^*(t_k), y_2^*(t_k), \ldots, y_H^*(t_k)\big)$ denote the state of the household at the k^{th} observation time point, where H is the number of distinct individuals observed in the household across the n observation time points and $y_j^*(t_k)$ denotes the observed infection status of individual j at time t_k and is encoded according to (3.5.1) ($j = 1, 2, \ldots, H$). When an event of arrival or departure occurs, we assume that the event must have occurred at a point approximately halfway in between t_k and t_{k+1}, and impute the time point $\tilde{t}_l \approx \frac{1}{2}(t_k + t_{k+1})$, where $l = 1, 2, \ldots, m$, and m is the number of time points imputed in the household over n observation time points. Note that when $h(t_{k+1}) > h(t_k)$, we set $\mathbf{z}(\tilde{t}_l) = \mathbf{x}(t_k)$ with the infectious status of individual j at time $\tilde{t}_l$ coded 3, *i.e.*, $\mathbf{z}(\tilde{t}_l) = \big(x_1^*(t_k), x_2^*(t_k), x_{j-1}^*(t_k), x_j^*(t_k) = 3, x_{j+1}^*(t_k) \ldots, x_H^*(t_k)\big)$. On the other hand, when $h(t_{k+1}) < h(t_k)$, we set $\mathbf{z}(\tilde{t}_l) = \mathbf{x}(t_k)$ with the infectious status of individual j at time $\tilde{t}_l$ coded 4, *i.e.*, $\mathbf{z}(\tilde{t}_l) = \big(x_1^*(t_k), x_2^*(t_k), x_{j-1}^*(t_k), x_j^*(t_k) = 4, x_{j+1}^*(t_k) \ldots, x_H^*(t_k)\big)$.

This data imputation scheme is illustrated in Table 3.5.1 for the open population SIS data in Table 3.3.1.

$$y_j^*(t_k) = \begin{cases} 0 & \textit{if susceptible}, \\ 1 & \textit{if infective}, \\ 2 & \textit{if status unknown}, \\ 3 & \textit{if individual joins the household}, \\ 4 & \textit{if individual exits household} \ , \\ 5 & \textit{if not in the population}. \end{cases} \tag{3.5.1}$$

	t_1	$\tilde{t}_1$	t_2	t_3	t_4	t_5	$\tilde{t}_2$	t_6	t_7	$\tilde{t}_3$	t_8	$\tilde{t}_4$	t_9
I_1	0	2	2	0	0	1	2	0	0	2	0	2	1
I_2	0	2	0	0	0	1	2	0	0	4	5	5	5
I_3	5	3	0	0	0	0	2	0	1	2	0	2	0
I_4	5	5	5	5	5	5	3	0	0	2	2	2	0
I_5	5	5	5	5	5	5	5	5	5	5	5	3	0

Table 3.5.1: Individual-based Data (IBD) with varying population sizes over time with imputed time points and coded according to (3.5.1).

The illustration on Table 3.5.1 shows that we impute a total of four $m(=4)$ time points, namely, $\tilde{t}_1$, $\tilde{t}_2$, $\tilde{t}_3$ and $\tilde{t}_4$, which are the unobserved time points at which individuals exit the household or at which individuals join the household. For example, $\tilde{t}_1$ is the first imputed observation time point at which individual 3 (or I_3) joins the household. Similarly, the third and fourth imputed time points $\tilde{t}_3$ and $\tilde{t}_4$ are the times at which the second individual (or I_2) leaves the household and at which the fifth individual (or

I_5) joins the household, respectively. Therefore, there are $H(=5)$ distinct individuals observed in the household for the $n(=9)$ observation time points. The data imputation procedures described above is repeated for each $i = 1, 2, \ldots, N$ households which we assume to be independent throughout this chapter.

In what follows, we shall now outline the implementation of MCMC algorithms for the open population SIS epidemic data.

3.5.2 Data Augmentation (MCMC)

We now follow the data augmentation steps described in Section 2.4.3 and proceed as follows. Given the augmented data $\mathbf{x} = (\mathbf{z}, \mathbf{y})$, obtain samples from the joint posterior $\pi(\boldsymbol{\theta}, \mathbf{x})$ by iteratively sampling

1. $\boldsymbol{\theta}$ `from` $\pi(\boldsymbol{\theta}|\mathtt{x} = (\mathtt{z}, \mathtt{y}))$ `and`

2. `z from` $\pi(\mathtt{z}|\mathtt{y}, \boldsymbol{\theta})$.

To update $\pi(\boldsymbol{\theta}|\mathbf{x} = (\mathbf{z}, \mathbf{y}))$, we need to choose appropriate prior distributions of the parameters, $\pi(\boldsymbol{\theta})$, calculate the likelihood function, $L(\boldsymbol{\theta}; \mathbf{x} = (\mathbf{z}, \mathbf{y}))$, obtain the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$, and then construct an efficient MCMC algorithm whose stationary distribution is our target density. Figure 3.5.1 shows the schematic representation of the model. Observe that by Figure 3.5.1 we have assumed that the observed data $\mathbf{y}$ is conditionally independent of $\boldsymbol{\theta}$ given the *complete* data $\mathbf{x} = (\mathbf{z}, \mathbf{y})$.

Priors

Given that our parameter values are positive and real-valued, we assign independent Gamma priors to $\boldsymbol{\theta} = \{\lambda, \beta, \gamma\}$ so that

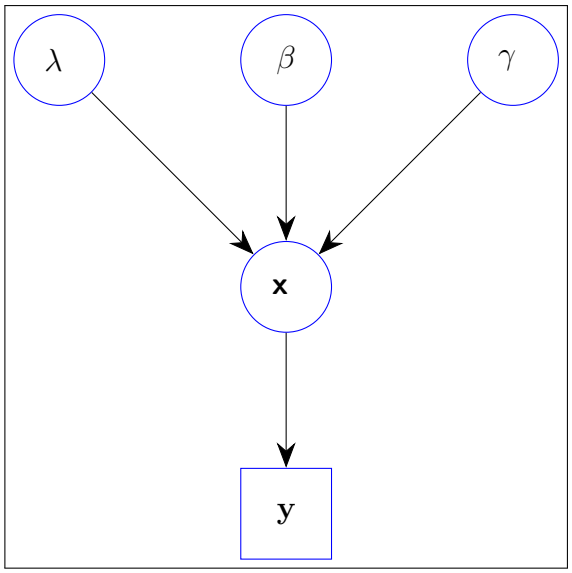

Figure 3.5.1: Schematic representation of the open population model.

$$\begin{aligned}
\lambda &\sim Gamma(A_\lambda, B_\lambda) \\
\beta &\sim Gamma(A_\beta, B_\beta) \\
\gamma &\sim Gamma(A_\gamma, B_\gamma) \quad . \qquad (3.5.2)
\end{aligned}$$

where $A_\lambda > 0$, $B_\lambda > 0$, $A_\beta > 0$, $B_\beta > 0$, $A_\gamma > 0$ and $B_\gamma > 0$ are hyper-parameters. Note that these priors can be chosen to be informative or uninformative. For example, we make the Gamma priors uninformative by choosing shape parameters to be small.

Posterior Distribution

The posterior distribution of the parameters given the data, $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = \pi(\boldsymbol{\theta}|\mathbf{x})$ (by conditional independence assumption) is then calculated according to (2.4.7) as

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\mathbf{x}) \quad &\propto \quad \left\{ \prod_{i=1}^{N} \prod_{k=2}^{n_i} \left\{ \pi\big(\mathbf{x}_i(t_{ik}) | \mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta}\big) \right\} \right\} \\
&\times \quad \lambda^{A_\lambda - 1} e^{-B_\lambda \lambda} \times \beta^{A_\beta - 1} e^{-B_\beta \beta} \times \gamma^{A_\gamma - 1} e^{-B_\gamma \gamma},
\end{aligned}$$

since the marginal distribution of the data $\pi(\mathbf{x})$ and the constants $B_\lambda^{A_\lambda}/\Gamma(A_\lambda)$, $B_\lambda^{A_\lambda}/\Gamma(A_\lambda)$ and $B_\lambda^{A_\lambda}/\Gamma(A_\lambda)$ are independent of $\boldsymbol{\theta}\{=(\lambda,\beta,\gamma)'\}$ and are not necessary for drawing samples from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$ using MCMC.

Next, update $\pi(\boldsymbol{\theta}|\mathbf{x})$ using adaptive Random Walk Metropolis with multivariate Gaussian proposal according to Algorithm 4.

Also, same as in Section 2.4.3, we update $\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$, for the three time points t_{ik-1}, t_{ik} and t_{ik+1}. First, obtain the probability of observing the complete data $\mathbf{x}$ given the observed data $\mathbf{y}$ and the model parameters $\boldsymbol{\theta}$ given by

$$\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y}) \quad \propto \quad \pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\theta}), \tag{3.5.3}$$

since $\mathbf{y}$ is conditionally independent of $\boldsymbol{\theta}$ given $\mathbf{x}$. Then, by the independent households assumption we have

$$\begin{aligned}\pi(\mathbf{x}=(\mathbf{z},\mathbf{y})|\boldsymbol{\theta},\mathbf{y}) \quad &\propto \quad \prod_{i=1}^{N}\prod_{k=2}^{n_i}\Big\{\pi\big(\mathbf{y}_i(t_{ik})|\mathbf{x}_i(t_{ik})\big) \\ &\times \quad \pi\big(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta}\big)\pi\big(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}),\boldsymbol{\theta}\big)\Big\}. \end{aligned}\tag{3.5.4}$$

Then to update $\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$, we update one $z_{ij}(t_{ik})$ at a time and calculate the following probabilities. First calculate the probability

$$\begin{aligned}\pi(z_{ij}(t_{ik})|\boldsymbol{\theta},\mathbf{x}_{-ij}(t_{ik})) \quad &\propto \quad \pi(\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta}) \\ &\times \quad \pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z_{ij}(t_{ik}),\boldsymbol{\theta}), \end{aligned}\tag{3.5.5}$$

where $\mathbf{x}_{-ij}(t_{ik})$ is the complete data vector for the infectious state of household i, $\mathbf{x}_i(\mathbf{t}_i)$, without $z_{ij}(t_{ik})$ at time t_{ik}.

Then, when an individual joins the household, we are only interested in the state transition from time point t_{ik} to t_{ik+1}. We calculate the probability

$$\pi(z_{ij}(t_{ik})|\boldsymbol{\theta}, \mathbf{x}_{-ij}(t_{ik})) \quad \propto \quad \pi(\mathbf{x}_i(t_{ik+1})|(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik})), \boldsymbol{\theta}). \tag{3.5.6}$$

Similarly, when an individual leaves the household, we are interested on the state transitions form time point t_{ik-1} to t_{ik}. We calculate the probability

$$\pi(z_{ij}(t_{ik})|\boldsymbol{\theta}, \mathbf{x}_{-ij}(t_{ik})) \quad \propto \quad \pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik})|(\mathbf{x}_i(t_{ik-1})), \boldsymbol{\theta}). \tag{3.5.7}$$

Finally, update $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ using the Independence Sampler steps described below.

3.5.3 Independence Sampler

We outline the Independence Sampler steps for updating $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ as follows.

Step 0: Initialize the complete data $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ by using methods similar to those described in Section 2.4.4. One of the methods suggested is to set $x_{ij}(t_{ik}) = y^*_{ij}(t_{ik})$, if $y^*_{ij}(t_{ik}) < 2$, and set $x_{ij}(t_{ik}) = 0$ or choose $x_{ij}(t_{ik})$ uniformly from $\{0, 1\}$, if $y^*_{ij}(t_{ik})$ is equal to 2 (status unknown), 3 (individual joining the household), or 4 (individual exiting the household), see (3.5.1). For $r \geq 0$, set the current state of the Markov chain $X^{(r)} = \big(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik})\big)$ $(k = 1, 2, \ldots, n_i; j = 1, 2, \ldots, h(t_k))$. Then, proceed as follows:

Step 1: Propose to switch states, *i.e.*, $z'_{ij}(t_{ik}) = 1 - z_{ij}(t_{ik})$, set $\mathbf{x}^{prop} = \big(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z'_{ij}(t_{ik})\big)$.

(a) If $y^*_{ij}(t_{ik}) = 2$, accept $z'_{ij}(t_{ik})$ with probability $\Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik}))$ given by

$$\min\left\{1, \frac{\pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z'_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}), \boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z'_{ij}(t_{ik}), \boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}), \boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik}), \boldsymbol{\theta})}\right\}. \tag{3.5.8}$$

(b) If $y^*_{ij}(t_{ik}) = 3$, accept $z'_{ij}(t_{ik})$ with probability

$$\Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik})) = \min\left\{1, \frac{\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z'_{ij}(t_{ik}), \boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik}), \boldsymbol{\theta})}\right\}. \tag{3.5.9}$$

(c) If $y^*_{ij}(t_{ik}) = 4$, accept $z'_{ij}(t_{ik})$ with probability $\Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik}))$

$$\Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik})) = \min\left\{1, \frac{\pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z'_{ij}(t_{ik}), |\mathbf{x}_i(t_{ik-1}), \boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}), \boldsymbol{\theta})}\right\}. \tag{3.5.10}$$

Step 2: For each acceptance probability, $\Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik}))$, calculated in ***Step 1***,

(a) Draw u from $U[0, 1]$.

(b) If $u \leq \Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik}))$,

(i) accept $z'_{ij}(t_{ik})$,

(ii) set $\mathbf{X}^{(r+1)} = \mathbf{x}^{prop}$

(c) If $u > \Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik}))$,

(i) do not accept $z'_{ij}(t_{ik})$,

(ii) set $\mathbf{X}^{(r+1)} = \mathbf{X}^{(r)}$

Step 3: Repeat ***Step 1*** and ***Step 2*** for all $i = 1, 2, \ldots, N$.

We summarize the Independence Sampler algorithm below:

Aggregate-based Data (ABD)

Updating $\pi(\tilde{\mathbf{z}}|\mathbf{y}, \boldsymbol{\theta})$ proceeds in similar way as described in Section 2.4.5.

For a given household with H distinct individuals observed across n observation time

Algorithm 8 Open population: Independence Sampler

1. Intialize $\mathbf{x}^{(0)}$.

2. Propose $\mathbf{x}^{prop} = \big(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z'_{ij}(t_{ik})\big)$

3. Accept $\mathbf{x}^{prop}$ with probability $\Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik}))$

   - $\min\left\{1, \frac{\pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik})=z'_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z'_{ij}(t_{ik}),\boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z_{ij}(t_{ik}),\boldsymbol{\theta})}\right\}$
   - $\min\left\{1, \frac{\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z'_{ij}(t_{ik}),\boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z_{ij}(t_{ik}),\boldsymbol{\theta})}\right\}$, if $y^*_{ij}(t_k) = 3$.
   - $\min\left\{1, \frac{\pi(\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z'_{ij}(t_{ik}),|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik}),x_{ij}(t_{ik})=z_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})}\right\}$, if $y^*_{ij}(t_k) = 4$.

4. Draw u from $U[0, 1]$

   - If $u \leq \Delta(z_{ij}(t_{ik}), z'_{ij}(t_{ik}))$, set $\mathbf{X}^{(r+1)} = \mathbf{x}^{prop}$
   - Otherwise, set $\mathbf{X}^{(r+1)} = \mathbf{X}^{(r)}$

5. Repeat 2 and 3 for all $i = 1, 2, \ldots, N$.

points. Let $\tilde{s}(t_k)$ denote the number of individuals observed in the household out of which $\tilde{y}(t_k)$ infectives are observed given that in fact, there are actually $\tilde{x}(t_k)(\geq \tilde{y}(t_k))$ infectives in the household of size $h(t_k) \geq 1$ at time t_k, where $\tilde{s}(t_k) \leq h(t_k)$. When we observe every member of the household at time t_k, then the number of infectives observed is the actual number of infectives in the household at that point in time, *i.e.*, when $\tilde{s}(t_k) = h(t_k)$, then $\tilde{y}(t_k) = \tilde{x}(t_k)$. However, when $\tilde{s}(t_k) < h(t_k)$, we augment $\tilde{y}(t_k)$ with $z(t_k) = \{0, 1, \ldots, h(t_k)\}$, the possible number of infectives in the household at the k^{th} observation time point. We then calculate $\pi(\tilde{y}|\tilde{s}, h(t_k), \tilde{z})$, the probability of observing the observed number of infectives $\tilde{y}(t_k)$ given everything else according to (2.4.16) as follows

$$\pi(\tilde{y}|\tilde{s}, h(t_k), \tilde{z}) = \frac{\binom{\tilde{z}}{\tilde{y}(t_k)}\binom{h(t_k)-\tilde{z}}{\tilde{s}(t_k)-\tilde{y}(t_k)}}{\binom{h(t_k)}{\tilde{s}(t_k)}}. \tag{3.5.11}$$

We now choose $\tilde{x}(t_k)$ from $\{0, 1, \ldots, h(t_k)\}$ with probability $\tilde{P}_m / \sum_{j=0}^{h(t_k)} \tilde{P}_j$, where $\tilde{P}_x$ is given in (2.4.17) as

$$\begin{aligned} \tilde{P}_x \quad &\propto \quad \prod_{k=2}^{n} \Big\{ \pi(\tilde{y}(t_k)|\tilde{s}(t_k), h(t_k), \tilde{x}(t_k)) \\ &\times \quad \pi\big(\tilde{x}(t_k)|\tilde{x}(t_{k-1}), \boldsymbol{\theta}\big)\pi\big(\tilde{x}(t_{k+1})|\tilde{x}(t_k), \boldsymbol{\theta}\big) \Big\}. \end{aligned} \tag{3.5.12}$$

As noted earlier the data augmentation scheme adopted here is essentially same as that described in Section 2.4.5 except that here we allow the household sizes to vary over time. In what follows, we shall discuss spatial epidemic model within households and also develop Bayesian inference approach for the estimation of the model parameters.

3.6 Spatial SIS Epidemic Model

In this section, we give a brief overview on some recent developments on spatial epidemic modelling and outline Bayesian inference framework for the parameters of the model.

3.6.1 Overview

Spatial epidemic models play a major role in accounting for spatial dependence in spatial epidemic data. In general, a spatial dataset contains information on the individual characteristics of interest as well as its location in space. A spatial dataset may be point-level or geostatistical, areal unit or lattice data, or point process data, see, Cressie (1993). Accounting for spatial dependence improves the estimates of the variations in estimates and makes inference and prediction more powerful (Haran, 2011). A popular approach to modelling spatial dependence is via Gaussian random fields (GRF) models which include Gaussian Processes (GP) and Gaussian Markov random fields (GMRF), see, for example, Haran (2011). Typically, the spatial dependence is modelled via GP using distance dependent parametric covariance function, $\Sigma(\Phi)$ say, when data are point-level. Spatial proximity is then measured in terms of the distance between two locations, s_i and s_j, say. For areal data where data are regionally aggregated, spatial dependence is modelled via GMRF using parameterised precision matrix (or inverse covariance matrix). A distance measure that can be employed for an areal level data is intercentroid distance between regions, but this may not be appropriate given that it is highly unlikely that all the regions considered would be regular. The use of Gaussian Markov random fields for areal data enables dependence to be specified in terms of adjacencies and neighborhoods, thereby giving rise to computationally efficient sparse covariance matrix. This is a major advantage of GMRF models. However, in this section we shall focus on Gaussian processes (GP) in that the data we consider here are point-level.

There are quite a few studies of spatial disease models which mainly assume that the correlation between the realizations of a latent process is the function of their separation distance, see, for example, Diggle et al. (1998), Keeling et al. (2001), Savill et al. (2006), Kypraios (2007), Jewell et al. (2009) and Deardon et al. (2010). Diggle et al. (1998) used a stationary Gaussian process with continuous index set, $D \subset \mathbb{R}$, to model spatial variations on the incidence of Campylobacter infections in north Lancashire and south Cumbria. The majority of the studies mentioned above focus mainly on farm to farm or individual to individual infection spread on the 2001 UK FMD using various distant dependent transmission kernels. For example, Keeling et al. (2001) used exponential-type transmission kernel. Jewell et al. (2009) used exponential-type kernel for High Pathogenic Avian Influenza H5N1 (HPAI) data. A Geometric change-point kernel employed by Deardon et al. (2010) satisfies

$$K(d_{i,j}, \Phi) = \begin{cases} k_0 & 0 < d_{ij} < \delta_0 \\ d_{ij}^b & \delta_0 < d_{ij} < \delta_{max} \\ 0 & otherwise, \end{cases} \tag{3.6.1}$$

where k_0, δ_0 and b are parameters with the maximum distance allowed, δ_{max}, equal to 30km. Also, Diggle et al. (1998) used a powered exponential kernel of type

$$\rho(u) = \exp\{-(\alpha u)^{\delta}\}\,, \tag{3.6.2}$$

where the parameter $\alpha > 0$, $\delta > 0$) are the parameters of the covariance function and u measures the distance between regions.

All the approaches mentioned above are for SIR (susceptible $\rightarrow$ infected $\rightarrow$ recovered) epidemics. We shall adapt an approach similar to that explored in Diggle et al. (1998) motivated by a rich data set on the spread of tick-borne diseases among Tanzania cattle. Modelling at an individual (cow) level gives rise to a household (farm) level.

3.7 Model Setup

In this section, we outline the model setup with all the relevant assumptions made. First we note that the model setup here is similar to that given in Section 3.4 except that here we assume that the background risk of infection λ is a function of a realization $\mathbf{A}$ of a Gaussian random field, $\mathcal{G}$. A Gaussian random field $\mathcal{G}$ (or GRF) is a random field or a stochastic process in an Euclidean space whose finite dimensional distribution has a multivariate Gaussian distribution completely specified by expectations and covariances. Let $s_1, s_2, \ldots, s_N$ denote the spatial locations of N households (or farms). Let $A(s_i) = A_i$ denote the realization of the GRF corresponding to household i at location s_i ($i = 1, 2, \ldots, N$), where the process $\{A(s) : s \in D \subset \mathbb{R}^d\}$ is a stationary Gaussian process with mean zero and covariance matrix $\Sigma(\Phi)$. Note that the index set D is fixed and continuous. Throughout, We have that

$$\mathbf{A}|\Phi \sim \mathcal{MVN}(\mathbf{0}, \Sigma(\Phi)), \tag{3.7.1}$$

where $\mathbf{A} = \big(A_1, A-2, \ldots, A_N\big)^T$, $\Phi = \{\kappa, \phi\}$ are the parameters of the covariance function, $\mathbf{0}$ is a vector of zeros of length N and $\Sigma(\Phi)$ is an $N \times N$ symmetric covariance matrix. It is well known in the literature that $\Sigma(\Phi)$ needs to be positive definite to avoid distributional impropriety, see, for example, Cressie (1993) and Haran (2011). Consequently, the covariance function $\Sigma(\Phi)$ is specified with a positive definite parametric covariance function. We adapt an exponential covariance function similar to that in Haran (2011) and is given by

$$(\Sigma(\Phi))_{ij} = \begin{cases} \kappa \exp(-\frac{d(i,j)}{\phi}) & \text{if } d(i,j) > 0, \\ \kappa + \psi & \text{if } d(i,j) = 0, \end{cases} \tag{3.7.2}$$

where $d(i,j)(= \|s_i - s_j\|)$ is the Euclidean distance between locations s_i and s_j, κ is

a scaling parameter, ϕ is the range of spatial dependence and ψ is the 'nugget' effect which measures the variance of non-spatial error. We choose Euclidean distance as the distance metric here in that there is no strong evidence against it, for example, there are no natural barriers such as lakes (see, Figure 3.10.1, for the data we consider here). Also, recent studies have shown that Euclidean distance works best when no major geographical barriers exist, see, for example, Savill et al. (2006). Note that (3.7.2) is a special case of the larger Matérn family which satisfies

$$Cov(d(i,j);\psi,\kappa,\nu) = \begin{cases} \frac{\kappa}{2^{\nu-1}\Gamma(\nu)}(2\nu^{1/2}d(i,j)/\phi)^{\nu}K_{\nu}(2\nu^{1/2}d(i,j)/\phi) & \text{if } d(i,j) > 0, \\ \kappa + \psi & \text{if } d(i,j) = 0, \end{cases} \tag{3.7.3}$$

where $K_{\nu}(d(i,j))$ is a modified Bessel function of order ν and where ν is a smoothness parameter and the smoothness of the process increases with ν. One advantage of the Matérn covariance kernel is that it allows for the estimation of the smoothness of the process, but this can be problematic for spatial realizations emanating from processes that are unlikely to be smooth. Studies have suggested the use of exponential covariance functions for spatial data, see, for example, Haran (2011).

Now given the realizations from the zero-mean stationary Gaussian process $\mathbf{A}(\mathbf{s})$, we define

$$\boldsymbol{\lambda}(\mathbf{s}) = \exp\{\ln|\mu| + \mathbf{A}(\mathbf{s})\}, \tag{3.7.4}$$

where $\boldsymbol{\lambda}(\mathbf{s}) = (\lambda(s_1), \lambda(s_2), \ldots, \lambda(s_N))$ are the spatially varying background risks of infection for locations $\mathbf{s} = (s_1, s_2, \ldots, s_N)$, and for some parameter μ.

Therefore, based upon the assumption that the data $\mathbf{y}$ are only partially observed and depends upon the augmented data $\mathbf{x}$, the complete data $\mathbf{x}$ depends upon $\mathbf{A}(\mathbf{s})$ through $\boldsymbol{\lambda}(\mathbf{s})$. Figure 3.7.1 shows the schematic representation of the dependence of our model.

Observe that we have assumed that local rate of infection β and the recovery rate γ are not spatially varying, whereas background risk intensity varies.

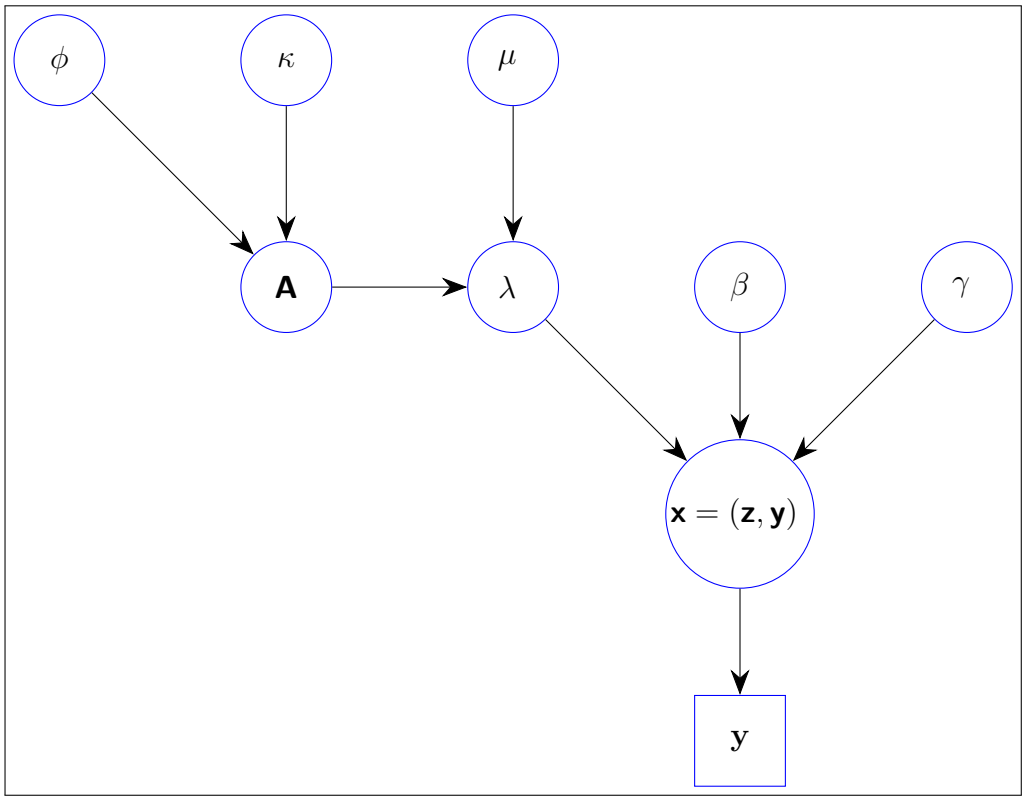

Figure 3.7.1: Schematic representation of the spatial model with incomplete data.

In Section 3.8, we develop Bayesian inference approach for efficient estimation of the key model parameters $\Phi = \{\kappa, \phi\}$, $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \beta, \gamma\}$ since $\lambda(\mathbf{s})$ depends upon μ and $\mathbf{A}(\mathbf{s})$. This approach shall be implemented in MCMC framework.

3.8 Bayesian Inference for Spatial SIS epidemic model

In this section, we utilize the flexibility of MCMC to develop algorithms that sample efficiently from the posterior distributions of interest and perform Bayesian inference on the desired model parameters.

Besag and Green (1993) gives a good review of the applicability and flexibility of MCMC

in spatial statistics. MCMC has been widely explored in disease mapping studies (example, Utazi et al. (2018)) and in epidemic modelling (for example, Gibson (1997), Diggle et al. (1998) and Jewell et al. (2009)). This popularity of MCMC in spatial statistics is due to its ability to efficiently sample from very complex models.

3.8.1 MCMC

As already mentioned above, let $\boldsymbol{\theta} = \{\mu, \beta, \gamma\}$ and $\Phi = \{\kappa, \phi\}$. Recall that we have a point-level spatial epidemic data $\mathbf{y}$ which are only partially observed and that the imputation of an additional information $\mathbf{z}$ makes the intractable likelihood function $\pi(\mathbf{y}|\boldsymbol{\theta})$ to become tractable through $\pi(\mathbf{x} = (\mathbf{z}, \mathbf{y})|\boldsymbol{\theta})$. Therefore, we need to construct MCMC algorithms to generate samples from the posterior distribution $\pi(\boldsymbol{\theta}, \Phi, \mathbf{A}(\mathbf{s}), \mathbf{s}, \mathbf{z}|\mathbf{y})$ by sampling iteratively from

1. $\pi(\Phi|\boldsymbol{\theta}, \mathbf{A}(\mathbf{s}), \mathbf{x} = (\mathbf{z}, \mathbf{y}))$,

2. $\pi(A_i|\mathbf{A}_{-i}, \boldsymbol{\theta}, \Phi, \mathbf{x} = (\mathbf{z}, \mathbf{y}))$, where $\mathbf{A}_{-i}$ denotes the vector $\mathbf{A}$ without its i^{th} element, $A_i = A(s_i)$,

3. $\pi(\boldsymbol{\theta}|\Phi, \mathbf{A}(\mathbf{s}), \mathbf{x} = (\mathbf{z}, \mathbf{y}))$ and

4. $\pi(\mathbf{z}|\boldsymbol{\theta}, \Phi, \mathbf{A}(\mathbf{s}), \mathbf{y})$.

Throughout we assume independent prior distributions.

We shall now exploit the conditional independence and mutual independence that exist in the model as shown in Figure 3.7.1. Observe that setting $\mathbf{A}$ and μ equal to zero reduces the model to the non-spatial model described in Figure 3.5.1. We now give the MCMC updating schemes utilized in this section and proceed as follows.

STEP 0: Initialize the values of $\boldsymbol{\theta} = \{\mu, \beta, \gamma\}$, $\Phi = \{\kappa, \phi\}$, $\mathbf{A}$ and $\mathbf{x}$. The starting values of $\boldsymbol{\theta}$, Φ and $\mathbf{A}$ may be arbitrarily, albeit with sensible values, chosen in the range specified by the prior distributions, $\pi(\boldsymbol{\theta})$ and $\pi(\Phi)$. The initial values of the augmented data $\mathbf{x}$ are chosen to be consistent with the observed data $\mathbf{y}$ using any of the methods described in the ***Step 0*** of the Independence Sampler algorithm in Section 3.5.3 above. Set $\boldsymbol{\theta}^{(0)} = \{\mu^{(0)}, \beta^{(0)}, \gamma^{(0)}\}$ and $\Phi^{(0)} = \{\kappa^{(0)}, \phi^{(0)}\}$ and obtain $\Sigma^{(0)}$, $\mathbf{A}^{(0)}(\mathbf{s})$ and $\boldsymbol{\lambda}^{(0)}(\mathbf{s})$ using (3.7.1), (3.7.2) and (3.7.4) respectively.

STEP 1: Updating $\pi(\Phi|\boldsymbol{\theta}, \mathbf{A}(\mathbf{s}), \mathbf{x} = (\mathbf{z}, \mathbf{y}), \mathbf{y})$ is essentially updating $\pi(\Phi|\mathbf{A})$ since Φ is independent of $\boldsymbol{\theta}$, $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{A}$. Therefore, we have

$$\pi(\Phi|\mathbf{A}) \propto \pi(\mathbf{A}|\Phi)\pi(\Phi), \tag{3.8.1}$$

where the likelihood function $\pi(\mathbf{A}|\Phi)$ is a zero mean multivariate Gaussian distribution with covariance matrix $\Sigma(\Phi)$, and $\pi(\Phi)$ is the joint prior distribution on the covariance function parameters $\Phi = \{\kappa, \phi\}$. Then, it follows that

$$\begin{aligned} \pi(\Phi|\mathbf{A}) &= (2\pi)^{-\frac{N}{2}} \det(\Sigma(\Phi))^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2} A^T \Sigma(\Phi)^{-1} \mathbf{A} \Big\} \\ &\times \pi(\Phi), \\ &\propto \det(\Sigma(\Phi))^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2} A^T \Sigma(\Phi)^{-1} \mathbf{A} \Big\} \\ &\times \pi(\Phi), \end{aligned} \tag{3.8.2}$$

where in principle, the prior distribution $\pi(\Phi)$ could be chosen to be any sensible distribution with respect to the range of the parameter values. Here, we assign Gamma distributed prior distribution to both Φ. In other words, $\pi(\Phi) \sim Gamma(A_\Phi, B_\Phi)$, where A_Φ and B_Φ are hyperparameters.

We then update $\pi(\Phi|\mathbf{A})$ using Random walk Metropolis (RWM) algorithms as follows. For $r \geq 0$, do the following:

(a) Propose $\Phi' = \{\kappa', \phi'\}$ from a bivariate Gaussian proposal distribution centered at the current value of Φ with a proposal covariance matrix Σ_Φ, *i.e.*, $\Phi' \sim \mathcal{N}_2(\Phi, \Sigma_{\Phi'})$.

(b) Accept $\Phi' = \{\kappa', \phi'\}$ with probability

$$\begin{aligned}\Delta(\Phi, \Phi') &= \min\left\{1, \frac{\pi(\Phi'|\mathbf{A})}{\pi(\Phi|\mathbf{A})}\right\}, \\ &= \min\left\{1, \frac{\det(\Sigma(\Phi'))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}A^T\Sigma(\Phi')^{-1}\mathbf{A}\right\}\pi(\Phi')}{\det(\Sigma(\Phi))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}A^T\Sigma(\Phi)^{-1}\mathbf{A}\right\}\pi(\Phi)}\right\}.\end{aligned} \tag{3.8.3}$$

(c) Draw u from $U[0, 1]$,

(d) If $u \leq \Delta(\Phi, \Phi')$,

(i) accept Φ',

(ii) set $\Phi^{(r+1)} = \Phi'$.

(e) If $u > \Delta(\Phi, \Phi')$

(i) do not accept Φ',

(ii) set $\Phi^{(r+1)} = \Phi^{(r)}$.

(f) If only samples from $\pi(\Phi|\mathbf{A})$ are desired, repeat steps (a) to (e) until samples of the desired size are obtained, otherwise proceed to ***STEP 2:***.

STEP 2: Updating $\pi(A_i|\mathbf{A}_{-i}, \boldsymbol{\theta}, \Phi, \mathbf{x}, \mathbf{y})$ essentially means updating $\pi(A_i|\mathbf{A}_{-i}, \boldsymbol{\theta}, \Phi, \mathbf{x})$ since $\mathbf{A}$ is independent of $\mathbf{y}$ given $\mathbf{x}$. Then we have

$$\pi(A_i|\mathbf{A}_{-i}, \boldsymbol{\theta}, \Phi, \mathbf{x}) \propto \pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{A})\pi(A_i|\mathbf{A}_{-i}, \Phi), \tag{3.8.4}$$

since $\mathbf{x}$ is conditionally independent of Φ given $\mathbf{A}$ and A_i is independent of $\boldsymbol{\theta}$. Then by mutually independent households assumption, we have

$$\pi(A_i|\mathbf{A}_{-i},\boldsymbol{\theta},\Phi,\mathbf{x}) \propto \pi(\mathbf{x}_i|\boldsymbol{\theta},A_i)\pi(A_i|\mathbf{A}_{-i},\Phi), \tag{3.8.5}$$

where the second term on the right hand side, $\pi(A_i|\mathbf{A}_{-i},\Phi)$, is a univariate Gaussian distribution which follows from the multivariate Gaussian distribution of $\pi(\mathbf{A}|\Phi)$. We derive the mean and variance of $\pi(A_i|\mathbf{A}_{-i},\Phi)$ as follows. Let $\Lambda = \Sigma(\Phi)^{-1}$, then we have

$$\begin{aligned}
\pi(\mathbf{A}|\Phi) &\propto \big|\mathbf{\Lambda}\big|^{\frac{1}{2}}\exp\big(-\frac{1}{2}\mathbf{A}^T\mathbf{\Lambda}\mathbf{A}\big),\\
&\propto \exp\Big\{-\frac{1}{2}\big(A_1^2\Lambda_{11}+2A_1\sum_{j\neq 1}^{N}A_j\Lambda_{ij}\big)\Big\},\\
&\propto \exp\Big\{-\frac{\Lambda_{11}}{2}\big(A_1^2+\frac{2A_1\sum_{j\neq 1}^{N}A_j\Lambda_{ij}}{\Lambda_{11}}\big)\Big\},\\
&\propto \exp\Big\{-\frac{\Lambda_{11}}{2}\Big(A_1^2+2A_1\sum_{j\neq 1}^{N}A_j\Lambda_{ij}+\Big(\sum_{j\neq 1}^{N}A_j\Lambda_{ij}\Big)^2\Big)\Big\},\\
&\propto \exp\Big\{-\frac{\Lambda_{11}}{2}\Big(A_1+\frac{\sum_{j\neq 1}^{N}A_j\Lambda_{ij}}{\Lambda_{11}}\Big)^2\Big\},\\
\Rightarrow A_i|\mathbf{A},\mathbf{\Phi} &\sim \mathcal{N}\Big(-\frac{\sum_{j\neq 1}^{N}A_j\Lambda_{ij}}{\Lambda_{ii}},\Lambda_{ii}^{-1}\Big).
\end{aligned} \tag{3.8.6}$$

Then, (3.8.5) can also be expressed as

$$\begin{aligned}
\pi(A_i|\mathbf{A}_{-i},\mathbf{x},\Phi,\boldsymbol{\theta}) &\propto \pi(\mathbf{x}_i|\boldsymbol{\theta},A_i)\pi(A_i|\mathbf{A}_{-i},\Phi),\\
&\propto \Big\{\prod_{k=2}^{n_i}\pi\big(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta},A_i\big)\Big\},\\
&\times \exp\Big\{-\frac{\Lambda_{ii}}{2}\Big(A_i+\frac{\sum_{j\neq 1}^{N}A_j\Lambda_{ij}}{\Lambda_{ii}}\Big)^2\Big\},
\end{aligned} \tag{3.8.7}$$

where the first term on the right hand side (second line) is the likelihood function for household i ($i = 1, 2, \ldots, N$). We shall now update $\pi(A_i|\mathbf{A}_{-i},\mathbf{x},\Phi,\boldsymbol{\theta})$. Since the second term of the second line of right hand side of (3.8.7) is a univariate Gaussian distribution, it is straightforward to simulate $A_i s$ using Independence Sampler steps. On the other

hand, the likelihood function, $\pi(\mathbf{x}_i|\boldsymbol{\theta}, A_i)$, is a complicated function of the transition probability matrix (Q-matrix) with no closed form. Therefore, we use Independence sampling steps and obtain samples from $\pi(A_i|\mathbf{A}_{-i}, \Phi)$ whilst using Metropolis-Hastings updates.

Independence Sampling steps

(a) Choose a new value A_i' from the univariate Gaussian distribution $A_i|\mathbf{A}_{-i}, \Phi \sim \mathcal{N}(U/V, 1/V)$, where $U = -\sum_{j\neq 1}^{N} A_j \Lambda_{ij}$ and $V = \Lambda_{ii}$, see, (3.8.6).

(b) Accept A_i' with probability

$$\begin{aligned} \Delta(A_i, A_i') &= \min\left\{1, \frac{\pi(\mathbf{x}_i|\boldsymbol{\theta}, A_i')/\pi(A_i'|\mathbf{A}_{-i}, \Phi)}{\pi(\mathbf{x}_i|\boldsymbol{\theta}, A_i)/\pi(A_i|\mathbf{A}_{-i}, \Phi)}\right\}, \\ &= \min\left\{1, \frac{\pi(\mathbf{x}_i|\boldsymbol{\theta}, A_i')}{\pi(\mathbf{x}_i|\boldsymbol{\theta}, A_i)}\right\}. \end{aligned} \tag{3.8.8}$$

(c) Draw u from $U[0, 1]$.

(d) If $u \leq \Delta(A_i, A_i')$,

(i) accept A_i',

(ii) set $\mathbf{A}' = (A_1, A_2, \ldots, A_{i-1}, A_i', A_{i+1}, \ldots, A_N)^T$,

(iii) set $\mathbf{A}^{(i+1)} = \mathbf{A}'$.

(e) If $u > \Delta(A_i, A_i')$,

(i) do not accept A_i',

(ii) set $\mathbf{A}' = (A_1, A_2, \ldots, A_{i-1}, A_i, A_{i+1}, \ldots, A_N)^T$,

(iii) set $\mathbf{A}^{(i+1)} = \mathbf{A}'$.

(f) Repeat steps (a) to (e) for all $i = 1, 2, \ldots, N$.

(g) If only samples from $\pi(A_i|\mathbf{A}_{-i}, \mathbf{x}, \Phi, \boldsymbol{\theta})$ are desired, repeat steps (a) to (f) until samples of the desired size are obtained, otherwise proceed to ***STEP 3***.

STEP 3: To update $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}, \mathbf{A}, \Phi)$ essentially means updating $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{A})$, since $\boldsymbol{\theta}$ is independent of Φ given $\mathbf{A}$, and conditionally independent of $\mathbf{y}$ given $\mathbf{x}$. Then, we have

$$\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{A}) \propto \pi(\mathbf{x}|\mathbf{A}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{3.8.9}$$

since $\mathbf{A}$ is independent of $\boldsymbol{\theta}$. Then by independent household assumption, we have

$$\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{A}) \propto \left\{ \prod_{i=1}^{N} \pi(\mathbf{x}_i|A_i, \boldsymbol{\theta}) \right\} \pi(\boldsymbol{\theta}), \tag{3.8.10}$$

where $\pi(\boldsymbol{\theta})$ is the joint prior distribution on the parameters $\boldsymbol{\theta} = (\mu, \beta, \gamma)$ and A_i is the Gaussian random fields realization corresponding to household i. Finally, like before, we assume independent prior distributions so that Equation (3.8.10) can be written as

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{A}) \quad &\propto \quad \prod_{i=1}^{N} \left\{ \prod_{k=2}^{n_i} \pi(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{ik}), A_i, \boldsymbol{\theta}) \right\} \\ &\times \quad \pi(\mu)\pi(\beta)\pi(\gamma), \end{aligned} \tag{3.8.11}$$

for N households each observed n_i times, where in principle any sensible prior distributions with respect to the valid range of the parameter values could be assigned to $\pi(\mu)$, $\pi(\beta)$ and $\pi(\gamma)$. In particular, we assign Gamma distributed priors to the parameters μ, β and γ, so that $\pi(\mu) \sim Gamma(A_\mu, B_\mu)$, $\pi(\beta) \sim Gamma(A_\beta, B_\beta)$ and $\pi(\gamma) \sim Gamma(A_\gamma, B_\gamma)$, where A_μ, B_μ, A_β, B_β, A_γ and B_γ are hyperparameters. It is now straightforward to update $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{A})$ via the Random walk Metropolis (RWM) algorithms with multivariate Gaussian proposal density given in Algorithm 4. Proceed to ***STEP 4***.

STEP 4: Update $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}, \Phi, \mathbf{A})$, where $\mathbf{z}$ is the additional imputed information which together with the partially observed data $\mathbf{y}$ gives the complete or the augmented data

$\mathbf{x}$. Here, we essentially update $\pi(\mathbf{x} = (\mathbf{z}, \mathbf{y})|\mathbf{y}, \boldsymbol{\theta}, \Phi, \mathbf{A}) = \pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}, \mathbf{A})$ due to conditional independence of $\mathbf{x}$ on Φ. Therefore, we have

$$\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}, \mathbf{A}) \propto \pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{A}), \tag{3.8.12}$$

since $\mathbf{y}$ is conditionally independent of $\boldsymbol{\theta}$ and $\mathbf{A}$ given $\mathbf{x}$. Then by mutually independent households assumption, the probability of observing the augmented data given $\boldsymbol{\theta}$, $\mathbf{y}$ and $\mathbf{A}$ satisfies

$$\begin{aligned} \pi(\mathbf{x} = (\mathbf{z}, \mathbf{y})|\boldsymbol{\theta}, \mathbf{y}, \mathbf{A}) \quad &\propto \quad \prod_{i=1}^{N} \prod_{k=2}^{n_i} \Big\{ \pi\big(\mathbf{y}_i(t_{ik})|\mathbf{x}_i(t_{ik})\big) \\ &\times \quad \pi\big(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{ik-1}), A_i, \boldsymbol{\theta}\big)\pi\big(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), A_i, \boldsymbol{\theta}\big) \Big\}. \end{aligned} \tag{3.8.13}$$

Now using the Independence Sampler steps outlined in Section 3.5.3 and summarized in Algorithm 8, update

$$\begin{aligned} \pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}, \Phi, \mathbf{A}) \quad &\propto \quad \prod_{i=1}^{N} \prod_{k=2}^{n_i} \Big\{ \pi(\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}), A_i, \boldsymbol{\theta}) \\ &\times \quad \pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}), x_{ij}(t_{ik}) = z_{ij}(t_{ik}), A_i, \boldsymbol{\theta}) \Big\}, \end{aligned} \tag{3.8.14}$$

and store the sampled $\mathbf{z}$ values if desired. Set $\mathbf{x} = (\mathbf{z}', \mathbf{y})$, where $\mathbf{z}'$ are the updated auxiliary data $\mathbf{z}$.

Summary:

From the foregoing, we see that one complete MCMC cycle is to go over the updating steps listed above from ***STEP 1*** to ***STEP 4*** after choosing the initial values for the parameters $(\boldsymbol{\theta}, \Phi)$, the Gaussian random fields realizations $(\mathbf{A})$ and the augmented data $(\mathbf{x} = (\mathbf{z}, \mathbf{y}))$. We now summarize the MCMC updating schemes and steps described above in Algorithm 9.

Algorithm 9 MCMC algorithms for Spatial epidemic models

1. Intialize Φ, $\boldsymbol{\theta}$, $\mathbf{A}$, $\mathbf{x} = (\mathbf{z}, \mathbf{y})$.

2. Update $\Phi|\mathbf{A}$ using RWM, then

   (a) propose Φ' from a multivariate Gaussian proposal distribution,

   (b) accept Φ' with probability $\Delta(\Phi, \Phi') = \min\{1, \pi(\Phi'|\mathbf{A})/\pi(\Phi|\mathbf{A})\}$.

3. Update $\pi(A_i|\mathbf{A}_{-i}, \boldsymbol{\theta}, \Phi, \mathbf{x})$ using

   (a) Independence Sampler,

   (b) draw A_i' from $A_i|\mathbf{A}_{-i}, \Phi \sim \mathcal{N}(U/V, 1/V)$, where $U = -\sum_{j\neq 1}^{N} A_j \Lambda_{ij}$ and $V = \Lambda_{ii}$, see, (3.8.6).

   (c) Accept A_i' with probability $\min\{1, \pi(\mathbf{x}_i|\boldsymbol{\theta}, A_i')/\pi(\mathbf{x}_i|\boldsymbol{\theta}, A_i)\}$.

   (d) Repeat steps (a) to (c) for all $i = 1, 2, \ldots, N$.

4. Update $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{A})$ using RWM,

   (a) Propose $\boldsymbol{\theta}' = (\mu', \beta', \gamma')$ from $\boldsymbol{\theta}' \sim \mathcal{MVN}(\boldsymbol{\theta}, \Sigma_\theta')$.

   (b) Accept the proposed value with probability $\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\{1, \pi(\boldsymbol{\theta}'|\mathbf{x}, \mathbf{A})/\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{A})\}$.

5. Update $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}, \mathbf{A}) \equiv \pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}, \mathbf{A})$ using the Independence Sampler steps outlined in Section and summarized in Algorithm 8.

6. Repeat steps (2) to (5) until samples of the desired size are obtained.

We incorporate steps to ensure that the MCMC algorithms are efficiently implemented. For all the Random walk Metropolis updates, we use the adaptive RWM strategies outlined in Section 2.4.2, which includes a pilot run and then using the posterior variance from the pilot run (which yields an acceptance rate close to the well known optimal acceptance rate of 23.4%) as the proposal variance for the main MCMC runs. This is called optimal shaping as it helps the algorithm to quickly learn the shape of the posterior distribution. Another MCMC efficiency improvement strategy is optimal scaling which scales the proposal variance by the scalar c, where an optimal value of $2.38^2/d$ is suggested in the literature for c, see, for example, Roberts and Rosenthal (2001), for a d-dimensional set of parameters.

In what follows, we shall illustrate the implementation of our MCMC algorithms using simulated data set and later applied to a real life data.

3.9 Simulated Data Example

In this section, we use simulated data sets to demonstrate the applicability of the MCMC algorithms developed in this chapter for both the non-spatial SIS open population model and the spatial SIS epidemic model. First, we outline the methodology employed in the simulation and later discuss the implementation of the MCMC algorithms to the simulated data sets as well as the results obtained.

Method:

We shall first describe the methodology employed in this example.

In both cases, we used same household structure and same time data. Also we used the same local rate of infection, $\beta = 0.40$, and same recovery rate, $\gamma = 0.55$, throughout as these are independent of the spatial locations of the households. Furthermore, we

simulate an SIS epidemic individual-based data (IBD) for $N = 100$ households with household sizes ranging from 1 to 5 with majority of the households sized 3 and above. The time data was such that the time difference range is $1 - 2$ weeks, with minimum (maximum) number of observations equal to 2 (8) visits. For the non-spatial model in which the background risk of infection is assumed to be independent of the spatial locations of the households, we set $\lambda = 0.65$ infections per week on average.

For the spatial SIS epidemic model, we first simulate the coordinates corresponding to $s_1, \ldots, s_{100}$ spatial locations of the households from the position (X, Y), such that (X,Y) is bivariate Gaussian distributed with mean $\boldsymbol{\mu}_{x,y}$ and covariance matrix $\Sigma_{x,y}$, where $\boldsymbol{\mu}_{x,y} = (0, 0)^T$ and $\Sigma_{x,y}$ is chosen to be

$$\begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

Figure 3.9.1 shows the distribution of the household locations. We set $\kappa = \mu = 1$ and induce spatial dependence by setting $\phi = 10$. Recall that ϕ is the range of spatial dependence, therefore spatial dependence decays as ϕ gets smaller.

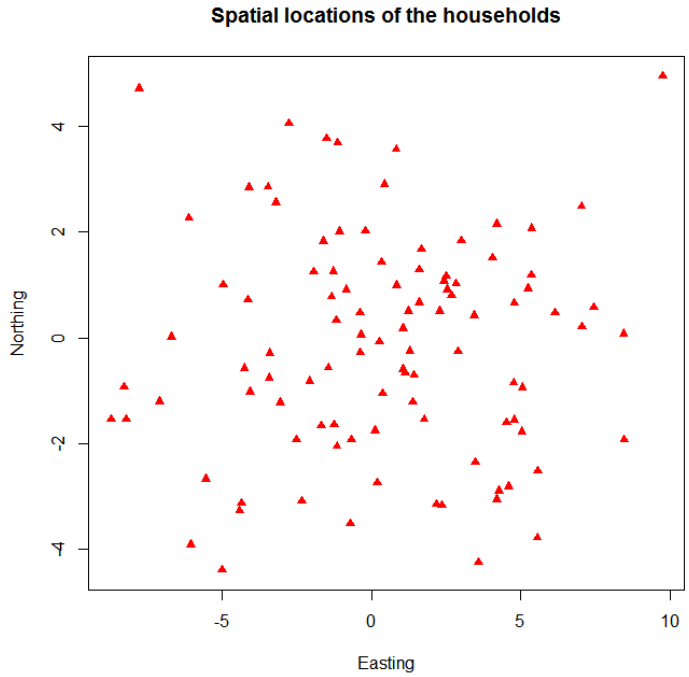

Figure 3.9.1: Spatial distribution of the $N = 100$ simulated households with each represented by a red shape.

Now we calculate the Euclidean distance matrix, $D = (\|d_{ij}\|)$, where

$$\|d_{ij}\| = \sqrt{(xx_i - xx_j)^2 + (yy_i - yy_j)^2}, \tag{3.9.1}$$

and obtain $\Sigma(\Phi) = (\sigma_{i,j})$, where $\sigma_{i,i} = \sigma_i^2$ and

$$\sigma_{i,j} = \exp(-d(i,j)/10). \tag{3.9.2}$$

Gaussian random fields realizations, $\mathbf{A} = (A_1, A_2, \ldots, A_{100})^T$, for the $N = 100$ households are then simulated from $\mathbf{A} \sim \mathcal{MVN}(\mathbf{0}, \Sigma(\Phi))$. Finally, the background risks of infection $\boldsymbol{\lambda}(\mathbf{s}) = (\lambda(s_1), \ldots, \lambda(s_{100}))$ are obtained from

$$\boldsymbol{\lambda}(\mathbf{s}) = \exp(\mathbf{A}(\mathbf{s})). \tag{3.9.3}$$

MCMC Implementation and Results:

We applied MCMC updating schemes as described in Algorithm 9. We used indepen-

dent gamma distributed priors throughout. In particular, we used $\mathcal{G}(1,1)$ for each of $\lambda, \beta, \gamma, \mu, \kappa$ and ϕ.

For the Random walk Metropolis updates, we used the adaptive RWM strategies for optimality. In particular, we used the covariance matrix obtained from the posterior distribution after 3 consecutive pilot runs of 1×10^3 iterations each, as the proposal variance for the main MCMC run. The main MCMC was run for 2×10^4 iterations after which we ignored the first 4×10^3 iterations representing 20% of the entire MCMC samples, as burn-in. Therefore, for future analysis, we used only used the GwM algorithms for the updates of $\pi(A_i|\mathbf{A}_{-i}, \boldsymbol{\theta}, \Phi, \mathbf{x})$. Throughout, the main MCMC diagnostic tool used is the traceplot of posterior density. Autocorrelation function (ACF) plots were used to examine the amount of correlation between the MCMC samples. Acceptance rates were close to the optimal value of 0.234.

For the non-spatial model, posterior mean (standard deviation) for λ, β and γ are $0.62(0.14), 0.38(0.08)$ and $0.65(0.10)$, respectively. These values are fairly close to the true parameter values of $(0.65, 0.4, 0.55)$ for (λ, β, γ) indicating that the MCMC algorithms performing well. Figure 3.9.2 are the traceplots with the ACF plots from the non-spatial open population model. The traceplots and the ACF plots also show that the MCMC is mixing well.

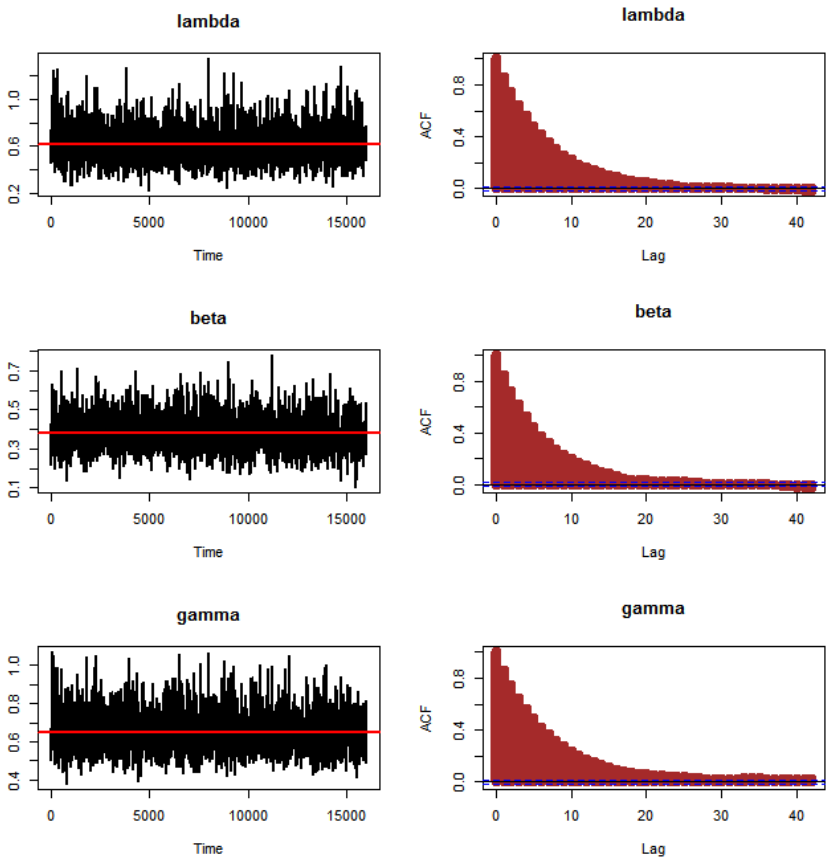

Figure 3.9.2: Traceplots (left) and ACF plots (right) for the non-spatial **simulated** open population model obtained after discarding the first 4×10^3 iterations as burn-in out of 2×10^4 iterations. Each of the red lines represents the mean of the corresponding parameter.

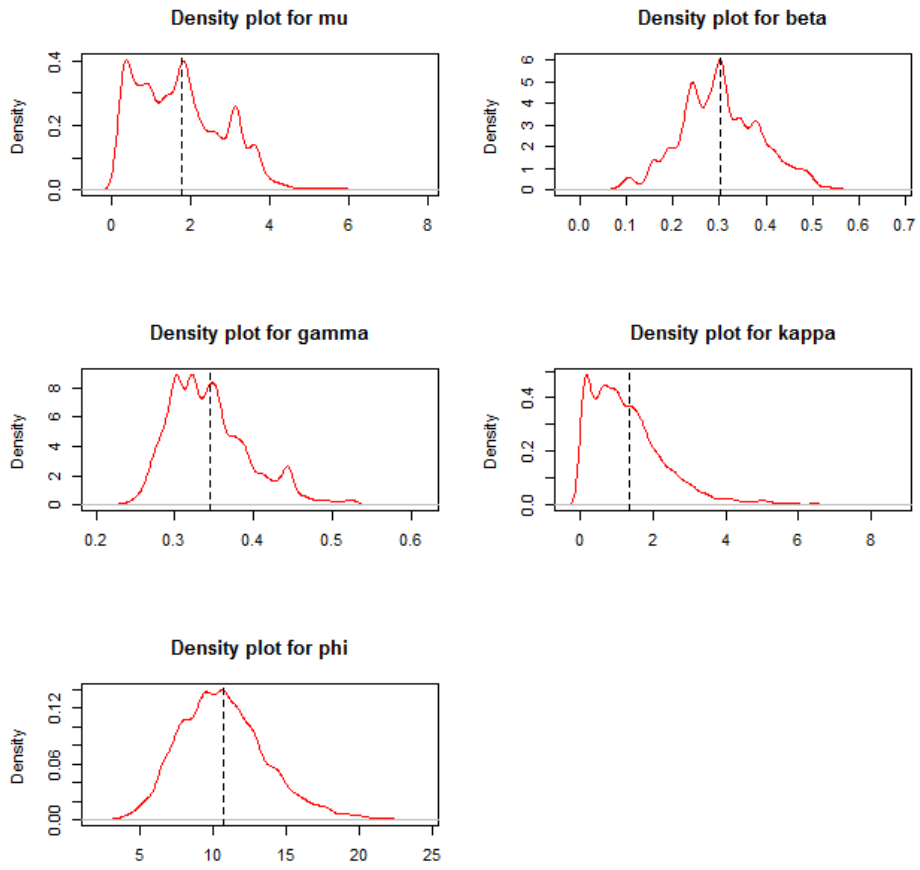

Figure 3.9.3: Posterior density plots of the spatial model with *true* $(\mu, \beta, \gamma, \kappa, \phi) = (1, 0.4, 0.55, 1, 10)$. The vertical lines are the means.

	Spatial model			Non-spatial		
Parameter	Mean	SD	ESS	Mean	SD	ESS
λ	-	-	-	0.615	0.142	1146
β	0.303	0.083	68	0.383	0.082	1277
γ	0.341	0.048	191	0.647	0.098	1129
μ	1.821	0.537	32	-	-	-
κ	1.590	1.104	74	-	-	-
ϕ	10.86	3.080	550	-	-	-

Table 3.9.1: Posterior Means, Standard Deviations (SD) and Effective Sample Sizes (ESS) for the Simulated data example for both spatial and non-spatial open population data obtained from the last 1.6×10^4 samples.

In what follows, we shall now demonstrate the implementation of our algorithms to a real life data set.

3.10 Application to the Tanzania Data

The data on the spread of *Theileria Parva*, a tick-borne disease, among Tanzania cattle population contains information on five strains of the disease namely, *T.parva, T.mutans, A.marginale, B.bigemina, B.bovis*. The data was collected over 11 observation time points with minimum (maximum) of 1(11) visit(s) with majority of the farms visited 3 times.

A total of 380 animals from 156 farms were observed across the 4 regions visited with minimum (maximum) farm size equal to 1(8). The majority of farms contained at most

4 animals. Farm locations were geocoded thus making available the spatial locations of the farms in terms of their longitude and latitude. The data is such that for every given farm, animals exit and join the farm at various point in time. The reasons for the varying population sizes are not clear as there are no information suggesting that in the data. However, possible causes of the varying population sizes over time include death of an animal naturally or disease related or an animal being sold, birth or acquisition of a new animal. The time data are such that the time difference between observations range from $3to11$ weeks. Our purpose here is not to analyze the entire data on the five strains of *Theileria Parva*, rather our aim is to illustrate how our MCMC algorithms could be applied in a real life situation. Figure 3.10.1 (top) displays the map of Tanzania showing the locations of the sampled farms in the four regions visited namely, Tanga, Korogwe, Kibaya and Mtindi. It is easy to see that the farms in Tanga (Figure 3.10.1, bottom) appear to be much closer together than the farms in the other regions. Also, unlike the other regions, there are no major geographical barriers such as lakes, very high mountains, etc, between the farms in Tanga region. This suggests that Euclidean distance is appropriate as the distance metric for the covariance function. Based upon the above reasons, we choose the farms in Tanga region as the case study farms here.

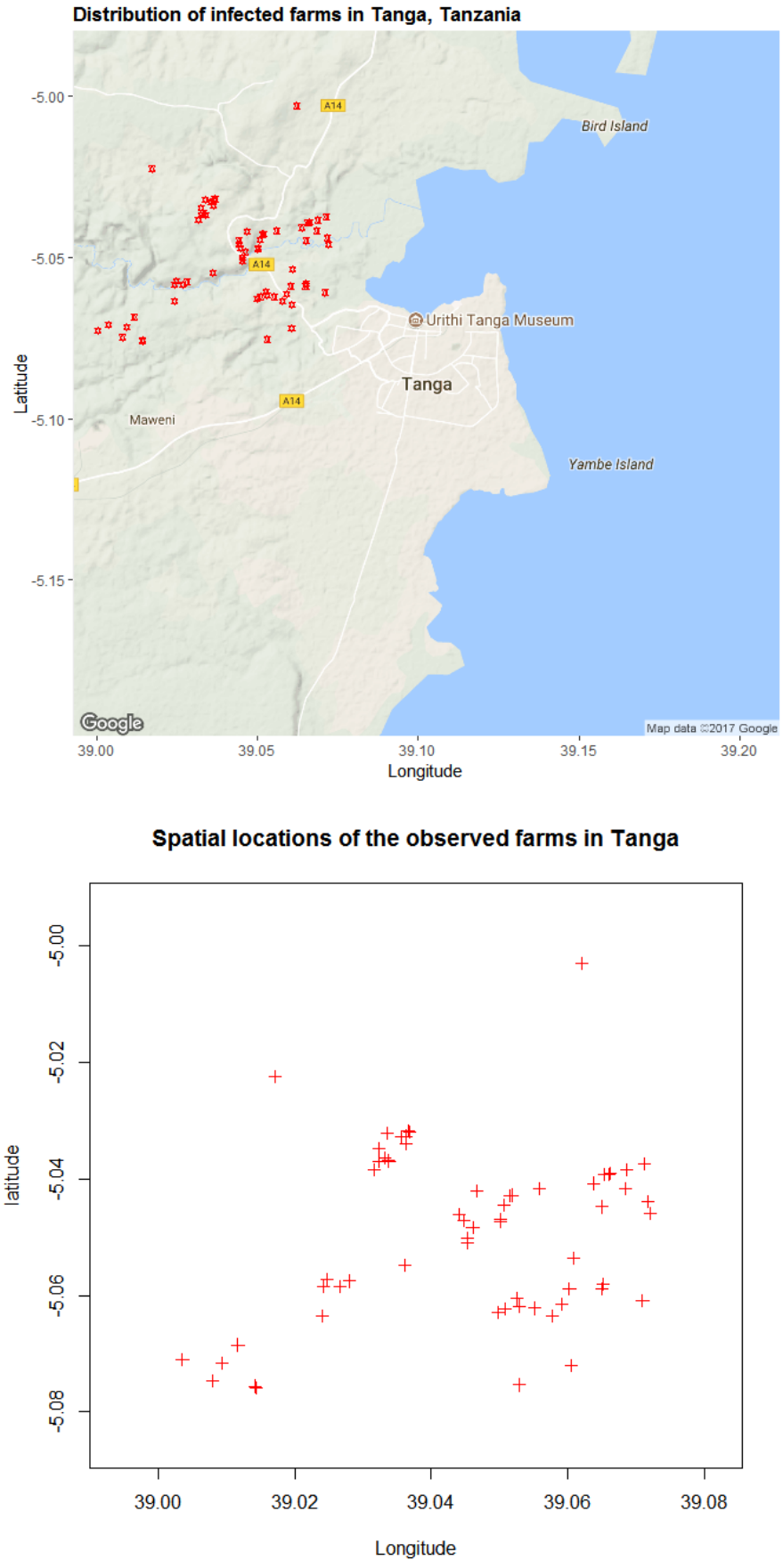

Figure 3.10.1: Map of Tanga (top), a town in Tanzania, and spatial distribution of the 62 observed farms in Tanga (bottom). Each red point represents an observed farm.

There are 64 farms with 174 (45.8% of the entire population) animals in total in all the farms visited in Tanga region with farm sizes range of $1-8$ animal(s) per farm. Majority of the farms in Tanga region contained between 1 and 3 animals. After data cleaning, we dropped data on two farms and proceed with the remaining $N = 62$ farms for the MCMC implementation. The two farms dropped were only visited once each and there is no contribution to the likelihood function by a single time point observation. We focus our attention on T.parva as the disease of primary interest. Next, we applied the data imputation strategy outlined in Section 3.5.1 using the appropriate data codes as specified in (3.5.1). In particular, we used 1 and 0 to denote presence (or true) and absence (or false) of the disease, respectively. Unknown disease statuses were coded 2. The number of imputed time points ranges from 2 to 9 with majority of the imputed time points less than 4. Given that the individual-based data are available, we analyse the data using the IBD framework.

Results:

Throughout, we used independent gamma distributed priors $\mathcal{G}(1,1)$ for each of the parameters $(\lambda, \beta, \gamma, \kappa, \phi)$. Thus we use the same priors as for the simulated data. We adopt the optimal scaling and optimal strategy to optimize our MCMC algorithms. For both the non-spatial and spatial case, the pilot runs informed us of good starting values for the main MCMC runs. Also, the pilot runs of 1×10^3 iterations each from 3 consecutive runs, informed us of the shape of the joint posterior distribution. Therefore, the variance of the posterior distribution from the pilot runs served as the proposal variance for the main MCMC runs. The algorithms were run for 2×10^4 further iterations and burn-in of 4×10^3 was taken. As before, the main diagnoistic tool employed for convergence checks is the traceplot, while the ACF is employed to check for autocorrelation between the sampled values. Figure 3.10.2 shows the traceplots and density plots for the non-spatial Tanzania data where we set both $\mathbf{A}$ and μ equal to zero and assume that λ is indepen-

dent of the households locations in space. The traceplots show evidence of nice mixing MCMC algorithms. The density plots show that marginal posteriors of both λ and γ are approximately symmetric, while the marginal posterior of β is asymmetric. The cause of this behavior in β is not clear at this stage, but the effect of large number of farms having farm sizes of 3 and below can not be completely ruled out, see, for example, Blake et al. (2009).

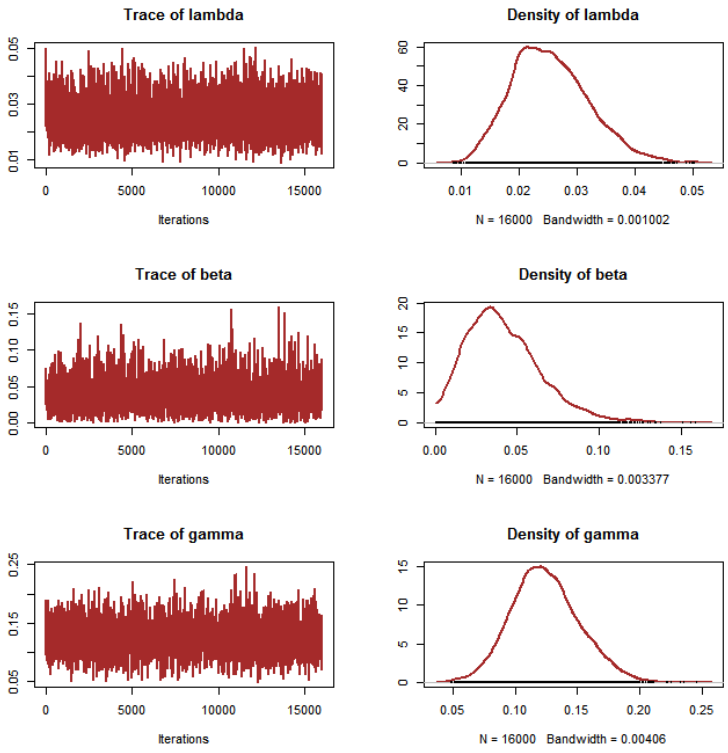

Figure 3.10.2: Traceplots and density plots for the **non-spatial** model parameters of the **Tanzania data application**.

Table 3.10.1 shows the posterior mean, standard deviations and effective sample sizes of the parameters from both spatial and non-spatial models. The values on the last columns of the table for the non-spatial model further support the observations from the trace and density plots. In particular, β has the least effective sample size albeit with moderate standard deviation.

	Spatial model			Non-spatial		
Parameter	Mean	SD	ESS	Mean	SD	ESS
λ	-	-	-	0.025	0.007	2283
β	0.033	0.011	318	0.041	0.022	1310
γ	0.074	0.011	563	0.124	0.027	2018
μ	0.015	0.003	242	-	-	-
κ	0.136	0.122	95	-	-	-
ϕ	1.096	0.318	157	-	-	-

Table 3.10.1: Posterior Means, Standard Deviations (SD) and Effective Sample Sizes (ESS) for the **Tanzania data application** for both spatial and non-spatial open population data based upon the last 1.6×10^4 samples.

We give the interpolation plots of the estimated Gaussian random fields realizations (a) with the corresponding spatially varying global risk of infection, $\boldsymbol{\lambda}$. The plots show spatial variation in the data with high risk areas located near the river (see also, Figure 3.10.1).

Furthermore, Figure 3.10.4 shows paired scatter plots for correlation, density and contours for the non-spatial data. This shows a low negative correlation between the local and global rates of infection. This suggests that high global force of infection does not imply high within farm disease transmission. As noted earlier, we would expect the reverse to be the case if there are more farms with higher number of animals. However, the correlation between the recovery rate γ and the global rate of infection γ is somewhat high suggesting that most recoveries are made when disease is contacted globally.

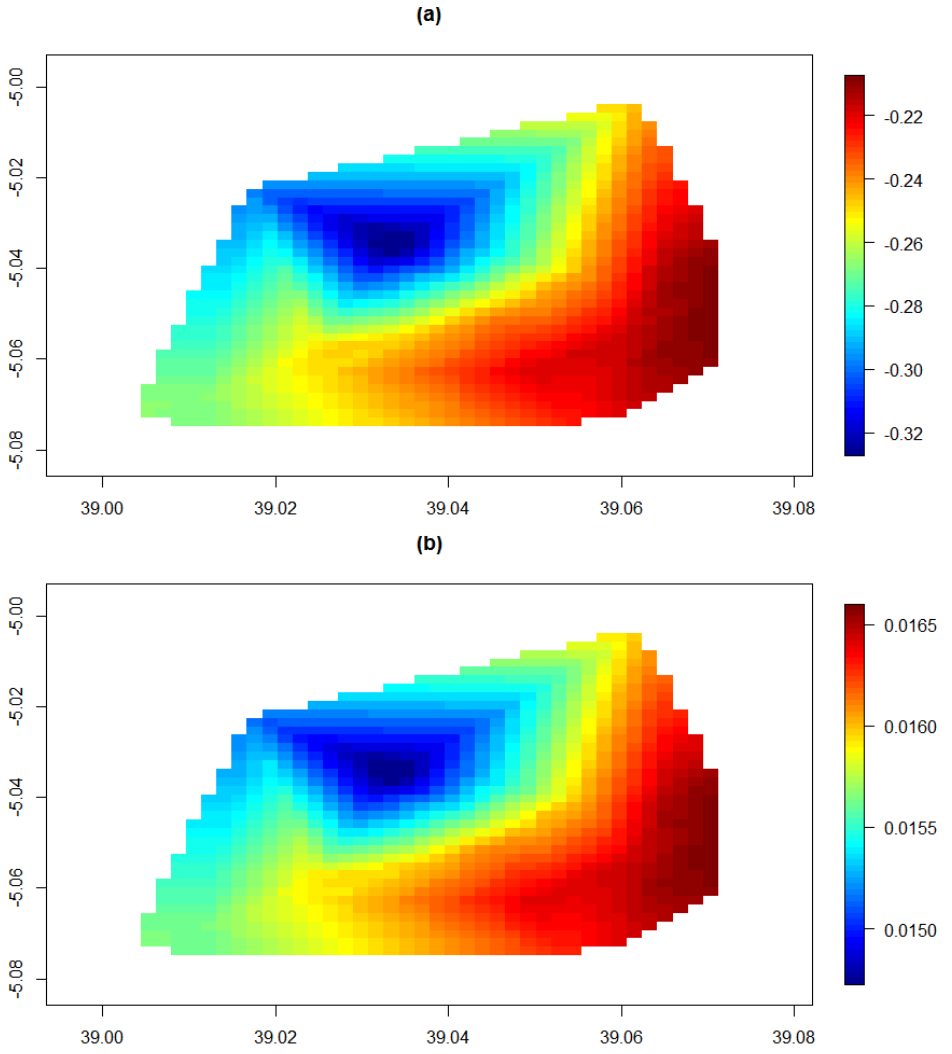

Figure 3.10.3: (a) Predicted Gaussian random fields realizations ($\hat{\mathbf{A}}$) and (b) predicted background risks of infection ($\hat{\boldsymbol{\lambda}} = \hat{\mu}\exp(\hat{\mathbf{A}})$) for the 62 farms observed in Tanga,**Tanzania**.

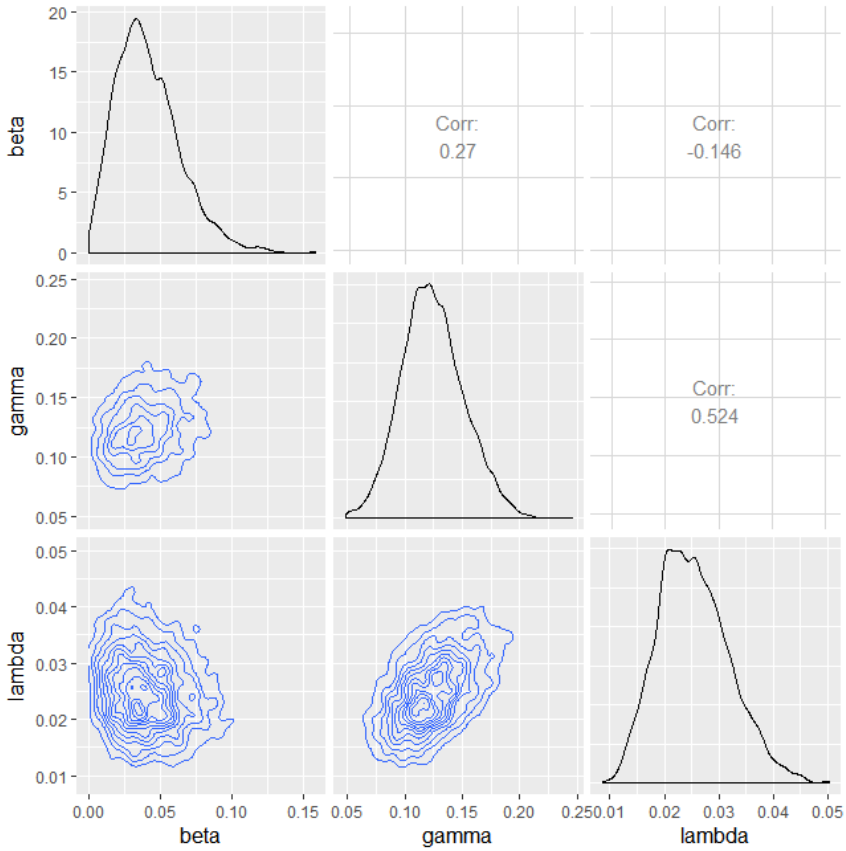

Figure 3.10.4: Paired scatter plots for the **non-spatial** model of the **Tanzania data application**.

Figure 3.10.5 shows paired scatter plots for correlation, density and contours for the spatial data. The density plots show that except κ, the other parameters have symmetric posterior distributions. Also, there is widespread evidence of very low or no correlation between the infection rate parameters and the covariance function parameters.

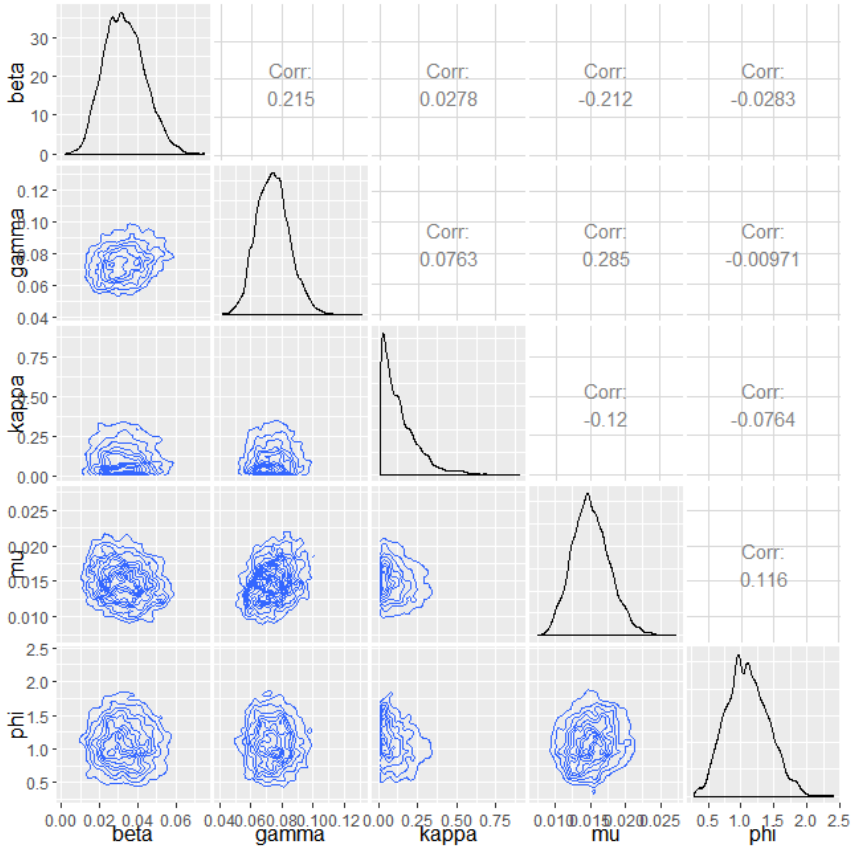

Figure 3.10.5: Paired scatter plots for the **spatial model** of the **Tanzania data application**.

3.11 Discussions

In this chapter, we have studied open population stochastic SIS epidemic model and spatial SIS epidemic model among a community of households.

We began with the non-spatial open population model which assumes that the infection rates, $\lambda > 0$ and $\beta > 0$, as well as the recovery rate, $\gamma > 0$, are independent of the spatial locations of the individual's household. This assumption simplifies the model and enables more straightforward implementation of the MCMC. Another assumption of the model is that the population size varies over time as individuals are allowed to join and exit the households at points in time. This later assumption adds to the complexity

of the model as it requires the computation of the infinitesimal transition rate matrix (G-matrix) each time the household size changes as a result of an individual joining or exiting a given household. A key computational burden is having to impute the state of the population when animals arrive and depart.

In Section 3.6 we studied a spatial SIS epidemic model which allows the background risk of infection λ to depend on the spatial locations of the individual's household whilst the local infection rate β and the recovery rate γ are space independent. In both the non-spatial and spatial stochastic SIS epidemic models, the overarching aim was to infer the model parameters using a Bayesian inference approach implemented in an MCMC framework. We developed easy-to-implement and efficient MCMC algorithms for estimation of the model parameters and these were successfully implemented using a simulated data set and effectively applied to a real life data set of a tick-borne diseases among Tanzania cattle.

Results from the non-spatial open population model show that our algorithms work well in terms of closeness of posterior parameter estimates (means) to the true parameter values.

On the other hand, results from the spatial model, *e.g.*, Figure 3.9.3, show that the spatial model performed comparatively poorly especially in the estimation of μ and κ, probably due to the problem of indentifiability. Further investigation is therefore required with the aim of fine-tuning the algorithms for optimal performance. However, the results obtained from the real life data example shows that our algorithms work well.

There are two key contributions of this chapter: first, MCMC algorithms which exploit extensive data augmentation schema for the estimation of the parameters of an open population household SIS epidemics were developed and successfully applied to both simulated data sets and real life data. Second, easy-to-implement MCMC algorithms

were developed for the estimation of spatial SIS epidemic model parameters. The flexibility of MCMC is fully utilized in this context given the complexity of spatial models, in general. The algorithms are found to work well upon application to both simulated data and real life data sets. We note that the MCMC algorithms developed here can easily be applied to a wide range of problems.

Chapter 4

Multiple Strains Model With Interactions

4.1 Motivation

In this Chapter, we introduce stochastic household-based SIS epidemic models for coinfecting diseases. These are an extension of the infectious disease model introduced in Chapter 2 for a single disease ($d = 1$). The models we develop here are generic and suitable for any number of diseases, d, but we will focus on the case $d = 2$ diseases. We begin by describing the two most prevalent forms of household-based SIS epidemic data for interacting diseases namely, the individual-based data (IBD) and the aggregate-based data (ABD). The ultimate aim is to develop Bayesian inferential tools in order to analyse such data and infer parameters. Markov Chain Monte Carlo (MCMC) algorithms are developed, tested with a simulated data and applied to a real life data set on tick-borne diseases among Tanzania cattle.

Coinfection occurs when a susceptible host becomes infected with two or more strains of a given pathogen (or with two or more pathogens each carrying different diseases.)

An individual host may acquire coinfection by being infected sequentially or simultaneously with different strains (or diseases). A major concern with disease coinfection is that coinfecting pathogens or strains usually interact with one another (Balmer and Tanner, 2011). Interactions within coinfection may lead to an increased susceptibility of the host to other infections due to waning immunity or decreased susceptibility of the host to similar strains due to cross-immunity. For example, infection with a strain of dengue fever has been found to enhance the transmission of another strain (see, for example, Ferguson et al. (1999)) and infection with HIV suppresses the immune system of the host making it more vulnerable for Tuberculosis transmission (see, for example, Newman and Ferrario (2013)). On the other hand, studies have observed the existence of cross-immunity between different subtypes of Influenza (see, for example, Epstein (2006)), with strong cross-protection existing among variants of antigenic drifts evolved from the same influenza subtype, see, for example, Barry et al. (2008). Understanding the transmission dynamics of disease coinfection is key to finding effective prophylactic and/or treatment measures to combat the diseases in an event of coepidemics, see, for example, Hoti et al. (2009) and Lipsitch (1997), for the use of vaccination in the prevention of Streptococcus pneumoniae and coinfection of Streptococcus pneumoniae and *Haemophilus influnzae*, respectively. Lipsitch (1997) observes that vaccination could offer full, partial or cross immunity to certain serotypes (strains) of the diseases. However, a serotype-targeted vaccine could give rise to an increased carriage of other out-competed non-target serotypes. This raises the question of how well diseases coinfection dynamics is understood.

There have been a few studies in the area of disease coinfection, see, for example, Slater et al. (2013) for coinfection of Malaria and Lymphatic Filariasis, Gao et al. (2016) for the coinfection of *Chlamydia trachomatis* and *pneumococcus*, Getahun et al. (2010) for coinfection of HIV and Tuberculosis, Sharp et al. (1997) for co-infection of multiple strains

of Influenza A viruses, and Neal and Huang (2015) for coinfection of four strains (HPV6, HPV11, HPV16, HPV18) of Human Papillomavirus (HPV) among a community of men who have sex with men (MSM), in addition to those already mentioned above. However, of all the studies mentioned above only Gao et al. (2016), Lipsitch (1997) and Neal and Huang (2015) studied disease coinfection within the SIS (susceptible $\rightarrow$ infective $\rightarrow$ susceptible) epidemics contexts, with only Neal and Huang (2015) employing Bayesian inference approach to estimate the parameters of the SIS model.

The model considered in this chapter is similar to those in Gao et al. (2016), Lipsitch (1997) and Neal and Huang (2015), but a number of differences exist. First whilst the model studied by Lipsitch (1997) and Neal and Huang (2015) assume that infected individuals recover at a constant rate which is serotype independent, the model we consider here assumes that the recovery rates of individuals are serotype dependent. This assumption makes sense in that it is unlikely that the infectivities of coinfecting diseases would be same. Also, the model by Gao et al. (2016) though assuming serotype dependent recovery rates, assumes that there is no simultaneous recovery of an individual infected with multiple strains, while the model we consider here allows individuals to recover simultaneously from multiple strains. Again, this assumption makes sense in that a successful treatment for a given strain may result to a simultaneous recovery from an immunologically similar strain.

Motivated by a rich set of tick-borne diseases data among Tanzanian cattle, which contains five *interacting* strains (*T.parva, T.mutans, A.marginale, B.bigemina, B.bovis*) of *Theileria Parva*, we seek to develop robust Bayesian inference approach for the analysis of the data. Note that we have already applied this data for a single disease case in Chapter 3. Ticks are known to be the most popular arthropods vector of both human and animal diseases with high rates of pathogenic coinfection which poses a global public health concern about the consequences of possible cotransmission to both human and

animal health (Moutailler et al., 2016). See, for example, Lou et al. (2017), Hersh et al. (2014), and Moutailler et al. (2016) for more on disease coinfection from tick-borne diseases.

We structure the remainder of this chapter as follows: data description is given in Section 4.2 for the two most prevalent household-based SIS epidemic data- the individual-based data (IBD) and aggregate-based data (ABD). We give the generic model setup including the construction of the various infinitesimal transition rate matrices (G-matrices) and calculation of the corresponding transition probability matrices (Q-matrices) in Section 4.3. We assume that the data are fully observed at a set of discrete time points and then relax this to allow the data to be only partially observed at the observation time points.

In Section 4.4, we give the procedures for the implementation of the MCMC algorithms with respect to IBD and ABD. First, we assume that data are fully observed and develop straightforward MCMC algorithms for the analyses of the fully observed household SIS data and later extend this to when data are only partially observed. This involves the development of an extensive data augmentation scheme Tanner and Wong (1987). Furthermore, in Section 4.5, we demonstrate how our approach is implemented using a simulated data set. In Section 4.6, the model and the MCMC algorithms developed are applied to a real-life tick-borne disease data. Finally, we give concluding remarks and discussions in Section 4.7.

4.2 Data Description

In this section, we describe the two forms of household-based SIS coepidemic data: the individual-based data (IBD) and the aggregate-based data (ABD).

4.2.1 Individual-based Data (IBD)

Given a household of size $h \geq 1$, the individual-based interacting diseases data holds information about the infectious status of each individual in the household with respect to $d > 1$ diseases. As in Section 2.2, we encode a susceptible 0 and an infective 1 so that each individual in a given household can belong to any of the 2^d possible states at a given point in time. Therefore, there are $2^{d\times h}$ possible states to which a household of size h can belong to at a point in time. We shall focus on the two-disease case ($d = 2$), so that there are 4 possible states $\{(0,0),(0,1),(1,0),(1,1)\}$ and 4^h states to which an individual and a households of size h can belong to at a given point in time, respectively. For $j = 1,\ldots,h$, and for $l = 1,2$, let $x_{jl}(t) \in \{0,1\}$ denote the infectious status of individual j for disease l. Also, let $\mathbf{x}_j(t) \in \{(0,0),(0,1),(1,0),(1,1)\}$ denote the infectious statuses of individual j for the 2 diseases, where

$$\mathbf{x}_j(t) = \begin{cases} (0,0) & \text{if susceptible to both diseases at time } t, \\ (0,1) & \text{if susceptible to disease 1 and infected with disease 2 at time } t, \\ (1,0) & \text{if infected with disease 1 and susceptible to disease 2 at time } t, \\ (1,1) & \text{if coinfected with both diseases 1 and 2 at time } t. \end{cases} \tag{4.2.1}$$

Then, the data $\mathbf{x}(t) = \big(\mathbf{x}_1(t), \mathbf{x}_2(t), \ldots, \mathbf{x}_h(t)\big)$ is the infectious state of the household at time t.

Now for $i = 1,2,\ldots,N$, let $h_i \geq 1$ and n_i denote the size and number of observation time points of household i, respectively. Also, for $k = 0,1,2,\ldots,n_i$, let t_{ik} denote the k^{th} observation time point of household i. Then, for $j = 1,2,\ldots,h_i$, it follows that

- $\mathbf{t}_i = \big(t_{i0}, t_{i1}, t_{i2}, \ldots, t_{in_i}\big)$ are the observation times of household i, where $t_{i0} = t_0 = 0$ (by independent households assumption with no time varying factors).

- $x_{ijl}(t_{ik}) \in \{0,1\}$ is the infectious status of the j^{th} individual in household i for the l^{th} disease at time point t_{ik}.

- $\mathbf{x}_{ij}(t_{ik}) \in \{0,1\}^2$ is the vector of infectious status of the j^{th} individual in household i for $d = 2$ diseases at time point t_{ik}.

- $\mathbf{x}_i(t_{ik}) \in \{0,1\}^{2\times h}$ is the infectious state of household i at time point t_{ik}.

- $\mathbf{x}_i(\mathbf{t}_i) = \big(\mathbf{x}_i(t_{i1}), \mathbf{x}_i(t_{i2}), \ldots, \mathbf{x}_i(t_{in_i})\big)$ are the i^{th} infectious statuses of household i at the set of time points, $\mathbf{t}_i$.

Therefore, $\mathbf{x}(\mathbf{t}) = \big(\mathbf{x}_1(\mathbf{t}_1), \mathbf{x}_2(\mathbf{t}_2), \ldots, \mathbf{x}_N(\mathbf{t}_N)\big)$ is the full IBD for N independent households. When we observe every individual at every given observation time point, we say that $\mathbf{x}(\mathbf{t})$ is completely observed and there are no missing values.

Usually, we only have partial observations of the data at the observation time points which leads to incomplete (partially observed) data. Let $\mathbf{y}_{ij}(t_{ik}) = \big(y_{ij1}(t_{ik}), y_{ij2}(t_{ik})\big)$ denote the partially observed infectious status of the j^{th} individual of household i for diseases 1 and 2 at time point t_{ik}. We assume that when an individual is unobserved at a given time point, their infectious status for all diseases is unknown. Here, we assume that an individual is missing completely at random (MCAR) (*see*, Rubin (1987)). Then

$$\mathbf{y}_{ij}(t_{ik}) = \begin{cases} \mathbf{x}_{ij}(t_{ik}) & \text{if observed,} \\ (2,2) & \text{if unobserved.} \end{cases} \tag{4.2.2}$$

Therefore, we have that

- $\mathbf{y}_i(t_{ik}) = \big(\mathbf{y}_{i1}(t_{ik}), \mathbf{y}_{i2}(t_{ik}), \ldots, \mathbf{y}_{ih}(t_{ik})\big)$ is the infectious state of household i at time point t_{ik}.

- $\mathbf{y}_i(\mathbf{t}_i) = \big(\mathbf{y}_i(t_{i1}), \mathbf{y}_i(t_{i2}), \ldots, \mathbf{y}_i(t_{in_i})\big)$ are the i^{th} household partially observed individual-based interacting diseases data over the set of time points, $\mathbf{t}_i$.

Then, $\mathbf{y}(\mathbf{t}) = \big(\mathbf{y}_1(\mathbf{t}_1), \mathbf{y}_2(\mathbf{t}_2), \ldots, \mathbf{y}_N(\mathbf{t}_N)\big)$ is the full partially observed individual-based interacting diseases data for N households over the set of observation time points, $\mathbf{t}$.

4.2.2 Aggregate-based Interacting Diseases Data (ABD)

As in Chapter 2, we consider cases where the disease status of individuals are aggregated at household level. That is, for $d = 2$, we know how many individuals in each of the categories $\{(0,0),(0,1),(1,0),(1,1)\}$ at a given timepoint. We call this the aggregate-based data (ABD).

For a household of size h at a given point in time, let

- n_0 denote the number of individuals not infected by any of the diseases,
- n_1 denote the number of individuals infected with disease 1 only,
- n_2 denote the number of individuals infected with disease 2 only,
- n_{12} denote the number of individuals infected with both diseases 1 and 2,

where $n_0 + n_1 + n_2 + n_{12} = h$. Then $\tilde{\mathbf{x}}(t) = \big(\tilde{x}_1(t), \tilde{x}_2(t), \tilde{x}_3(t), \tilde{x}_4(t)\big) = (n_0, n_1, n_2, n_{12})$ is the state of a given household at time t. In other words, individuals of the household are divided into four categories (category 1 for no infection, category 2 for infection with disease 1, category 3 for infection with disease 2, and category 4 for coinfection with both diseases) at a given point in time. Hence, $\tilde{x}_j(t)$ is the number of individuals in category j at time t $(j = 1, 2, 3, 4)$. Then for $i = 1, 2, \ldots, N$ and for $k = 1, 2, \ldots, n_i$

- $\tilde{x}_{ij}(t_{ik})$ are the number of individuals of household i in category j,
- $\tilde{\mathbf{x}}_i(t_{ik}) = \big(\tilde{x}_{i1}(t_{ik}), \tilde{x}_{i2}(t_{ik}), \tilde{x}_{i3}(t_{ik}), \tilde{x}_{i4}(t_{ik})\big)$ is the state of household i at time t_{ik},
- $\tilde{\mathbf{x}}_i(\mathbf{t}_i) = \big(\tilde{\mathbf{x}}_i(t_{i1}), \tilde{\mathbf{x}}_i(t_{i2}), \ldots, \tilde{\mathbf{x}}_i(t_{in_i})\big)$ are the SIS aggregated disease coinfection data of household i at the sets of time $\mathbf{t}_i$.

Therefore, the data $\tilde{\mathbf{x}}(\mathbf{t}) = \big(\tilde{\mathbf{x}}_1(\mathbf{t}_1), \tilde{\mathbf{x}}_2(\mathbf{t}_2), \ldots, \tilde{\mathbf{x}}_N(\mathbf{t}_N)\big)$ are the full aggregate-base SIS coinfection data for N households at time $\mathbf{t}$. When every individual is observed, $\tilde{\mathbf{x}}$ is said to be complete. However, in practice, this is rarely the case as most infectious disease data are only partially observed or incomplete. When only a subset of the households is observed, we let

- $\tilde{y}_{ij}(t_{ik})$ are the observed number of individuals of household i in category j,
- $\tilde{\mathbf{y}}_i(t_{ik}) = \big(\tilde{y}_{i1}(t_{ik}), \tilde{y}_{i2}(t_{ik}), \tilde{y}_{i3}(t_{ik}), \tilde{y}_{i4}(t_{ik})\big)$ is the observed state of household i at time t_{ik},
- $\tilde{\mathbf{y}}_i(\mathbf{t}_i) = \big(\tilde{\mathbf{y}}_i(t_{i1}), \tilde{\mathbf{y}}_i(t_{i2}), \ldots, \tilde{\mathbf{y}}_i(t_{in_i})\big)$ are the observed SIS aggregated disease coinfection data of household i at the sets of time $\mathbf{t}_i$,
- $\tilde{\mathbf{y}}(\mathbf{t}) = \big(\tilde{\mathbf{y}}_1(\mathbf{t}_1), \tilde{\mathbf{y}}_2(\mathbf{t}_2), \ldots, \tilde{\mathbf{y}}_N(\mathbf{t}_N)\big)$ is the full partially observed ABD for N households over the set of time points data $\mathbf{t}$,

where $\sum_{j=1}^{4} \tilde{\mathbf{y}}_j(t) = \tilde{y}_1(t) + \tilde{y}_2(t) + \tilde{y}_3(t) + \tilde{y}_4(t) \leq h$. When $\sum_{j=1}^{4} \tilde{\mathbf{y}}_j(t) < h$, we say that the data $\tilde{\mathbf{y}}$ is incomplete. In Section 4.4.4, we discuss in details the implementation of data augmentation schema for the analysis of a partially observed aggregated SIS disease coinfection data.

4.3 Generic Model Setup

In this Section, we provide the details of the model construction including the infinitesimal rate matrices (G-matrices) and the calculation of the corresponding transition probability matrices (Q-matrices) for both IBD and ABD. First, we give the generic model setup and later describe separately the construction of the G-matrices for IBD and ABD beginning with IBD.

Given a population of M individuals endemic with two disease strains (disease 1 and

disease 2). Let the population be divided into N non-overlapping households with N_h households of size $h \geq 1$ such that $\sum_{h=1}^{H} hN_h = M$ and $\sum_{h=1}^{H} N_h = N$, where H is the maximum household size. Individuals in each household are further divided into one of four (4) mutually exclusive epidemiological sub-classes, namely, state S (susceptible to both diseases or no infection), state I_1 (infection with disease 1), state I_2 (infection with disease 2), and state I_{12} (coinfection with both diseases). For $l = 1, 2$, we make the following assumptions

- the households are mutually independent,
- there exists a disease-specific global force of infection $\lambda_l > 0$,
- events of infection and recovery can happen sequentially or simultaneously,
- within-household infection transmission happens at a disease-specific rate $\beta_l > 0$,
- recovery from disease happen at disease-specific a rate $\gamma_l > 0$,
- within-household simultaneous infection with both diseases happens at rate $\beta_{12} > 0$,
- simultaneous global infection with both diseases happens at rate $\lambda_{12} > 0$,
- simultaneous recovery from both diseases happens at rate $\gamma_{12} > 0$,
- recovery from disease does not confer immunity so a recovered individual immediately returns to the susceptible state and may be reinfected. all contacts are made according to mutually independent Poisson point processes.

In addition, let $\phi_{12} = \phi_{21} = \phi$ denote the relative risk of an individual acquiring diseases 1 and 2 sequentially compared to an individual who is susceptible to both diseases. The parameter ϕ can take the following values: $\phi = 0$ (no coinfection), $\phi = 1$ (the two

diseases behave independently), $\phi > 1$ (increased risk), and $\phi < 1$ (reduced risk). Figure 4.3.1 shows a schematic representation of the stochastic SIS disease confection model.

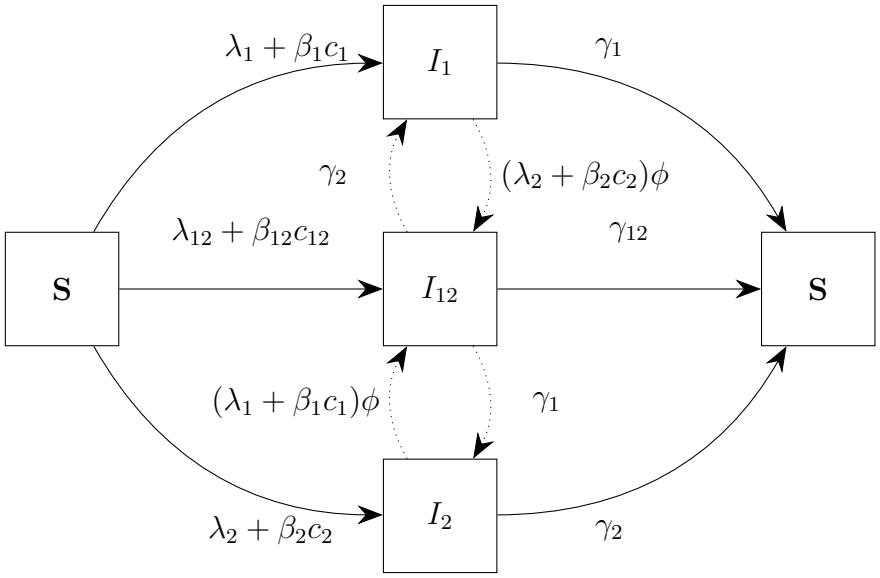

Figure 4.3.1: Schematic representation of the two diseases SIS epidemic model with interaction.**S** = susceptible; I_1 = infected with diseases 1; I_2 = infected with disease 2; I_{12} = infected with both diseases, where c_l is the number of individuals infected with strain l. Note that the transition rates given in this diagram are for the IBD case with $c_1 = |I_1| + |I_{12}|$, $c_2 = |I_2| + |I_{12}|$ and $c_{12} = |I_{12}|$.

G-matrix for diseases coinfection IBD

Given a household of size h, let $u_{jl} \in \{0, 1\}$ denote the infectious status of individual j for disease l, then $\mathbf{u}_j \in \{0, 1\}^2$ are the infectious statuses of individual j for $d = 2$ diseases at a point in time ($l = 1, 2; j = 1, 2, \ldots, h$). Therefore, $\mathbf{u}_j = \big(u_{j1}, u_{j2}\big)'$ and $\mathbf{u}(t) = \big(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_h\big)'$, where $\mathbf{u}(t)$ is the infectious status of the household at time t. Let $\mathbf{e}_{jl}$ denote a vector of length $2 \times h$ in which only the l^{th} element of its j^{th} component (of length $d = 2$) is equal to 1 and the rest are zeros. Then, we have the following transitions

Infection:

1. For $l = 1, 2$, if $u_{jl} = 0$

   - the household state transition $\mathbf{u} \to \mathbf{u} + \mathbf{e}_{jl}$ corresponds to the infection of the j^{th} individual with disease l.

2. For $l, l' = 1, 2$ $(l \neq l')$, if $u_{jl} = 0$ and $u_{jl'} = 1$,

   - the household transition $\mathbf{u} \to \mathbf{u} + \mathbf{e}_{jl}$ corresponds to the infection of the j^{th} individual by disease l having been infected with disease l' at present.

3. For $l, l' = 1, 2$ $(l \neq l')$, if $u_{jl} = 0$ and $u_{jl'} = 0$,

   - the household transition $\mathbf{u} \to \mathbf{u} + (\mathbf{e}_{jl} + \mathbf{e}_{jl'})$ corresponds to the infection of the j^{th} individual with both diseases l and l' simultaneously.

Recovery:

1. For $l = 1, 2$, if $u_{jl} = 1$

   - the household state transition $\mathbf{u} \to \mathbf{u} - \mathbf{e}_{jl}$ corresponds to the recovery of the j^{th} individual from disease l.

2. For $l, l' = 1, 2$ $(l \neq l')$, if $u_{jl} = 1$ and $u_{jl'} = 1$,

   - the household transition $\mathbf{u} \to \mathbf{u} - \mathbf{e}_{jl}$ corresponds to the recovery of the j^{th} individual from disease l having been infected with both diseases l and l' at present.

3. For $l, l' = 1, 2$ $(l \neq l')$, if $u_{jl} = 1$ and $u_{jl'} = 1$,

- the household transition $\mathbf{u} \to \mathbf{u} - (\mathbf{e}_{jl} + \mathbf{e}_{jl'})$ corresponds to the recovery of the j^{th} individual from both diseases l and l' simultaneously.

Now let

- c_l denote the number of individuals presently infected with disease l,
- c_{12} denote the number of individuals presently co-infected with diseases 1 and 2,

where

$$c_l = \sum_{j=1}^{h} u_{jl}, \tag{4.3.1}$$

and

$$c_{12} = \sum_{j=1}^{h} u_{j1} u_{j2}. \tag{4.3.2}$$

Then, provided that $u_{jl} = 0$, the infinitesimal rate of moving from state $\mathbf{u}$ to state $\mathbf{u}+\mathbf{e}_{jl}$ is given by

$$\left\{\lambda_l + \beta_l c_l\right\} \phi_{ll'}^{u_{jl'}}, \tag{4.3.3}$$

where u_{jl} is the infectious status of individual j for disease l. Therefore, we define the 4^h *by* 4^h infinitesimal rate matrix, $G^{(h)} = \left(g_{uv}^{(h)}\right)$ for IBD as follows.

$$g_{\mathbf{uv}}^{(h)} = \begin{cases} \lambda_1 + \beta_1 c_1 & \text{if } u_{j1} = 0,\ u_{j2} = 0 \text{ and } \mathbf{v} = \mathbf{u} + \mathbf{e}_{j1}, \\ \lambda_2 + \beta_2 c_2 & \text{if } u_{j1} = 0,\ u_{j2} = 0 \text{ and } \mathbf{v} = \mathbf{u} + \mathbf{e}_{j2}, \\ \lambda_{12} + \beta_{12} c_{12} & \text{if } u_{j1} = 0,\ u_{j2} = 0 \text{ and } \mathbf{v} = \mathbf{u} + (\mathbf{e}_{j1} + \mathbf{e}_{j2}), \\ \big(\lambda_1 + \beta_1 c_1\big)\phi_{21} & \text{if } u_{j1} = 0,\ u_{j2} = 1 \text{ and } \mathbf{v} = \mathbf{u} + \mathbf{e}_{j1}, \\ \big(\lambda_2 + \beta_2 c_2\big)\phi_{12} & \text{if } u_{j1} = 1,\ u_{j2} = 0 \text{ and } \mathbf{v} = \mathbf{u} + \mathbf{e}_{j2}, \\ \gamma_1 & \text{if } u_{j1} = 1, \\ \gamma_2 & \text{if } u_{j2} = 1 \text{ and } \mathbf{v} = \mathbf{u} - \mathbf{e}_{j2}, \\ \gamma_{12} & \text{if } u_{j1} = 1,\ u_{j2} = 1 \text{ and } \mathbf{v} = \mathbf{u} - (\mathbf{e}_{j1} + \mathbf{e}_{j2}), \\ -\sum\limits_{\mathbf{w} \neq \mathbf{u}} g_{\mathbf{uw}}^{(h)} & \text{if } \mathbf{v} = \mathbf{u}, \\ 0 & \text{Otherwise,} \end{cases} \tag{4.3.4}$$

for $\lambda_1, \lambda_2, \lambda_{12}, \beta_1, \beta_2, \beta_{12}, \gamma_1, \gamma_2, \gamma_{12}, \phi_{12}, \phi_{21} > 0$ and $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w} \in \mathcal{S}$. Observe that setting $\lambda_{12} = \beta_{12} = \gamma_{12} = 0$ and $\phi = 1$ makes our model equivalent to the one disease case ($d = 1$) model introduced in Chapter 2 and extended in Chapter 3, as the diseases would behave independently.

G-matrix for diseases coinfection ABD

For a household of size h and for $l = 1, 2$, let $\tilde{u}_l$ denote the number of individuals currently infected with disease l. Also, let $\tilde{u}_{12}$ denote the number of individuals currently coinfected with diseases 1 and 2 and let $\tilde{u}_0$ denote the number of individuals currently not infected (or coinfected) by any of the two diseases (or by both diseases). Then the vector $\tilde{\mathbf{u}} = \big(\tilde{u}_{12}, \tilde{u}_1, \tilde{u}_2, \tilde{u}_0\big)$ is the infectious state of a given household at time t, where

- $\tilde{u}_{12}$ is the number of individuals currently co-infected with both diseases 1 & 2,
- $\tilde{u}_1$ is the number of individuals currently infected with diseases 1 only,

- $\tilde{u}_2$ is the number of individuals currently infected with diseases 2 only,
- $\tilde{u}_0$ is the number of individuals currently not infected at all,

such that $\sum \tilde{\mathbf{u}} = \tilde{u}_{12} + \tilde{u}_1 + \tilde{u}_2 + \tilde{u}_0 = h$. Thus, the transitions allowed are as follows.

$$\tilde{\mathbf{u}} \to \begin{cases} (\tilde{u}_{12}, \tilde{u}_1 - 1, \tilde{u}_2, \tilde{u}_0 + 1): & \textit{recovery from disease } 1, \\ (\tilde{u}_{12}, \tilde{u}_1, \tilde{u}_2 - 1, \tilde{u}_0 + 1): & \textit{recovery from disease } 2, \\ (\tilde{u}_{12} - 1, \tilde{u}_1, \tilde{u}_2, \tilde{u}_0 + 1): & \textit{simultaneous recovery from diseases } 1 \ \& \ 2, \\ (\tilde{u}_{12} - 1, \tilde{u}_1, \tilde{u}_2 + 1, \tilde{u}_0): & \textit{recovery from disease } 1 \textit{ having been infected with } 1 \ \& \ 2, \\ (\tilde{u}_{12} - 1, \tilde{u}_1 + 1, \tilde{u}_2, \tilde{u}_0): & \textit{recovery from disease } 2 \textit{ having been infected with } 1 \ \& \ 2, \\ (\tilde{u}_{12}, \tilde{u}_1 + 1, \tilde{u}_2, \tilde{u}_0 - 1): & \textit{infection with disease } 1, \\ (\tilde{u}_{12}, \tilde{u}_1, \tilde{u}_2 + 1, \tilde{u}_0 - 1): & \textit{infection with disease } 2, \\ (\tilde{u}_{12} + 1, \tilde{u}_1, \tilde{u}_2, \tilde{u}_0 - 1): & \textit{simultaneous infection with diseases } 1 \ \& \ 2, \\ (\tilde{u}_{12} + 1, \tilde{u}_1, \tilde{u}_2 - 1, \tilde{u}_0): & \textit{co-infection by disease } 1 \textit{ having been infected with } 2\ , \\ (\tilde{u}_0 + 1, \tilde{u}_1 - 1, \tilde{u}_2, \tilde{u}_0): & \textit{co-infection by disease } 2 \textit{ having been infected with } 1. \end{cases} \tag{4.3.5}$$

Then we have the following transition rates:

Infection:

1. The infinitesimal rate of infection with disease 1 is given by

$$\big(\lambda_1 + \beta_1 \tilde{u}_1\big)(\tilde{u}_0 + \tilde{u}_2). \tag{4.3.6}$$

2. The infinitesimal rate of infection with disease 2 is given by

$$\big(\lambda_2 + \beta_2 \tilde{u}_2\big)(\tilde{u}_0 + \tilde{u}_1). \tag{4.3.7}$$

3. The rate of simultaneous coinfection with both diseases is

$$\big(\lambda_{12} + \beta_{12} \tilde{u}_{12}\big)\tilde{u}_0. \tag{4.3.8}$$

4. Provided that $\tilde{u}_2 \geq 1$, the rate of infection with disease 1 having been infected with disease 2 is

$$\big(\lambda_1 + \beta_1 \tilde{u}_1\big)\tilde{u}_2\phi. \tag{4.3.9}$$

5. Provided that $\tilde{u}_1 \geq 1$, the rate of infection with disease 2 having been infected with disease 1 is

$$\big(\lambda_2 + \beta_2 \tilde{u}_2\big)\tilde{u}_1\phi. \tag{4.3.10}$$

Recovery:

1. Provided that $\tilde{u}_1 \geq 1$, the rate of recovery from disease 1 is given by

$$\gamma_1 \tilde{u}_1, \tag{4.3.11}$$

2. Provided that $\tilde{u}_2 \geq 1$, the rate of recovery from disease 2 is given by

$$\gamma_2 \tilde{u}_2, \tag{4.3.12}$$

3. Provided that $\tilde{u}_{12} \geq 1$,

   - the rate of simultaneous recovery from both diseases 1 and 2 is

$$\gamma_{12} \tilde{u}_{12}, \tag{4.3.13}$$

   - the rate of recovery from disease 1 having been infected with both diseases 1 and 2 is

$$\gamma_1 \tilde{u}_{12}, \tag{4.3.14}$$

   - the rate of recovery from disease 2 having been infected with both diseases 1 and 2 is

$$\gamma_2 \tilde{u}_{12}. \tag{4.3.15}$$

Therefore, given that there are $\binom{h+2^d-1}{h}$ possible states to which a household of size h can belong to at any point in time, we define a $\binom{h+3}{h} \times \binom{h+3}{h}$ infinitesimal rate matrix $G^{(h)} = (g^{(h)}_{\tilde{\mathbf{u}}\tilde{\mathbf{v}}})$ of an ABD as follows:

$$g^{(h)}_{\tilde{\mathbf{u}}\tilde{\mathbf{v}}} = \begin{cases} \gamma_1 \tilde{u}_1 & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12}, \tilde{u}_1 - 1, \tilde{u}_2, \tilde{u}_0 + 1) \\ \gamma_2 \tilde{u}_2 & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12}, \tilde{u}_1, \tilde{u}_2 - 1, \tilde{u}_0 + 1) \\ \gamma_{12} \tilde{u}_{12} & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12} - 1, \tilde{u}_1, \tilde{u}_2, \tilde{u}_0 + 1) \\ \gamma_1 \tilde{u}_{12} & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12} - 1, \tilde{u}_1, \tilde{u}_2 + 1, \tilde{u}_0) \\ \gamma_2 \tilde{u}_{12} & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12} - 1, \tilde{u}_1 + 1, \tilde{u}_2, \tilde{u}_0) \\ \big(\lambda_1 + \beta_1 \tilde{u}_1\big)(\tilde{u}_0 + \tilde{u}_2) & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12}, \tilde{u}_1 + 1, \tilde{u}_2, \tilde{u}_0 - 1) \\ \big(\lambda_2 + \beta_2 \tilde{u}_2\big)(\tilde{u}_0 + \tilde{u}_1) & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12}, \tilde{u}_1, \tilde{u}_2 + 1, \tilde{u}_0 - 1) \\ \big(\lambda_{12} + \beta_{1,2} \tilde{u}_{1,2}\big) \tilde{u}_0 & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{1,2} + 1, \tilde{u}_1, \tilde{u}_2, \tilde{u}_0 - 1) \\ \big(\lambda_1 + \beta_1 \tilde{u}_1\big) \tilde{u}_2 \phi & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12} + 1, \tilde{u}_1, \tilde{u}_2 - 1, \tilde{u}_0) \\ \big(\lambda_2 + \beta_2 \tilde{u}_2\big) \tilde{u}_1 \phi & \text{if } \tilde{\mathbf{v}} = (\tilde{u}_{12} + 1, \tilde{u}_1 - 1, \tilde{u}_2, \tilde{u}_0) \\ -\sum\limits_{\tilde{\mathbf{w}} \neq \tilde{\mathbf{u}}} g^{(h)}_{\tilde{\mathbf{u}}\tilde{\mathbf{w}}} & \text{if } \tilde{\mathbf{v}} = \tilde{\mathbf{u}}, \\ 0 & \text{Otherwise,} \end{cases} \tag{4.3.16}$$

for $\lambda_1, \lambda_2, \lambda_{12}, \beta_1, \beta_2, \beta_{12}, \gamma_1, \gamma_2, \gamma_{12}, \phi > 0$; $\tilde{\mathbf{u}}$, $\tilde{\mathbf{v}}$, $\tilde{\mathbf{w}} \in \mathcal{S}$.

4.3.1 Transition Probability Matrix (Q-matrix)

As outlined in Section 2.3.1, we calculate the transition probability matrices $Q^{(h)} = (q^{(h)}_{\mathbf{ij}})$ as a matrix exponential given by $\mathbf{Q} = \exp(t\mathbf{G})$.

Note that the Q-matrix here with its corresponding G-matrix is a $4^h \times 4^h$ and a $\binom{h+3}{h} \times \binom{h+3}{h}$ transition probability matrix for IBD and ABD, respectively.

4.4 Bayesian Inference on household-based SIS interacting diseases model

In this section, we develop a Bayesian inference approach to infer the parameters of an SIS diseases coinfection model. Specifically, we develop MCMC algorithms for the analysis of a disease coinfection data with respect to the individual-based data (IBD) and the aggregate-based data (ABD). Throughout, for ease of exposition, we shall focus our descriptions on closed (constant) population SIS, although it is straightforward to extend this to allow varying population sizes over time by following the methods described in Section 3.5. We shall begin with when data are completely observed and later extend this to when data are only partially observed.

4.4.1 Inference on Completely Observed Household SIS Data

Setup

As before, let $\mathbf{x}$ denote the data generated from the parametric model with parameters

$$\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_{1,2}, \beta_1, \beta_2, \beta_{1,2}, \gamma_1, \gamma_2, \gamma_{1,2}, \phi).$$

For the purpose of inference, we need to draw samples from the posterior distribution of the parameters given data, $\pi(\boldsymbol{\theta}|\mathbf{x})$, using MCMC algorithms. When data are completely observed, it is straightforward to employ RWM to draw samples from the joint posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Given that the parameters values are rates except the relative risk ϕ, and all parameters are positive, the Gamma distribution is a natural choice of prior distribution. That is, for $j = 1, 2, \ldots, 10$, the prior distribution on the parameters is

$$\boldsymbol{\theta}_j \sim Gamma(A_{\boldsymbol{\theta}_j}, B_{\boldsymbol{\theta}_j})$$

. (4.4.1)

where $A_{\boldsymbol{\theta}_j} > 0$ and $B_{\boldsymbol{\theta}_j} > 0$ are hyper-parameter. Then we calculate the likelihood function, $\pi(\mathbf{x}|\boldsymbol{\theta})$ as

$$\begin{aligned} L(\boldsymbol{\theta};\mathbf{x}) := \pi(\mathbf{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{N}\prod_{k=1}^{n_i}\Big\{\mathbb{P}\big(X_i(t_{ik}) = \mathbf{x}_i(t_{ik})|X_i(t_{i,k-1}) = \mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta}\big)\Big\} \\ &= \prod_{i=1}^{N}\prod_{k=1}^{n_i}\Big\{\pi\big(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta}\big)\Big\}, \end{aligned} \tag{4.4.2}$$

where $\pi(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta})$ is the probability of being in state $\mathbf{x}_i(t_{ik})$ at time point t_{ik} from state $\mathbf{x}_i(t_{i,k-1})$ at time point $t_{i,k-1}$.

The posterior distribution of the parameters given the data, $\pi(\boldsymbol{\theta}|\mathbf{x})$ is thus given by

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{x}) &\propto \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto \left\{\prod_{i=1}^{N}\prod_{k=1}^{n_i}\Big\{\mathbb{P}\big(X_i(t_{ik}) = \mathbf{x}_i(t_{ik})|X_i(t_{i,k-1}) = \mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta}\big)\Big\}\right\} \\ &\times \prod_{j=1}^{10}\pi(\boldsymbol{\theta}_j) \\ &\propto \left\{\prod_{i=1}^{N}\prod_{k=1}^{n_i}\Big\{\pi\big(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{i,k-1}), \boldsymbol{\theta}\big)\Big\}\right\} \\ &\times \prod_{j=1}^{10}\boldsymbol{\theta}_j^{A_{\boldsymbol{\theta}_j}-1}e^{-B_{\boldsymbol{\theta}_j}\boldsymbol{\theta}_j}. \end{aligned} \tag{4.4.3}$$

Finally, using the RWM algorithms described in Algorithm 4 propose $\boldsymbol{\theta}^{prop}$ from a multivariate Gaussian distribution with mean $\boldsymbol{\theta}^{curr}$ (the current values of the parameters) and a proposal covariance matrix Σ. Apply the adaptive schemes given in Section 2.4 for the MCMC optimality. Then with probability

$$\alpha(\boldsymbol{\theta}^{curr}, \boldsymbol{\theta}^{prop}) \quad = \quad \min\left\{1, \frac{\pi(\mathbf{x}|\boldsymbol{\theta}^{prop})\pi(\boldsymbol{\theta}^{prop})}{\pi(\mathbf{x}|\boldsymbol{\theta}^{curr})\pi(\boldsymbol{\theta}^{curr})}\right\}, \tag{4.4.4}$$

accept $\boldsymbol{\theta}^{prop}$. Note that the MCMC steps described here are generic and can be applied to both IBD and ABD since the primary focus at this stage is on updating the model parameters and not the data given that the data is assumed to be fully observed.

4.4.2 Inference on Partially Observed co-epidemics

In this section, we give a Bayesian inference approach for partially observed household-based SIS co-epidemics with respect to the data forms considered here, IBD and ABD.

4.4.3 Bayesian Inference for Partially Observed IBD

Let $\mathbf{y}$ denote the partially observed data whose likelihood function given the parameters $\boldsymbol{\theta}$, $\pi(\mathbf{y}|\boldsymbol{\theta})$, is intractable. Let $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ denote the complete data so that the likelihood function $\pi\big(\mathbf{x} = (\mathbf{y}, \mathbf{z})|\boldsymbol{\theta}\big)$ becomes tractable, where $\mathbf{z}$ are additional imputed information. In Sections 2.4 and 3.5.3, we outlined how to obtain samples from the joint posterior $\pi(\boldsymbol{\theta}|\mathbf{z}, \mathbf{y})$ in the case of a single disease and in the case involving varying population sizes, respectively. In both cases, the *data augmentation* scheme involves alternating between updating $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ and $\pi\big(\boldsymbol{\theta}|\mathbf{x} = (\mathbf{z}, \mathbf{y})\big)$.

Here, we follow the descriptions in Sections 2.4 and 3.5.3 and proceed as follows. For $i = 1, 2, \ldots, N$, if an individual is observed, set $\mathbf{x}_{ij}(t_k) = \mathbf{y}_{ij}(t_k)$, otherwise, impute the missing information $\mathbf{z}_{ij} \in \{0, 1\}^2$ by either setting $\mathbf{x}_{ij}(t_k) = \mathbf{x}_{ij}(t_{k-1})$ or choosing $x_{ijl}(t_k)$ uniformly from $\{0, 1\}$, where $l = 1, 2$. Then calculate the probability of observing the

complete data $\mathbf{x}$ given $\mathbf{y}$ and the parameters $\boldsymbol{\theta}$, $\pi(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ by

$$\begin{aligned}\pi(\mathbf{x}=(\mathbf{z},\mathbf{y})|\mathbf{y},\boldsymbol{\theta}) &\propto \prod_{i=1}^{N}\prod_{k=1}^{n_i}\Big\{\pi\big(\mathbf{y}_i(t_{ik})|\mathbf{x}_i(t_{ik})\big)\\ &\times \pi\big(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta}\big)\pi\big(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}),\boldsymbol{\theta}\big)\Big\}. \quad (4.4.5)\end{aligned}$$

Then, assuming that k is neither the first nor the last timepoint, we have

$$\begin{aligned}\pi(\mathbf{z}_{ij}(t_{ik})|\boldsymbol{\theta},\mathbf{x}_{-ij}(t_{ik})) &\propto \pi(\mathbf{x}_i(t_{ik}),\mathbf{x}_{ij}(t_{ik})=\mathbf{z}_{ij}(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})\\ &\times \pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik}),\mathbf{x}_{ij}(t_{ik})=\mathbf{z}_{ij}(t_{ik}),\boldsymbol{\theta}) \quad (4.4.6)\end{aligned}$$

where $\mathbf{x}_{-ij}(t_{ik})$ is the complete data vector for the infectious state of household i excluding $\mathbf{x}_{ij}(t_{ik})$.

Independence Sampler

To update $\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$, we develop an Independence Sampler algorithms similar to that described in Algorithm 8 and propose to switch states and set $\mathbf{z}_{ij}^{prop}(t_{ik}) = 1-\mathbf{z}_{ij}(t_{ik})$ (or $z_{ijl}^{prop} = 1-z_{ijl}, l=1,2$), so that $\mathbf{x}_i^{prop}(t_{ik}) = \big(\mathbf{x}_i(t_{ik}),\mathbf{x}_{ij}(t_{ik})=\mathbf{z}_{ij}^{prop}(t_{ik})\big)$, for every $\mathbf{y}_{ij}(t_k)=(2,2)$. In other words, we propose to explore other possibilities in {(0,0), (0,1), (1,0), (1,1)}. Then, accept $\mathbf{z}_{ij}^{prop}(t_{ik})$ with the probability which depends on the three time points t_{ik-1}, t_{ik} and t_{ik+1} and given by

$$\alpha \leftarrow min\Big\{1, \frac{\pi(\mathbf{x}_i^{prop}(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i^{prop}(t_{ik}),\boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik})),\boldsymbol{\theta})}\Big\}. \quad (4.4.7)$$

In what follows, we shall give Bayesian inference approach for an SIS coepidemics data of the ABD form.

4.4.4 Inference on Partially Observed ABD

We outline data augmentation scheme for a partially observed coepidemics ABD within the SIS contexts in household settings.

Let $\tilde{\mathbf{y}} = \tilde{\mathbf{y}}(\mathbf{t})$ denote the partially observed aggregate-based data at time $\mathbf{t}$ from our parametric model with parameters $\boldsymbol{\theta}$. The likelihood function $\pi(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ is rarely tractable. However, given the complete data $\tilde{\mathbf{x}} = (\tilde{\mathbf{z}}, \tilde{\mathbf{y}})$, the likelihood $\pi(\tilde{\mathbf{x}} = (\tilde{\mathbf{z}}, \tilde{\mathbf{y}})|\boldsymbol{\theta})$ becomes tractable, where $\tilde{\mathbf{z}}$ are imputed values. For a household of size h and which is observed n times, as noted earlier, the observed infectious state of the household at time t_k, $\tilde{\mathbf{y}}(t_k)$ is said to be only partially observed when $\tilde{y}_s = \sum_{j=1}^{4} \tilde{y}_j(t_k) < h$. We proceed as follows. Whenever $\tilde{y}_s < h$ (or incomplete data), set $\tilde{\mathbf{x}}(t_k) = \tilde{\mathbf{y}}(t_k) + \tilde{\mathbf{z}}(t_k)$, otherwise set $\tilde{\mathbf{x}}(t_k) = \tilde{\mathbf{y}}t_k$ (no need for data imputation). Here, the imputed data $\tilde{\mathbf{z}} = \tilde{\mathbf{z}}(t_k)$ are a multinomial random vector taking values from the sample space

$$S = \left\{ \tilde{\mathbf{z}} \in \mathbb{Z}^4 : 0 \leq \tilde{z}_j \leq \tilde{y}_r, j = 1, \ldots, 4, \text{and} \sum_{j=1}^{4} \tilde{z}_j = \tilde{y}_r \right\}, \tag{4.4.8}$$

where $\tilde{y}_r (= h - \tilde{y}_s)$ is the number of unobserved individuals in the household of size h at a given point in time.

Therefore, the probability of observing the complete data $\tilde{\mathbf{x}} = (\tilde{\mathbf{z}}, \tilde{\mathbf{y}})$, given the observed data $\tilde{\mathbf{y}}$ and the parameters $\boldsymbol{\theta}$ is given by

$$\begin{aligned}
\pi(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}, \boldsymbol{\theta}) &\propto P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \boldsymbol{\theta})\pi(\tilde{\mathbf{x}}|\boldsymbol{\theta}) \\
&\propto P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})\pi(\tilde{\mathbf{x}}|\boldsymbol{\theta}) \quad (\textit{by conditional independence}) \\
&\propto \prod_{i=1}^{N} \prod_{k=2}^{n_i} \Big\{ P\big(\tilde{\mathbf{y}}_i(t_{ik})|\tilde{\mathbf{x}}_i(t_{ik})\big) \\
&\times \pi\big(\tilde{\mathbf{x}}_i(t_{ik}) = (\tilde{\mathbf{z}}_i(t_{ik}) + \tilde{\mathbf{y}}_i(t_{ik}))|\tilde{\mathbf{x}}_i(t_{ik-1}), \boldsymbol{\theta}\big) \\
&\times \pi\big(\tilde{\mathbf{x}}_i(t_{ik+1})|\tilde{\mathbf{x}}_i(t_{ik}) = (\tilde{\mathbf{z}}_i(t_{ik}) + \tilde{\mathbf{y}}_i(t_{ik})), \boldsymbol{\theta}\big) \Big\},
\end{aligned} \tag{4.4.9}$$

by the independent households assumption, and where the probability of observing the observed data given the complete data for household i at timepoint t_{ik}, $P\big(\tilde{\mathbf{y}}_i(t_{ik})|\tilde{\mathbf{x}}_i(t_{ik})\big)$

is given by

$$P\big(\tilde{\mathbf{y}}_i(t_{ik})|\tilde{\mathbf{x}}_i(t_{ik})\big) = \frac{\prod_{j=1}^{4}\binom{\tilde{x}_i j}{\tilde{y}_{ij}}}{\binom{\sum_{j=1}^{4} x_{ij}}{\sum_{j=1}^{4} y_{ij}}} \tag{4.4.10}$$

where $\tilde{x}_{ij} = \tilde{x}_{ij}(t_{ik})$ is the number of individuals in category j in household i at time point t_{ik}.

Independence Sampler

We utilize Independence Sampler algorithms to update $\pi(\tilde{\mathbf{z}}|\tilde{\mathbf{y}}, \boldsymbol{\theta})$ as follows. Whenever $\sum_{j=1}^{4} \tilde{y}_j(t_k) < h$, propose $\tilde{\mathbf{z}}^{prop}$ from (4.4.8), such that $\sum_{j=1}^{4} \tilde{z}_j^{prop} = \tilde{y}_r$. Then, set $\tilde{\mathbf{x}}_i^{prop}(t_{ik}) = \tilde{\mathbf{z}}_i^{prop}(t_{ik}) + \tilde{\mathbf{y}}_i(t_{ik})$. Accept the proposed value with probability

$$\alpha = min\left\{1, \frac{P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}^{prop})\pi(\tilde{\mathbf{x}}^{prop}|\boldsymbol{\theta})}{\tilde{P(\mathbf{y}}|\tilde{\mathbf{x}}^{curr})\pi(\tilde{\mathbf{x}}^{curr}|\boldsymbol{\theta})}\right\}. \tag{4.4.11}$$

SUMMARY

We shall now give a generic step by step summary of the MCMC updating schemes developed in this section and proceed as follows.

Step 0: Initialize $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_{1,2}, \beta_1, \beta_2, \beta_{1,2}, \gamma_1, \gamma_2, \gamma_{1,2}, \phi)$ and $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ (for IBD) or $\tilde{\mathbf{x}} = (\tilde{\mathbf{z}}, \tilde{\mathbf{y}})$ (for ABD). The starting values of $\boldsymbol{\theta}$ are chosen to be positive values since apart from ϕ, the rest are rates, and $\phi \geq 0$. For an individual-based data (IBD), initial values of the partially observed data $\mathbf{x}$ are chosen as described under Section 4.4.3, while the initial values of $\tilde{\mathbf{x}}$ are chosen as described under Section 4.4.4. Note that when data are completely observed, no initialization is required for the data as there is no need for data augmentation and interest is only on $\pi(\boldsymbol{\theta}|\mathbf{x})$ or $\pi(\boldsymbol{\theta}|\mathbf{x})$ as the case may be.

STEP 1: Update $\pi(\boldsymbol{\theta}|\mathbf{x}=(\mathbf{z},\mathbf{y}))$ using Random walk Metropolis and propose $\boldsymbol{\theta}'$ from a multivariate Gaussian proposal density centered at the current value of $\boldsymbol{\theta}$ and with a proposal covariance matrix Σ. Accept the proposed value with the probability

$$\alpha(\boldsymbol{\theta}^{curr},\boldsymbol{\theta}^{prop}) \leftarrow min\left\{1, \frac{\pi(\mathbf{x}|\boldsymbol{\theta}^{prop})\pi(\boldsymbol{\theta}^{prop})}{\pi(\mathbf{x}|\boldsymbol{\theta}^{curr})\pi(\boldsymbol{\theta}^{curr})}\right\}.$$

STEP 2: Update $\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$ for IBD via Independence Sampler steps and propose to switch states by setting $\mathbf{z}_{ij}^{prop}(t_{ik}) = 1 - \mathbf{z}_{ij}(t_{ik})$ (or $z_{ijl}^{prop} = 1 - z_{ijl}, l = 1,2$), exploring all the possibilities in $\{(0,0),(0,1),(1,0),(1,1)\}$, so that $\mathbf{x}_i^{prop}(t_{ik}) = \big(\mathbf{x}_i(t_{ik}), \mathbf{x}_{ij}(t_{ik}) = \mathbf{z}_{ij}^{prop}(t_{ik})\big)$. Accept the proposed values with probability

$$\alpha \;\leftarrow\; min\left\{1, \frac{\pi(\mathbf{x}_i^{prop}(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i^{prop}(t_{ik}),\boldsymbol{\theta})}{\pi(\mathbf{x}_i(t_{ik})|\mathbf{x}_i(t_{ik-1}),\boldsymbol{\theta})\pi(\mathbf{x}_i(t_{ik+1})|\mathbf{x}_i(t_{ik})),\boldsymbol{\theta})}\right\}.$$

STEP 3: Update $\pi(\tilde{\mathbf{z}}|\tilde{\mathbf{y}},\boldsymbol{\theta})$ for ABD via Independence Sampler steps and propose a new value for $\tilde{\mathbf{z}}^{prop}$ to be a multinomial random vector from the sample space define in (4.4.8), such that $\sum_{j=1}^{4}\tilde{z}_j^{prop} = \tilde{y}_r$, where $y_r = h - \sum_{j=1}^{4}\tilde{y}_j(t_k)$. Set $\tilde{\mathbf{x}}_i^{prop}(t_{ik}) = \tilde{\mathbf{z}}_i^{prop}(t_{ik}) + \tilde{\mathbf{y}}_i(t_{ik})$ and accept $\tilde{\mathbf{x}}_i^{prop}(t_{ik})$ with probability given by (4.4.11).

STEP 4: Repeat steps 1 and 2 (for IBD) or steps 1 and 3 (for ABD), until samples of the desired size are obtained.

4.5 Simulated Data Example

In this section we use simulated data sets to demonstrate the implementation of the MCMC algorithms developed in this chapter and compare the accuracy of the posterior estimates from the IBD-based and ABD-based MCMC algorithms.

4.5.1 Methodology

Using two different sets of *true* parameter values (*see*, Table 4.5.1), first we assume that data are completely observed and simulate 4 data sets, 2 (IBD and ABD) for each set of true parameter values,.

Parameter	SET 1	SET 2
λ_1	0.90	0.03
λ_2	0.80	0.05
λ_{12}	0.60	0.20
β_1	1.20	0.04
β_2	1.10	0.07
β_{12}	0.70	0.30
γ_1	0.50	0.03
γ_2	0.60	0.02
γ_{12}	0.40	0.10
ϕ	0.50	1.50

Table 4.5.1: True parameter value used for the simulation of the 16 data sets.

We simulate 4 data sets and allow 4 levels of missingness in the data, namely, 30%, 60% and 90%, for each set of true parameter values and for both IBD and ABD, whilst assuming data are missing completely at random (MCAR).

Each data set contains the same number of households, $N = 100$. The parameter values were chosen such that there is a low relative risk $\phi = 0.5$ for SET I and a high relative risk $\phi = 1.5$ for SET 2. Our reason for this choice of parameter values is to see how the performance of the MCMC algorithm is affected by various ranges of the parameter val-

ues. In all cases, we assumed that there is coinfection between the two diseases. We also assume that the coinfecting diseases interact and do not behave independently, hence $\phi \neq 1$. Throughout, we used same household structure and sizes, same time difference data with same number of observation time points.

Household sizes range from 1 to 4 with the majority of the households having between 2 and 4 individuals. The minimum (maximum) observation time difference is 1 (3), while the minimum (maximum) number of observation time points is 2 (5).

For $i = 1, 2, , N$, first we simulate the IBD data and proceed as follows. Choose $\mathbf{x}_i(t_{i0})$, the initial state of household i from the $\{0,1\}^{2\times h_i}$ possible states to which the household can belong to at any point in time. Then, for each set of parameters, we calculate the G-matrix and the corresponding Q-matrix according to (4.3.4) and $\mathbf{Q}_{\mathbf{t}}^{(h)} = \exp(t\mathbf{G}^{(h)})$, respectively. Then, sample $\mathbf{x}_i(t_{ik})$, the infectious status of household i at time t_{ik} from row $\mathbf{x}_i(t_{ik-1})$ of the transition probability matrix, $\mathbf{Q}_{\mathbf{t}}^{(h)}$. Record the the data each time for the n_i times the household was observed, where $n_i \in Z^+$ is chosen uniformly from $\{2,3,4,5\}$. This procedure was repeated for all $i = 1, 2, \ldots, 100$. After successful generation of the IBD data from the methods described above, the corresponding aggregate-based data (ABD) was obtained by setting

$$\tilde{x}_l = \sum_{j=1}^{h} u_{jl}, \tag{4.5.1}$$

where $\tilde{x}_l$ is the number of individuals in category l and u_{jl} is the infectious status of individual j for category l, for $l = 1, 2, 3, 4$. Recall that category 1 is no infection; category 2 is infected with disease 1; category 3 is infected with disease 2; and category 4 is infected with both diseases. Then the vector $\tilde{\mathbf{x}}_i(t_{ik}) = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4)$ is the ABD infectious state of household i at time t_{ik}.

4.5.2 MCMC Implementation

We assigned independent $Gamma(1,1)$ prior distribution to each of the parameters. For the completely observed data, it is straightforward to implement the Random walk Metropolis (RWM) algorithms with multivariate Gaussian proposal distribution described in Algorithm 4 to sample from $\pi(\boldsymbol{\theta}|\mathbf{x})$ for IBD or from $\pi(\boldsymbol{\theta}|\tilde{\mathbf{x}})$ for ABD, where

$$\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_{1,2}, \beta_1, \beta_2, \beta_{1,2}, \gamma_1, \gamma_2, \gamma_{1,2}, \phi).$$

We choose the proposal covariance matrix $\Sigma = \mathbf{I}\sigma_{ii}^2$, where $\sigma_{ii}^2 = 0.005$ for all $i = 1, 2, \ldots, 10$. This choice of proposal variance was found to yield acceptance rate close to the optimal value of 23.4% from 3 pilot runs with 1×10^3 iterations for each. Then we used the variance of the posterior distribution from the third pilot run as the proposal covariance matrix for the main 5×10^4 MCMC runs.

4.5.3 Results

The main convergence diagnostic tools used were traceplots and autocorrelation function (ACF) plots. We present the results obtained after discarding 1×10^4 iterations as burn-in as follows. Figure 4.5.1 shows the trace plots of the sojourn history of the chain for the completely observed IBD. The traceplot indicates that the Markov chain was mixing well.

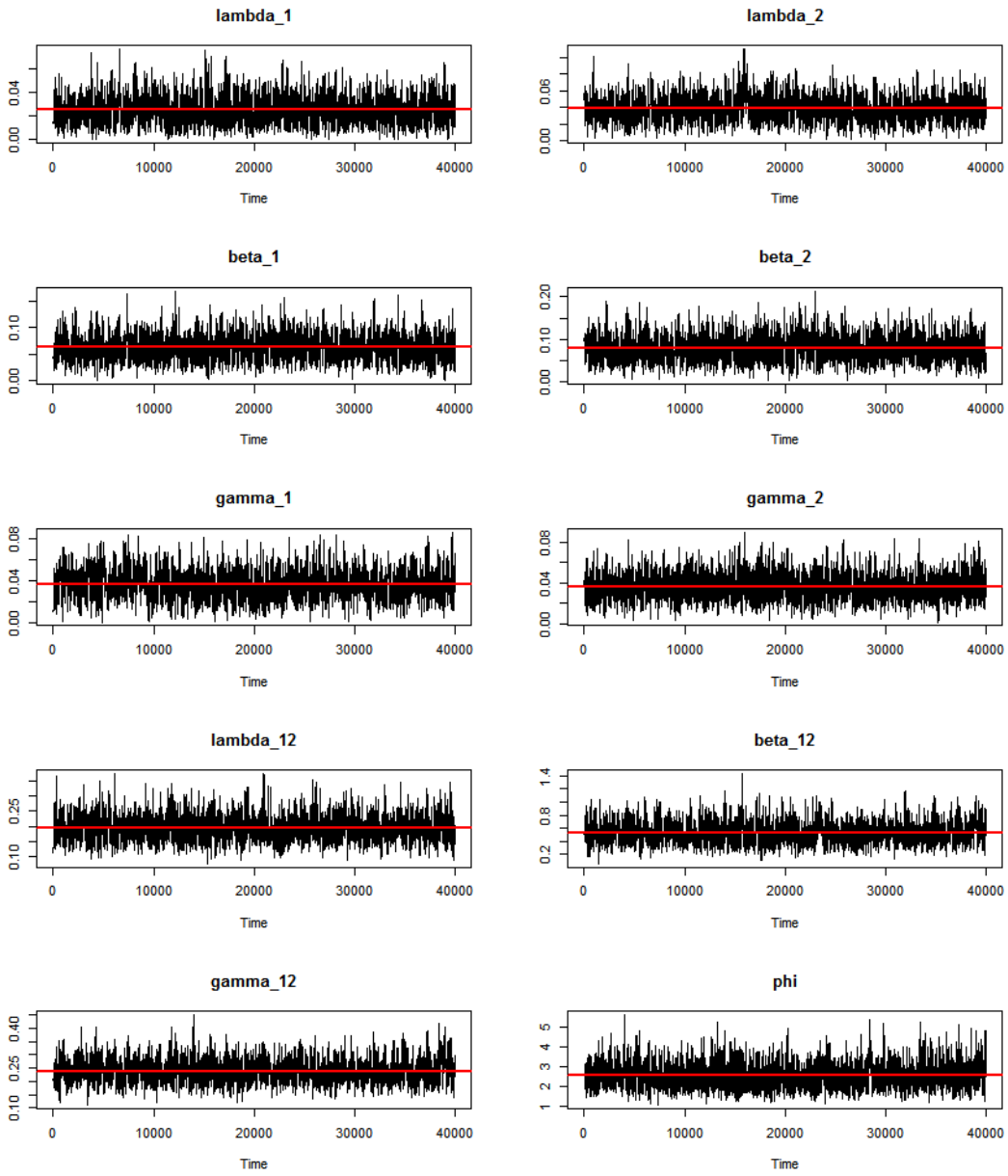

Figure 4.5.1: Traceplots of the completely observed IBD from SET 2 parameters for the last 4×10^4 iterations after a burn-in period of 1×10^4 iterations. The red lines are the corresponding posterior means of the parameters.

Table 4.5.2 compares the posterior means, standard deviations (SD) and the effective sample sizes (ESS) for the completely observed (or 0% missingness) co-epidemic data with reference to the individual-based data (IBD) and the aggregate-based (ABD) obtained from parameters in SET 2.

Parameters	**Mean**	**SD**	**ESS**
	IBD (ABD)	IBD (ABD)	IBD (ABD)
$\lambda_1 = 0.03$	0.025 (0.027)	0.011 (0.012)	1320(964)
$\lambda_2 = 0.05$	0.038 (0.042)	0.014 (0.016)	1110(1386)
$\beta_1 = 0.04$	0.062 (0.068)	0.023 (0.026)	1233(888)
$\beta_2 = 0.07$	0.078 (0.099)	0.028(0.034)	1428(988)
$\gamma_1 = 0.03$	0.036 (0.029)	0.013 (0.014)	942(1610)
$\gamma_2 = 0.02$	0.036 (0.036)	0.012 (0.014)	1725(930)
$\lambda_{12} = 0.20$	0.196 (0.200)	0.039 (0.042)	1219(1559)
$\beta_{12} = 0.30$	0.536 (0.646)	0.159 (0.191)	1057(1385)
$\gamma_{12} = 0.10$	0.238 (0.279)	0.043 (0.051)	1088(1491)
$\phi = 1.50$	2.555 (2.475)	0.593 (0.576)	1516(486)

Table 4.5.2: Posterior Means, Standard Deviations (SD) and Effective Sample sizes (ESS) for observed data for parameters **SET 2**, and for 0% **missing**

The results on the table show that both ABD- and IBD-based algorithms performed equally well when there are no missing data. We note that the ABD-based algorithm was approximately 2 times faster to run than the IBD-based algorithms. This is because the ABD contains fewer states than the IBD. Also, we explore the performance of the algorithms when data are only partially observed. Tables 4.5.3 and 4.5.4 compare the performance of the MCMC algorithms for parameters SET 1 for when missing data are 30% and 90%, respectively.

Parameters	**Mean**	**SD**	**ESS**
	IBD (ABD)	IBD (ABD)	IBD (ABD)
$\lambda_1 = 0.90$	1.093 (2.401)	0.381 (1.780)	632(54)
$\lambda_2 = 0.80$	0.582 (0.798)	0.187 (0.619)	942(440)
$\beta_1 = 1.20$	0.816 (1.387)	0.321 (0.886)	987(85)
$\beta_2 = 1.10$	0.631(1.083)	0.218 (0.603)	901(471)
$\gamma_1 = 0.50$	0.563 (1.635)	0.158 (1.181)	468(28)
$\gamma_2 = 0.60$	0.451 (0.788)	0.108(0.498)	660(316)
$\lambda_{12} = 0.60$	0.352(1.125)	0.291(0.938)	802(277)
$\beta_{12} = 0.70$	0.718 (0.876)	0.630 (0.860)	297(415)
$\gamma_{12} = 0.40$	0.150 (0.552)	0.139 (0.474)	639(502)
$\phi = 0.50$	0.494 (0.709)	0.084 (0.186)	931(129)

Table 4.5.3: Posterior Means, Standard Deviations (SD) and Effective Sample sizes (ESS) for observed data for parameters **SET 1**, and for 30% **missing**

Parameters	**Mean**	**SD**	**ESS**
	IBD (ABD)	IBD (ABD)	IBD (ABD)
$\lambda_1 = 0.90$	1.069 (0.123)	0.308 (0.304)	861(193)
$\lambda_2 = 0.80$	0.592 (0.118)	0.186 (0.229)	917(148)
$\beta_1 = 1.20$	0.843 (0.093)	0.314 (0.099)	798(406)
$\beta_2 = 1.10$	0.647(0.152)	0.226 (0.129)	837(452)
$\gamma_1 = 0.50$	0.544 (13.57)	0.123 (5.854)	1073(2)
$\gamma_2 = 0.60$	0.439 (15.72)	0.094(7.055)	1535(2)
$\lambda_{12} = 0.60$	0.307(61.79)	0.251(31.10)	1307(2)
$\beta_{12} = 0.70$	0.617 (0.555)	0.551 (0.734)	784(82)
$\gamma_{12} = 0.40$	0.139(0.518)	0.128 (0.504)	1161(274)
$\phi = 0.50$	0.479 (21.33)	0.081 (11.48)	758(2)

Table 4.5.4: Posterior Means, Standard Deviations (SD) and Effective Sample sizes (ESS) for observed data for parameters **SET 1**, and for 90% **missing**

The results show that IBD-based algorithms performed better in both cases, while the performance of the ABD-rapidly deteriorates. In both cases, it can be seen that the IBD-based algorithm is fairly stable.

As in SET 1, a similar result was obtained with SET 2, see, Tables 4.5.5 and 4.5.5. Throughout, as the percentage of missing data increases, the ABD-based algorithm performed very poorly

Parameters	**Mean**	**SD**	**ESS**
	IBD (ABD)	IBD (ABD)	IBD (ABD)
$\lambda_1 = 0.03$	0.025 (0.030)	0.011 (0.039)	1107(431)
$\lambda_2 = 0.05$	0.039 (0.046)	0.015 (0.093)	871(107)
$\beta_1 = 0.04$	0.063 (0.050)	0.023 (0.047)	960(199)
$\beta_2 = 0.07$	0.079 (0.064)	0.027(0.058)	1874(246)
$\gamma_1 = 0.03$	0.036 (2.646)	0.013 (0.920)	1364(4)
$\gamma_2 = 0.02$	0.037 (3.330)	0.013 (1.163)	1558(6)
$\lambda_{12} = 0.20$	0.196 (19.09)	0.039 (10.74)	1506(2)
$\beta_{12} = 0.30$	0.527 (5.219)	0.154 (3.419)	1294(2)
$\gamma_{12} = 0.10$	0.239 (0.137)	0.043 (0.123)	1188(135)
$\phi = 1.50$	2.563 (12.76)	0.626 (2)	801(486)

Table 4.5.5: Posterior Means, Standard Deviations (SD) and Effective Sample sizes (ESS) for observed data for parameters **SET 2**, and for 60% **missing**

Parameters	**Mean**	**SD**	**ESS**
	IBD (ABD)	IBD (ABD)	IBD (ABD)
$\lambda_1 = 0.03$	0.026 (1.863)	0.012 (1.456)	779(268)
$\lambda_2 = 0.05$	0.039 (3.194)	0.015 (4.379)	1954(17)
$\beta_1 = 0.04$	0.063 (0.248)	0.023 (0.243)	1246(383)
$\beta_2 = 0.07$	0.082 (1.016)	0.030(1.244)	743(29)
$\gamma_1 = 0.03$	0.036 (1.611)	0.013 (0.738)	1479(214)
$\gamma_2 = 0.02$	0.037 (5.975)	0.012 (11.183)	1334(7)
$\lambda_{12} = 0.20$	0.198 (2.203)	0.040 (1.767)	1109(251)
$\beta_{12} = 0.30$	0.538 (0.720)	0.164 (0.633)	1082(510)
$\gamma_{12} = 0.10$	0.240 (0.661)	0.045 (0.652)	1327(524)
$\phi = 1.50$	2.545 (0.962)	0.595 (0.506)	1272(52

Table 4.5.6: Posterior Means, Standard Deviations (SD) and Effective Sample sizes (ESS) for observed data for parameters **SET 2**, and for 90% **missing**

4.6 Application to the Tanzania Cattle Data

In this section we apply the models developed in this chapter in a real life situation and demonstrate the implementation of our MCMC algorithms to a rich set of data on tick-borne diseases among Tanzania cattle.

4.6.1 Data and Methods

The data contains information on the spread of 5 coinfecting tick-borne diseases (*T.parva, T.mutans, A.marginale, B.bigemina, B.bovis*) among Tanzania cattle. The longitudinal

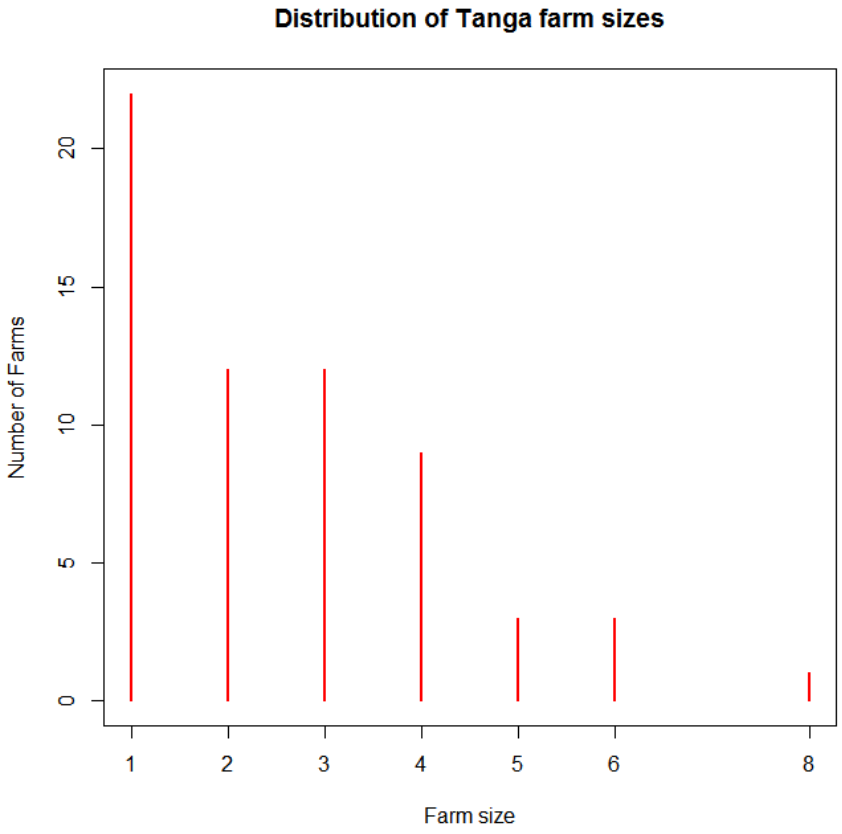

Figure 4.6.1: Distribution of the farms in Tanga town according to their sizes.

data was collected over 11 observation timepoints. The data were collected from four(4) regions of Tanga, Mtindi, Korogwe and Kibaya. For our purposes, we shall focus on the coinfection among the farms in Tanga.

There are 62 farms in Tanga with farm sizes range of 1 to 8. Figure 4.6.1 shows the number of farms N_n which have exactly n animals in the farm.

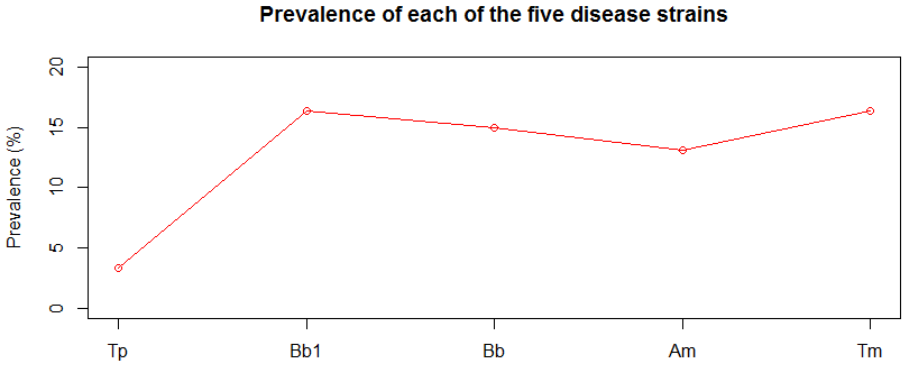

Figure 4.6.2: Observed Prevalence plot for the five strains of ticks from the Tanzanian data.

4.6.2 MCMC Implementation and Results

We assign independent $Gamma(1,1)$ described in Algorithm 4 to sample from $\pi(\boldsymbol{\theta}|\mathbf{x})$ for IBD or from $\pi(\boldsymbol{\theta}|\tilde{\mathbf{x}})$ for ABD. Convergence diagnostics used are mainly traceplots and ACF plots. Figure 4.6.3 shows a traceplot from individual-based data for coinfection of *T.mutans* vs *B.bovis*, and is obtained from 1×10^4 iterations after a burn-in period of 2×10^3. The traceplot shows that the MCMC algorithms were mixing well.

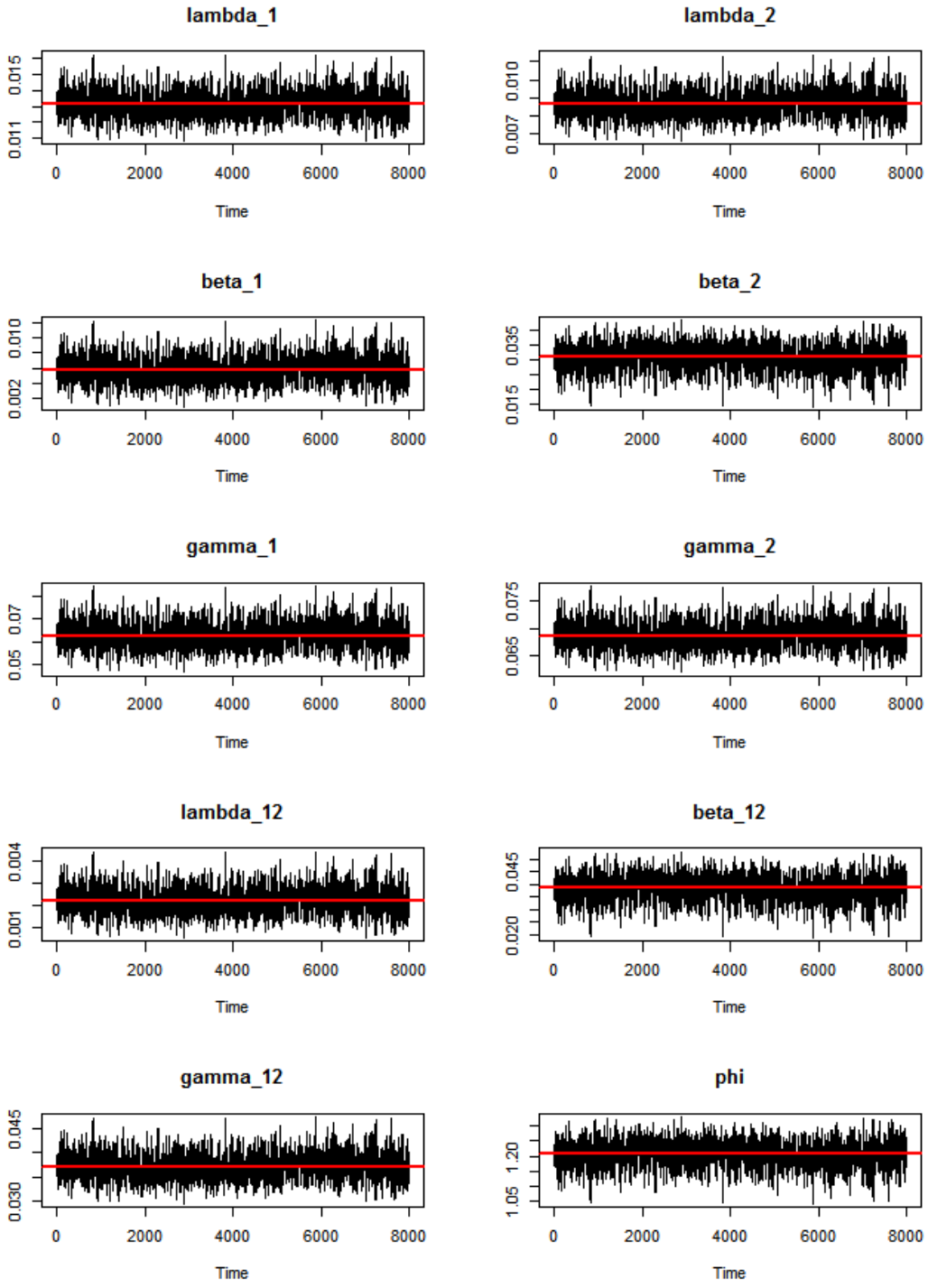

Figure 4.6.3: **Real life application:** Traceplots of posterior distribution of coinfection of *T.mutans* vs *B.bovis* using individual-based data (IBD), obtained from 1×10^4 iterations after a burn-in period of 2×10^3 iterations. The red lines are the posterior means of the corresponding parameters.

4.6.3 Results

In this section, we present the results from the real-life application of our approach. Table 4.6.3 shows the posterior estimates (means and standard deviations) of the relative risks for the 10 possible pairwise combination of the diseases.

Serotype	***Tm***	***Am***	***Bb***	***Bb1***
Tp	0.847 (0.051)	1.126 (0.020)	1.301 (0.147)	1.219 (0.0058)
Tm		1.868 (0.273)	1.519 (0.149)	1.208 (0.0432)
Am			1.271 (0.111)	0.873 (0.1300)
Bb				1.051 (0.0100)

Table 4.6.1: Posterior Relative risk (ϕ) **mean** (**Standard deviation**) obtained from the individual-based data (IBD) of the Tanzania tick-borne diseases for the 10 ten possible combinations of the disease pairs, and from 5×10^4 iterations after a burn-in period of 1×10^4 iterations. ***Tp*** = *T.parva*; ***Tm*** = *T.mutans*; ***Am*** = *A.marginale*; ***Bb*** = *B.bigemina*; ***Bb1*** = *B.bovis*.

The results show that infection with *T.parva*, for example, will lead to a reduced risk of getting infected with *T.mutans* with $\hat{\phi} = 0.847 < 1$, while it also leads to an increased risk of the animal getting infected with *B.bigemina* or with *B.bovis* ($\hat{\phi} > 1$). Observe that *B.bigemina* and *B.bovis* will evolve independently with $\hat{\phi} \approx 1$ (see also, Figure 4.6.4).

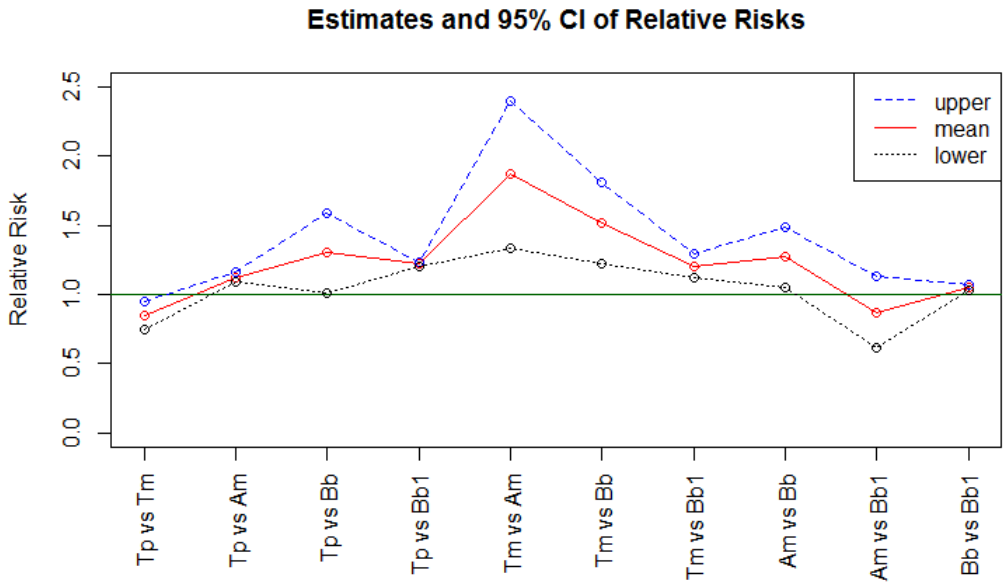

Figure 4.6.4: **Real life application:** Relative Risks estimates plots from the individual-based data (IBD) for the 10 possible pairwise disease combinations with 95% Credible Interval. Tp = *T.parva*; Tm = *T.mutans*; Am = *A.marginale*; Bb = *B.bigemina*; Bb1 = *B.bovis*.

4.7 Discussions

In this chapter, we studied the transmission dynamics of SIS epidemics with coinfection in a households setting. Throughout, we considered $d = 2$ diseases, but it is fairly straightforward to extend this to cases involving $d > 2$ diseases. This will involve making additional assumptions and estimating more parameters.

To validate our model and the MCMC algorithms developed here, we first used a simulated data set, with $d = 2$ and then applied it to a real life situation via the Tanzania data.

With the simulated data set, we were able to demonstrate the implementation of MCMC algorithms. The posterior estimates obtained with both IBD- and ABD- were close to the true parameter values, when data are fully observed. The ABD has fewer states

transitions and this makes it easier to work with. However, as the proportion of missing values increases, the MCMC algorithms based upon ABD rapidly deteriorates, while the IBD is robust all through. In addition, our models and MCMC algorithms were successfully applied to a real-life data on tick-borne diseases among Tanzania cattle. Given that the Tanzania tick-borne disease data allows varying population sizes over time, it was straightforward to extend our model to allow for varying population sizes over time by following the methods developed in Chapter 3.

A major challenge encountered in this chapter is in the calculation of the transition rate matrix. For example, for the individual-based data, for a farm size of $h = 8$ the G-matrix is a 65536×65536 ($2^{d\times h} \times 2^{d\times h}$) matrix of transition rates. This places a huge burden on the computer memory. However, this problem can be circumvented using computer clusters which allows easy computation.

A key contribution of this chapter is the development of Bayesian inference approaches for the analysis of SIS coepidemic model implemented via MCMC framework. The method developed here can easily be applied to any number of diseases and also work well when population is constant or varying.

Chapter 5

Conclusions and Future Works

In this chapter, we present a chapter by chapter summary of the major problems tackled in this thesis, highlighting some salient points and limitations. We also suggest some follow up studies in order to surmount some identified challenges.

5.1 Closed Population SIS Household Model

The main work in Chapter 2 of this thesis develops Bayesian inference approach for the estimation of parameters of stochastic household-based SIS epidemic models implemented in Markov chain Monte Carlo (MCMC) framework. Two most prevalent SIS household epidemic data were considered throughout, namely, the individual-based data (IBD) and the aggregate-based data (ABD). The IBD is more informative, but is more complex to handle unlike the less informative ABD. An extensive simulation study carried out shows that the MCMC algorithms developed with respect to both data types worked well especially when the data are complete. However, when the data are allowed to be only partially observed, interesting behaviors of the algorithms were observed. The overarching aim for allowing incomplete data case is to completely mimic what actually happens in reality as most infectious disease data are rarely complete. Given that dif-

ferent methods of data collection would give rise to different forms of missingness, we considered three different forms of missing SIS household-based epidemic data. Firstly, we considered a case where data (individuals' infection statuses) are allowed to miss randomly at various time points. Secondly, we considered a missing data due to a randomly selected individual missing completely in all the observation time points. Thirdly, we considered a missing data due to a randomly selected observation time point missing completely. Following efficient data imputation strategies, robust MCMC algorithms were developed and data successfully analyzed. Two different approaches were adopted for the third form of missing data. First, we deleted each unobserved time point and then treated the rest of the data as completely observed. In this case, there was no need for data augmentation. The second approach was to impute the missing time point with its observation. It was found that the former outperformed the latter. The most robust missing data form was found to be the first type which assumes that individuals miss randomly at observation time points. On the other hand, the second data form which assumes that randomly selected individuals miss completely at all time points, has the worst performance. In general, the IBD-based MCMC shows better performance than the ABD as the proportion of missing data increases.

One major problem encountered in this chapter is computational. Given that there are 2^h possible states for a household of size h, computational burden for the calculation of the transition probability matrix (Q-matrix) grows as $h\infty$. For example, when $h = 10$, the Q-matrix contains contains 1048576 entries and this places a huge demand on the computer memory. Parallel computing provides a relief for moderate h. In most practical situations, especially among animals population, it is possible to have a single farm with much lager farm size, say 20, and the applicability of our methods may suffer a huge setback. It would be interesting to develop more efficient methods for computing such high-dimensional matrices. Although the data imputation approaches used in this

chapter appeared to work really well, there could be several other better ways to do this and obtain even much better results. It would be interesting to explore this further. Finally, the model studied in Chapter 2 could be extended to include demography such as births and deaths.

5.2 Open population, Spatial SIS

Chapter 3 was divided into two main parts. The first part develops inference methods using Bayesian paradigm for open population stochastic household SIS epidemics. The second part incorporates spatial element into the modelling. In both cases, we adopted the framework developed in Chapter 2 where we assume that individuals are contacted locally and and there also exists a global force of infection. Unlike the closed population epidemics, household size of the open population model varies over time and this created computational complexity as the Q-matrix needs to be calculated at each observation time point whenever the household size changes. On the other hand, the spatial SIS model assumes that the global force of infection depends on some Gaussian random fields realizations. Both models were found to work well with real life data, but the spatial model performed comparatively poorly than the non-spatial open population model with simulated datasets. However, we note that a further investigation is required to further optimize the spatial model algorithms, and possibly extend the models to allow for the inclusion of demographic parameters such as birth and death rates.

5.3 Coinfection

Chapter 4 of this thesis contains a study on the infection of a host with multiple pathogens or by multiple strains of a given pathogen. Understanding transmission dynamics of coinfection is an important factor towards defining a stable control approach to

its spread. The Bayesian inference methods via MCMC developed in 4 was successfully applied to both simulated and real life datasets. This could also be extendnd to allow for demographic elements. Note that we have considered only pairwise interactions. Therefore, it would be interesting to extend this to cases with multiple coinfection, *i.e.*, where $d \geq 3$ pathogens or strains of pathogen can infect either simultaneously or singly. This would be a direct extension of the model in Chapter 4, with more parameters included into the modelling. Again, the problem of computational burden due to the calculation of the Q-matrix needs to be addressed first before a substantial progress could be made.

Bibliography

Addy, C., I. Longini, and M. Haber (1991). A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics 47*, 961–974.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*. New York second edition: Hafner Press [Macmillan Publishing Co Inc].

Ball, F. (1999). Stochastic and deterministic models for sis epidemics among a population partitioned into households. *Mathematical Biosciences 156*, 41–67.

Ball, F., T. Britton, T. House, V. Isham, D. Mollison, L. Pellis, and G. Scalia-Tomba (2015). Seven challenges for metapopulation models of epidemics, including household models. *Epidemics 10*, 63–67.

Ball, F. and O. Lyne (2001). Stochastic multitype sir epidemic among a population partitioned into households. *Advances in Applied Probability 33*, 99–123.

Ball, F., D. Mollison, and G. Scalia-Tomba (1997). Epidemics with two levels of mixing. *Annals of Applied Probability 7*, 46–89.

Ball, F. and P. Neal (2004). Poisson approximations for epidemics with two levels mixing. *The Annals of Probability 32*(1), 1168–1200.

Balmer, O. and M. Tanner (2011). Prevalence and implications of multiple-strain infections. *The Lancet Infectious Diseases 11*(11), 868–878.

Barry, J. M., C. Vibond, and L. Simonsen (2008). Cross-protection between successive waves of the 1918–1919 infleunza pandemic: epidemiological evidence from us army camps and from britain. *Journal of Infectious Diseases 198*(10), 1427–1434.

Bartlett, M. S. (1949). Some evolutionary stochastic processes. *J. Roy. Statist. Soc. Ber. B 11.*

Becker, N. and K. Dietz (1995). The effect of the household distribution on transmission and control of highly infectious diseases. *Math. Biosci. 127*, 207–219.

Becker, N. G. (1993). Parametric inference for epidemic models. *Math. Biosci 117*, 239–259.

Bernardo, J.-M. and A. F. M. Smith (1994). *Bayesian theory.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Chichester: John Wiley & Sons Ltd.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. In M. de Mathématique et de Physique de L'Académie Royale des Sciences Paris (Ed.), *Histoire de L'Académie Royale des Sciences (1766)*, pp. 1–45.

Besag, J. and P. Green (1993). Spatial statistics and bayesian computation (disc: P53-102). *Journal of the Royal Statistical Society, Series B: Statistical Methodology 55*, 25–37.

Blake, I., M. Burton, R. Bailey, and A. e. a. Solomon (2009). Estimating household and community transmission of ocular *Chlamydia trachomatis. PLoS Negl Trop Dis 3*(3).

Britton, T. (2009). Stochastic epidemic models: a survey arxiv:0910.4443v1[math.pr].

Britton, T. and P. Neal (2010). The time to extinction for an sis-household-epidemic model. *Journal of Mathematical Biology 61*(6).

Britton, T. and P. D. O'Neill (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scand J Statist 29*(3), 375–390.

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman and Hall/CRC.

Cauchemez, S., F. Carrat, C. Viboud, A. Valleron, and P. Bo:elle (2004). A bayesian mcmc approach to study transmission of influenza:application to household longitudinal data. *Statistics in Medicine 23*, 3469–3487.

Clancy, D., P. D. O'Neill, and P. Pollet (2001). Approximations for the long-term behavior of an open population epidemic model. *Methodology and Computing in Applied Probability 3*(1), 75–95.

Cressie, N. (1993). *Statistics for Spatial Data*. New York, 2nd: John Wiley & sons.

Daley, D. and J. Gani (1999). *Epidemic Modelling: An Introduction*. Cambridge: Cambridge University Pressl.

Deardon, R., S. P. Brooks, B. Grenfell, M. J. Keeling, M. J. Tildesley, N. J. Savill, D. Shaw, and M. E. J. Woolhouse (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica 20*, 239–261.

Demiris, N. and P. O'Neill (2005a). Bayesian inference for epidemics with two levels of mixing. *Scandinavian Journal of Statistics 32*, 265–280.

Demiris, N. and P. O'Neill (2005b). Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal of the Royal Statistical Society, Series B 67*, 731–746.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B 39*(1), 1–38.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998). Model-based geostatistics (disc: P326-350). *Journal of the Royal Statistical Society, Series C: Applied Statistics 47*, 299–326.

Economou, A., A. Gómez-Corral, and M. López-García (2015). A stochastic sis epidemic model with heterogeneous contacts. *Physica A 421*, 78–97.

Epstein, S. L. (2006). Prior h1n1 influenza infection and susceptibility of cleveland family study participants during the h2n2 pandemic of 1957: an experiment of nature. *Journal of Infectious Diseases 193*(1), 49–53.

Ferguson, N., R. Anderson, and S. Gupta (1999). The effect of anti-boddy enhancement on the transmission dynamics and persistence of multi-strain pathogens. *Proceedings of the National A cademy of Sciences of the United States of America 96*(2), 790–794.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo, 2nd ed.* Boca Raton, FL: Chapman and Hall/CRC.

Gao, D., T. Porco, and S. Ruan (2016). Coinfection dynamics of two diseases in a single host population. *J. Math. Annal. Appl. 442*, 171–188.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc. 85*(410), 398–409.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Tr. Pat. An. Mach. Int 6*, 721–741.

Getahun, H., C. Gunneberg, R. Granich, and P. Nunni (2010). Hiv infection–associated tuberculosis:the epidemiology and the response. *Clin Infect Dis (Suppl 3) 50*, 201–207.

Gibson, G. (1997). Markov chain monte carlo methods for ftting spatiotemporal stochastic models in plant epidemiology. *Applied Statistics 46*(2), 215–233.

Gibson, G. and E. Renshaw (1998). Estimating parameters in stochastic compartmental models using markov chain methods. *IMA J. Math. Appl. Med. Biol 15*, 19–40.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov chain Monte Carlo in practice. Interdisciplinary Statistics*. London: Chapman Hal.

Goldstein, E., K. Paur, C. Fraser, E. Kenah, J. Wallinga, and M. Lipsitch (2009). Reproductive numbers, epidemic spread and control in a community of households. *Mathematical Biosciences 221*, 11–25.

Greenhalgh, Y. Liang, and X. l. Mao (2016). Sde sis epidemic model with demographic stochasticity and varying population size. *Applied Mathematics and Computation 276*, 218–238.

Greenwood, P. and L. Gordillo (2009). Stochastic epidemic modeling. In G. Chowell, J. M. Hyman, L. M. A. Bettencourt, and C. Castillo-Chavez (Eds.), *Mathematical and Statistical Estimation Approaches in Epidemiology*, pp. 31–52. Springer, Dordrechtr.

Grimmett, G. and D. Stirzaker (2001). *Probability and Random Processes Third Edition*. Oxford: Oxford University Press.

Haario, H., E. Saksman, and J. Tamminem (1999). Adaptive proposal distribution for random walk metropolis algorithm. *Comput Stat 14*, 375–395.

Haario, H., E. Saksman, and J. Tamminem (2005). Componentwise adaptation for high dimensional mcmc. *Comput Stat 20*, 265–274.

Hamer, W. H. (1906). Epidemic diseases in england. *The Lancet 1*, 733–739.

Haran, M. (2011). Gaussian random field models for spatial data. In S. Brooks, A. Gelman, G. Jones, and X. Meng (Eds.), *Handbook of Markov chain Monte Carlo*. Boca Raton, Florida, USA: Chapman and Hall/CRC.

Hastings, W. (1970). Monte carlo sampling using markov chains and their applications. *Biometrika 57*, 97–109.

Hersh, M. H., R. S. Ostfeld, D. J. McHenry, M. Tibbetts, J. L. Brunner, M. E. Killilea, K. LoGuidice, K. A. Schmidt, and F. Keesing (2014). Co-infection of blacklegged ticks with *Babesia microti* and *Borrelia burgdorferi* is higher than expected and acquired from small mammals hosts. *Plos ONE 9*(6).

Hethcote, H. W. (1976). Qualitative analyses of communicable disease models. *Math Biosci 28*, 335–356.

Hethcote, W. and P. van den Driessche (1995). An sis epidemic model with variable population size and delay. *Journal of Mathematical Biology 34*, 177–194.

Hoti, F., P. Erasto, T. Leino, and K. Auranen (2009). Outbreaks of *Streptococcus pneumoniae* in day care cohorts in finland- implication for elimination and transmission. *BMC Infectious Diseases 9*(102).

House, T. and M. J. Keeling (2008). Deterministic epidemic models with explicit household structure. *Mathematical Biosciences 213*, 29–39.

Isham, V. (2005). Stochastic models for epidemics. In O. S. S. Series (Ed.), *Papers in Honour of Sir David Cox on His 80th Birthday*, pp. 1–31. Oxford University Press.

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

Keeling, M. J., M. E. J. Woolhouse, R. M. May, G. Davies, and B. T. Grenfell (2003). Modelling vaccination strategies against foot-and-mouth disease. *Nature 421*, 136–142.

Keeling, M. J., M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell (2001). Dynamics of the 2001 uk foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science 294*(5543), 813–818.

Kermack, W. O. and A. G. McKendrick (1927). Contributions to mathematical theories of epidemics part i. *Proceedings of the royal society of London A115*, 700–721.

Kypraios, T. (2007). Efficient bayesian inference for partially observed stochastic epidemics and a new class of semiparametric time series models. ph.d. thesis, department of mathematics and statistics, lancaster university, lancaster.

Lipsitch, M. (1997). Vaccination against colonizing bacteria with multiple serotypes. *Proceedings of the National Academy of Scineces 94*(12), 6571–6576.

Longini, I. and J. S. Koopman (1982). Household and community transmission parameters from final distribution of infections in households. *Biometrics 38*(1), 115–126.

Longini, I. M., A. Nizam, S. Xu, K. Ungchusak, and W. Hanshaoworakul (2005). Containing pandemic at the source. *Science 309*, 1083–1087.

Lou, Y., L. Liu, and D. Gao (2017). Modelling co-infection os *IXODES* tick-borne pathogens. *Mathematical Biosciences and Engineering 14*(4 & 5), 1301–1316.

Marion, G., G. Gibson, and E. Renshaw (2003). Estimating likelihoods for spatio-temporal models using importance sampling. *Statistics and Computing 13*.

McKendrick, A. G. (1926). Applications of mathematics to medical problems. *Proceedings of Edinburgh Mathematical Society 44*, 98–130.

Metropolis, N., A. W. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys. 21*, 187–191.

Moutailler, S., C. V. Moro, E. Vaumourin, L. Michelet, F. H. Tran, E. Devillers, J.-F. Cosson, P. Gasqui, V. T. Van, P. Mavingui, G. Vaourc'h, and M. Vayssier-Taussat (2016). Co-infection of ticks: The rule rather than the exception. *PLOS Neglected Tropical Diseases 10*(3).

Neal, P. (2006). Stochastic and deterministic analysis of sis household epidemics. *Adv. Appl. Prob. 38*, 943–968.

Neal, P. (2012). Efficient likelihood-free bayesian computation for household epidemics. *Stat Comput 22*, 1239–1256.

Neal, P. (2014). Endemic behaviour of sis epidemics with general infectious period distribution. *Adv. Appl. Prop 46*, 241–255.

Neal, P. and T. C. L. Huang (2015). Forward simulation markov chain monte carlo with applications to stochastic epidemic models. *Scandinavian Journal of Statistics 42*(2), 378–396.

Neal, P. and T. Kypraios (2015). Exact bayesian inference via data augmentation. *Stat Comput 25*, 333–347.

Neal, P. and G. Roberts (2005). A case study in non-centering for data augmentation: stochastic epidemics. *Stat. Comput 15*(4), 315–327.

Neal, P. and G. Roberts (2006). Optimal scaling for partially updating mcmc algorithm. *Ann. Appl. Probab. 16*, 475–515.

Neal, P. and F. Xiang (2017). Collapsing of non-centered parameterized mcmc algorithms with applications to epidemic models. *Scandinavian Journal of Statistics 44*, 81–96.

Newman, M. E. J. and C. R. Ferrario (2013). Interacting epidemics and coinfection on contact networks. *PloS ONE 8*(8).

O'Neill, P. D. (1996). Strong approximations for some open population epidemic models. *Journal of Applied Probability 33*(2), 448–457.

O'Neill, P. D. (2009). Bayesian inference for stochastic multitype epidemics in structured populations using sample data. *Biostatistics 10*(4), 779–791.

O'Neill, P. D., D. Balding, N. G. Becker, M. Eerola, and D. Mollison (2000). Analysis of infectious disease data from household outbreaks by markov chain monte carlo methods. *Journal of the Royal Statistical Society 49*(4), 517–542.

O'Neill, P. D. and N. G. Becker (2001). Inference for epidemic when susceptibility varies. *Biostatistics 2*(1), 99–108.

O'Neill, P. D. and G. O. Roberts (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A 162*.

Robert, C. P. and G. Casella (1999). *Monte Carlo statistical methods*. Springer-Verlag, Newyork: Monte Carlo statistical methods. Springer Texts in Statistics. Springer-Verla.

Roberts, G., A. Gelman, and W. Gilks (1997). Weak covergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Prob. 7*, 110–120.

Roberts, G. and J. Rosenthal (2001). Optimal scaling for various metropolishastings algorithms. *Statistical Science 16*, 351–367.

Roberts, G. and J. Rosenthal (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *J. Appl. Probab. 44*, 458–475.

Roberts, G. O. and J. Rosenthal (2010). Examples of adaptive mcmc. *J Comp Graph Stat 8*, 349–367.

Ross, R. (1911). *The Prevention of Malaria 2nd edition*. Murray: London.

Ross, R. (1915). Some apriori pathometric equations. *British Med J 1*, 546–546.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. In W. Series (Ed.), *Probability and Mathematical Statistics:Applied Probability and Statistics*. New York: John Wiley & Sons Inc.

Savill, N. J., D. J. Shaw, R. Deardon, M. J. Tildesley, M. J. Keeling, M. E. Woolhouse, S. P. Brooks, and B. T. Grenfell (2006). Topographic determinants of foot and mouth disease transmission in the uk 2001 epidemic. *BMC Vet Res 2*.

Sharp, G. B., Y. Kawaoka, D. J. Jones, W. J. Bean, S. P. Pryor, V. Hinshaw, and R. G. Webster (1997). Coinfection of wild ducks by influenza a viruses: Distribution patterns and biological significance. *Journal of Virology 71*, 6128–6135.

Sherlock, C., P. Fearnhead, and G. O. R. (2010). The random walk metropolis: Linking theory and practice through a case study. *Statistical Science 25*(2).

Slater, H. C., M. Gambhir, P. E. Parham, and E. Micheal (2013). Modelling co-infection with malaria and lymphatic filariasis. *PloS Comput Bio 9*(6).

Streftaris, G. and G. J. Gibson (2004). Bayesian inference for stochastic epidemics in closed populations. *Statiscall Modellling 4*, 63–75.

Swendsen, R. H. and S. Wang (1987). Nonuniversal critical dynamics in monte carlo simulations. *Physical Review Letters 58*, 86–88.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc. 82*(398), 528–550.

Utazi, E. C., E. O. Afuecheta, and C. C. Nnanatu (2018). A bayesian latent process spatiotemporal regression model for areal count data. *Spatial and Spatio-temporal Epidemiology 25*, 25–37.

Xiang, F. and P. Neal (2014). Efficient mcmc for temporal epidemics via parameter reduction. *Computational Statistics and Data Analysis 80*, 240–250.