Efficient Bayesian Inference for Partially Observed Stochastic Epidemics

and

A New Class of Semi–Parametric Time Series Models

Theodore Kypraios, MSc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University,

June 2007.

Efficient Bayesian Inference for Partially Observed Stochastic Epidemics and A New class of Semi−Parametric Time Series Models

Theodore Kypraios, MSc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University,

June 2007.

Abstract

This thesis is divided in two distinct parts. In the first part we are concerned with developing new statistical methodology for drawing Bayesian inference for partially observed stochastic epidemic models. In the second part, we develop a novel methodology for constructing a wide class of semi−parametric time series models.

First, we introduce a general framework for the heterogeneously mixing stochastic epidemic models (HMSE) and we also review some of the existing methods of statistical inference for epidemic models. The performance of a variety of centered Markov Chain Monte Carlo (MCMC) algorithms is studied. It is found that as the number of infected individuals increases, then the performance of these algorithms deteriorates. We then develop a variety of centered, non−centered and partially non−centered reparameterisations. We show that partially non−centered reparameterisations often offer more efficient MCMC algorithms than the centered ones.

The methodology developed for drawing efficiently Bayesian inference for HMSE is then applied to the 2001 UK Foot-and-Mouth disease outbreak in Cumbria. Unlike other existing modelling approaches, we model stochastically the infectious period of each farm assuming that the infection date of each farm is typically

unknown. Due to the high dimensionality of the problem, standard MCMC algorithms are inefficient. Therefore, a partially non−centered algorithm is applied for the purpose of obtaining reliable estimates for the model's parameter of interest. In addition, we discuss similarities and differences of our findings in comparison to other results in the literature.

The main purpose of the second part of this thesis, is to develop a novel class of semi−parametric time series models. We are interested in constructing models for which we can specify in advance the marginal distribution of the observations and then build the dependence structure of the observations around them. First, we review current work concerning modelling time series with fixed non−Gaussian margins and various correlation structures. Then, we introduce a stochastic process which we term a latent branching tree (LBT). The LBT enables us to allow for a rich variety of correlation structures. Apart from discussing in detail the tree's properties, we also show how Bayesian inference can be carried out via MCMC methods. Various MCMC strategies are discussed including non−centered parameterisations. It is found that non−centered algorithms significantly improve the mixing of some of the algorithms based on centered reparameterisations. Finally, we present an application of this class of models to a real dataset on genome scheme data.

Acknowledgments

I feel extremely lucky that I met Petros Dellaportas during my undergraduate studies in Athens University of Economic and Business. Apart from being a great teacher, Petros was the first person who advised me to do a PhD and in particular to come to the UK and be supervised by Gareth Roberts. I will always be grateful to Petros for this suggestion and his valuable guidance all these years.

Working with Gareth has been really exciting. Apart from his energy and his enthusiasm for research, his support and guidance throughout my PhD have been invaluable. I very much enjoyed discussions on the intuition behind the work in this thesis and many other subjects. I am grateful to Gareth for trying hard and succeeding in a finding me a scholarship without which I would not be able to study for a PhD. Gareth has been encouraging and very patient while I was writing up this thesis and provided me with very constructive comments. After 4 years now, I believe that I made a very good friend. For all these and many more I would like to thank him.

It has been a great pleasure that I had the chance to study in Lancaster where I had the chance to exchange ideas and collaborate with many interesting people. In particular, I would like to thank Paul Fearnhead for his support and guidance through my PhD. I am also grateful to him for many interesting discussions and for providing me the genome scheme data used in the second part of this thesis.

Furthermore, I feel very lucky that coming to the UK, the first person I met was Omiros Papaspiliopoulos. Apart from being a very good friend, Omiros has

been very supportive and his help during my PhD has been very important. I would also like to thank my friends Alexandros Beskos, Kostas Kalogeropoulos and Nikos Demiris for many fruitful discussions. Pete Neal and Chris Jewell have been exciting collaborators and I am very happy that I had the chance to work with them.

In addition, many other people made my time in Lancaster enjoyable. I would particularly like to mention my officemates Jamie Kirkham, Chris Sherlock, Mark Latham and Rosemeire Fiaccone for suffering silently when half of the Greek community in Lancaster was visiting me in the office. Furthermore, I would like to mention the friends of mine who I played music with (Dimitris, Kostas, Kostas and Manolis) and those who have suffered listening to us (too many to mentioned here). Also, a big thank to my flatmates throughout the years I spent on campus: Alexandra, Vanessa, Dina, Nik and George - I will never forget the moments we shared together. Although there are people out there who sharing the time with has been very rewarding, there are too many to be mentioned here. My special thanks to Anastasia Lykou for printing for me the final version of this thesis!

I would also like to thank Omiros Papaspiliopoulos, Paul Fearnhead, Simon Preston, Pete Neal and Chris Jewell for useful comments on earlier versions of this thesis. I would also like to thank Thomas Miles in DEFRA who provided us for the data on 2001 UK Foot-and-Mouth disease outbreak.

Last but not least, my family was extremely supportive throughout this period, despite me being abroad - many thanks to them. Since my my very best friends from Rhodes (Thanasis, Leandros, Mixalis and Stefanos) also considered as 'family', they also deserve to be thanked for their patience and support for all the years we know each other.

Finally, I would like to thank Lancaster University for its financial support during my PhD.

Declaration

I hereby declare that this thesis is my own work, except otherwise stated, and has not been submitted in substantially the same form for the award of a higher degree elsewhere.

The computationally intensive algorithms in Chapters 2, 3 and 5 were coded in the C programming language and used the GNU Scientific Library for generating random variables. All other computational work in this thesis was carried out in the R statistical environment. All computer code included in this thesis was my own.

Theodore Kypraios

‘*Οταν θελεις κατι πολυ, σημαινει οτι δεν το εχεις.*’

List of Tables

List of Figures

Contents

Chapter 1

Introduction

1.1 Motivation

During the last two decades, sampling-based methods for performing Bayesian inference have been widespread. The need for considering realistic models to adequately explain particular phenomena has lead to inferential problems which involve multidimensional analytically intractable integrations. However, such integrations can be easily managed by using Monte Carlo methods which are particularly appropriate within this framework, (see for example, Smith and Roberts, 1993). Suppose, we have a probability density $\pi(x)$, corresponding to some random variable, X and a function f of interest. It is often the case that we might be interested in evaluating integrals of the following form:

$$\mathbb{E}_\pi(f) = \int_x f(x)\pi(x)\ \mathrm{d}x \tag{1.1}$$

Suppose that $\pi(x)$ is multidimensional and analytical calculations are impossible. However, we are able to draw a sequence of values, X_i, such that X_i are identically and independently distributed (i.i.d.) with density π. Then it is true that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right] = \mathbb{E}_\pi(f) \tag{1.2}$$

and by the *strong law of large numbers* if we take n to be large enough we could approximate the desired expectation by:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx \mathbb{E}_\pi(f) \tag{1.3}$$

Furthermore, we might also use the Central Limit Theorem (CLT), given that π admits a variance for the function $f(x)$, say σ^2, to see how accurate this estimate might be:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}_\pi(f))}{\sigma\sqrt{n}} \sim N(0,1) \tag{1.4}$$

Therefore, the computational challenge which has to be faced is how to draw samples from π which will be used in (1.3). Techniques which attempt to draw directly from $\pi(x)$ have been shown to have limited applicability. Instead, a large collection of powerful, iterative computational algorithms which are general and easy to implement, have found a great success within the statistical community since early 1990s. These methods are known as *Markov Chain Monte Carlo* (MCMC) and the main idea goes back to 1953 in the particle Physics literature (Metropolis et al., 1953). Then it was generalised in statistical context by Hastings (1970). Nevertheless it is much later with Gelfand and Smith (1990) that the statistical community became aware of the potential of MCMC for Bayesian inference. Since then, the use of Bayesian methods for applied statistical modelling has increased rapidly.

MCMC methods enable us to draw a sequence $X_n, n = 1, 2, \ldots$, which although neither independent nor identically distributed, still satisfies (1.3). The idea behind MCMC is the following: for a given distribution π, on an arbitrary state space $\mathcal{X}$, construct a Markov chain with the same state space and stationary distribution π. Then under mild conditions, a Markov chain's sample path X_n is an approximate and dependent random sample from π. Asymptotic results ensure for instance,

distributional convergence of the realisations, i.e.

$$X_n \xrightarrow{d} \pi$$

where $\xrightarrow{d}$ denotes the convergence in distribution. In addition, they ensure consistency of "ergodic averages", for any integrable scalar function f,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \int_{\mathcal{X}} f(x)\pi(x)\ \mathrm{d}x, \qquad \text{as } n \to \infty, \quad \textit{almost surely}$$

The dependence among the simulated values plays a very significant role in terms of the efficiency of an MCMC algorithm. The "ergodic averages" such as in (1.3) can become very unstable and converge very slowly to their strong limited values in the presence of very high serial correlation in the $\{X_n\}$ series.

Therefore, the motivation behind this thesis is to provide a general methodology for constructing efficient MCMC algorithms so as to reduce the serial dependence and obtain more reliable results.

1.2 Structure of the Thesis

This thesis is divided into two discrete parts. The first part is mainly concerned with drawing Bayesian inference for stochastic epidemic models. The focus is to construct and analyse a class of non−centered parameterisations which can improve the speed of the convergence of the Gibbs sampler (see Section 1.6.1 for definition) and other related MCMC algorithms. This part consists of three chapters and which are outlined below.

- **Chapter 2.** In this chapter, we will first explain why understanding the spread of an infectious disease is an important issue in order to prevent major outbreaks. We will also provide a historical background on deterministic and stochastic models which have been used throughout the literature. We shall

briefly review the previous work in epidemic modelling by mainly focusing on the general stochastic epidemic model and describing existing approaches for drawing classical (frequentist) and Bayesian inference for its associated parameters.

Furthermore, we introduce a more general and realistic model to capture the dynamics of infectious diseases. We will demonstrate how standard methods can be applied for inferential purposes and also show via an illustrative example that they can be problematic in some cases. Therefore, we will mainly focus how to develop a class of *centered* and $non-centered$ reparameterisations in order to obtain more robust and efficient algorithms.

- **Chapter 3.** This chapter is mainly concerned with modelling the 2001 UK Foot-and-Mouth (FMD) outbreak from a fully Bayesian perspective. First, we will refer to the previous work on modelling the FMD outbreak and then adopting the methodology presented in Chapter 2 we will focus on describing the transmission's dynamics of the disease. Moreover, we will compare our findings with those presented in the literature already.

- **Chapter 4.** In the final chapter of Part I, we discuss various extensions of methods and applications for partially observed stochastic epidemics. This chapter also includes a first attempt to provide a real-time risk assessment tool for a potential Avian Influenza outbreak in the poultry industry of the UK.

In the second part of the thesis we introduce a wide class of $semi-$parametric time series models based on an underlying stochastic process, which we term *latent branching tree*. The motivation behind this chapter is to develop a general methodology to construct time series with pre-specified marginal distributions of the observations and build the correlation structure around them. The structure of this chapter is as follows:

- **Sections 5.1, 5.2, 5.3**. In the beginning of this chapter we will briefly review the literature on constructing time series models with fixed margins outside the Gaussian context with a specific correlation structure. Then, we present some motivating examples of time series that we will be interested in modelling via our class of models.

- **Sections 5.4, 5.5, and 5.6.** In these sections, the construction of a latent branching tree based on diffusions is given and the general properties of the tree are discussed. We will refer to the nature of realisations obtained via the proposed stochastic process by focusing on their marginal distribution and their corresponding dependence structure.

- **Sections 5.7 and 5.8.** We show how we can simulate a latent branching tree *exactly* without the need of discretisation of the diffusions processes which are chosen to build the tree. We demonstrate how Bayesian inference can be conducted for the parameters of interest via MCMC methods. Moreover, we describe in detail alternative MCMC strategies, including non−centered parameterisations, so as to improve the efficiency of the standard algorithms.

- **Sections 5.10 and 5.11.** In these sections we first present a simulation study to illustrate the performance of the proposed class of models. Then we apply our methodology to analyse some real genome scheme data.

- **Sections 5.13 and 5.12.** Finally, we summarize the advantages of the proposed methodology and also discuss further extensions regarding generalisations of the existing methods and also extensions which are motivated by real applications.

1.3 Bayesian Inference

In this section we will describe the fundamentals of Bayesian inference. A rigorous and a more detailed approach can be found in Bernardo and Smith (1994).

1.3.1 Bayes' Theorem

Bayesian inference, similarly to likelihood inference, requires a sampling model that produces the *likelihood*, the conditional distribution of the data given the parameters. Then, the Bayesian approach will additionally place a *prior* distribution on the model parameters. The likelihood and the prior are then combined using Bayes theorem to derive the *posterior distribution*. The posterior distribution is the conditional distribution of the (unknown) parameters, denoted by $\boldsymbol{\theta}$ given the data, denoted by $\boldsymbol{Y}$. All Bayesian inference arises from the posterior distribution.

Adopting a Bayesian approach, a prior distribution is assigned to $\boldsymbol{\theta}$ and we are interested in deriving explicitly or sampling from the posterior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$. In the case of a continuous state space, the posterior turns out to be:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{\pi(\boldsymbol{\theta})L(\boldsymbol{Y}|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta})L(\boldsymbol{Y}|\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}} \tag{1.5}$$

We refer to this formula as the *Bayes'* theorem. The integral in the denominator is essentially a normalising constant and its calculation has traditionally been a severe obstacle in Bayesian computation. In Section 1.6, we will demonstrate how we can avoid its calculation using MCMC methods. In terms of a discrete state space the integral is substituted with a sum over the sample space of θ. In this thesis we are mainly concerned with continuous state spaces and therefore in the rest of this chapter we omit the corresponding results for the discrete state spaces.

Bayes' theorem can be used sequentially. Suppose that we have collected two independent data samples, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$.

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{Y}_1, \boldsymbol{Y}_2) &\propto L(\boldsymbol{Y}_1, \boldsymbol{Y}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\
&\propto L(\boldsymbol{Y}_2|\boldsymbol{\theta}) \times L(\boldsymbol{Y}_1|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \\
&\propto L(\boldsymbol{Y}_2|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}|\boldsymbol{Y}_1)
\end{aligned}$$

In other words, this means that we can obtain the full posterior of $\boldsymbol{\theta}$ given the full

dataset by first evaluating $\pi(\boldsymbol{\theta}|\boldsymbol{Y}_1)$ and then treating it as a prior for the second dataset $\boldsymbol{Y}_2$. Thus, we have a natural setting when the data arrive sequentially over time.

1.3.2 Priors

The choice of the prior distribution has drawn a considerable attention in the Bayesian community (see for example, Bernardo and Smith, 1994). In this section we briefly present some of the most popular approaches for choosing the priors. Additionally to the priors we mention here there exist the so called elicited priors, created using an experts opinion. However, elicitation methods go beyond the scope of this thesis and we shall not give more details here.

It is possible to select a distribution which is *conjugate* to the likelihood, that is, one that leads to a posterior belonging to the same family as the prior. Morris (1983) showed that exponential families, where likelihood functions often belong, do in fact have conjugate priors, so that this approach will typically be available in practice. The great advantage of such a prior is that can be more computationally convenient than others.

In many practical situations prior information about $\boldsymbol{\theta}$ is not available. Therefore, the need of specifying *non-informative* priors is essential. In other words, we would like to define a prior distribution $\pi(\boldsymbol{\theta})$ that contains very little information about the parameter of interest, $\boldsymbol{\theta}$ and argue that the information contained in the posterior about it, comes almost entirely from the data. Summarizing, we should always choose a prior for the parameter of interest very carefully.

1.3.3 Posterior Distribution

Having obtained the posterior distribution for the parameters of interest we have all the information that the data contain for the parameters. A natural first step is to plot the density function to visualise the current state of our knowledge. In

addition, we can obtain summaries of our posteriors which can give us all the information that can be obtained using a frequentist approach to inference. In this section we will mention the most commonly used in practice, point estimation and interval estimation.

Point estimation is readily available through $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$. The most commonly used location measures are the mean, the median and the mode of the posterior distribution since they all have appealing properties. Depending on the shape of the posterior distribution one of the aforementioned measures can be used.

In the case of a continuous parameter space $\boldsymbol{\Theta}$, a $100 \times (1-\alpha)\%$ credibility set for $\boldsymbol{\theta}$ is a subset of $\boldsymbol{\Theta}$ which satisfies the following:

$$1-\alpha \leq \mathbb{P}(C|\boldsymbol{Y}) = \int_C \pi(\boldsymbol{\theta}|\boldsymbol{Y}) \, \mathrm{d}\boldsymbol{\theta} \tag{1.6}$$

where integration is replaced by summation for discrete components of the parameter.

One of the most attractive credibility sets, is the *highest posterior density region* defined as:

$$C = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \pi(\boldsymbol{\theta}|\boldsymbol{Y}) \geq q(\alpha)\} \tag{1.7}$$

where $q(\alpha)$ is the largest constant satisfying $\pi(C|\boldsymbol{Y}) \geq 1-\alpha$. This credibility set consists of the most likely $\boldsymbol{\theta}$ values. Nevertheless, it can be hard to compute such integrals analytically and therefore numerical methods should be applied. On the other hand, a much easier and commonly used approach is to calculate the equal tail credibility set by simply taking the $\alpha/2-$ and $1-\alpha/2-$ quantiles of $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ which equals to the highest posterior density set for symmetric unimodal densities. Nevertheless, this is not the case for highly skewed distributions.

1.4 Bayesian Inference for Missing Data Problems

Let $\boldsymbol{Y}$ denote the observed data, $\boldsymbol{X}$ the missing data and $\boldsymbol{\theta}$ the parameters in the model. The statistical models considered in this thesis share a common structure: the distribution of $(\boldsymbol{Y}, \boldsymbol{X})$ is specified and depends on the parameter $\boldsymbol{\theta}$. Nevertheless, only $\boldsymbol{Y}$ is observed, and therefore $\boldsymbol{X}$ is treated as *missing data*. The pair of $(\boldsymbol{X}, \boldsymbol{Y})$ is often known as the *augmented or complete data*. The term "missing data" can either be interpreted as data which for some reason we failed to collect or data which are not available to us. On the other hand, in many cases, especially in models with latent variables, random effects, or hidden stochastic processes, we would never be able to observe $\boldsymbol{X}$.

By adopting a Bayesian approach, the conditional distribution of the parameter (in a continuous state space) given the observed data is given up to proportionality as follows:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) \propto \pi(\boldsymbol{\theta}) \int_{\boldsymbol{X}} \pi(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{X} \tag{1.8}$$

This means that in order to perform posterior inference for $\boldsymbol{\theta}$ we need to find the marginal distribution of the observed data given the parameters. In practice, in many complex statistical models used nowadays, for example in econometrics, geostatistics and engineering, the integral $\int_{\boldsymbol{X}} \pi(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{X}$ is neither analytically or numerically feasible.

Nevertheless, powerful iterative sampling schemes have been developed which allow us to sample from the joint posterior $(\boldsymbol{\theta}, \boldsymbol{X})$ by sampling iteratively the two conditionals $\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\theta}$ and $\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}$. This methodology is known as *data augmentation* (see Tanner and Wong, 1987) and is described in detail in Section 1.6.2. Once samples have been obtained from the joint distribution, $\pi(\boldsymbol{\theta}, \boldsymbol{X})$, then sampling based posterior inference for $\boldsymbol{\theta}$ (or $\boldsymbol{X}$) can be easily performed using Monte Carlo methods as for example in Ripley (1987), Gelfand and Smith (1990) and Smith

and Roberts (1993).

1.5 Conditional Independence

We say that two variables X and θ are independent, and we write $X \perp \theta$, when any information received for θ does not alter uncertainty about X, see Dawid (1979):

$$\pi(X|\theta) = \pi(X)$$

The concept of conditional independence is very important in this thesis. The centered and the non−centered parameterisations which are introduced in 2 and 5 are defined in terms of the conditional independence structure they impose between the missing data and the parameters.

Following Dawid (1979) who develops the theory of conditional independence in the statistical context, the random variables Y and θ are said to be conditionally independent given another variable X, when they are independent in their joint distribution conditional on $X = x$, for any value of x. That is

$$\pi(Y, \theta|X) = \pi(Y|X)\pi(\theta|X).$$

Marginally though, when X is unknown Y and θ could be dependent. The conditional independence is often expressed in terms of factorisation of the joint density of X, Y, θ. A compact and illustrative way of expressing conditional independence statements is by means of graphical models and such an approach is often adopted in this thesis. See Whittaker (1990) for an introduction to graphical modelling.

1.6 Markov Chain Monte Carlo Methods

Markov chain Monte Carlo methods are employed to (approximately) draw samples from a specific distribution π say, which is often called as *target distribution*. π is typically multidimensional and in the application we will be concerned in this thesis, is the joint posterior distribution of the parameters and the missing data in a hierarchical model.

In this section we will present the main idea and review some well known MCMC algorithms. There is a vast literature about the theory, methodology, implementation and applications of MCMC. Currently available texts on the subject include, for, example Gilks et al. (1996), Tanner (1996), Robert and Casella (1999) and Roberts and Tweedie (2006).

We shall briefly describe some of the MCMC algorithms most relevant for our purposes. For more details, we refer to the aforementioned books for details. The main idea behind MCMC methods has already been mentioned; for a given target distribution π, MCMC methods construct a Markov chain $\{X_n\}$ which has π as an invariant measure. Mild conditions ensure that π is also a limiting distribution of the chain, whatever the initial value X_0. Such Markov chains, are called ergodic. Most of the MCMC algorithms used in practice satisfy the condition which ensure convergence to the invariant distribution π. An essential task in designing an MCMC algorithm is to ensure that π is invariant which is mostly achieved using the idea of reversibility.

From a statistical perspective, the convergence in distribution of the Markov chain to π is exploited to estimate expectations under the invariant measure. More details about convergence results can be found in Roberts and Tweedie (Chapter 8, 2006). In Bayesian analysis, π is a posterior distribution and most inference problems come down to calculating expectations, (see for example, Gelfand and Smith, 1990). Therefore MCMC is a very powerful tool for posterior inference, although it can be easily applied outside the Bayesian context.

Having ensured the convergence to stationarity, the question which is of interest, is the speed at which an MCMC algorithms converges. This practically determines how much time we should "run" the chain before the simulated values are assumed to be drawn from π. A related concern is the dependence among the simulated values. Even if we start at stationarity by sampling $X_0 \sim \pi$, the Markov chain will generate exact but dependent samples from π. High dependence among the sample can often lead in very slow convergence of the ergodic average estimates to the expectations under π. The effect of the dependence among the sample is discussed in more detail in Section 1.10.

1.6.1 Gibbs Sampler

The Gibbs sampler decomposes the state space $\mathcal{X}$ as $X_1 \times \mathcal{X}_2 \times \cdots \mathcal{X}_k, k > 2$ and simplifies a complicated multi-dimensional simulation into a collection of k smaller dimensional which are often more manageable. Often, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{X}_i = \mathbb{R}^{r_i}$ and $\sum_i r_i = d$. The factorisation of the space is usually naturally suggested by the statical model which is considered. We adopt the following notation; we write $x = (x^{(1)}, \ldots, x^{(k)})$ for an element of $\mathcal{X}$ where denote by $x^{(i)} \in \mathcal{X}_i$, for all $1 \leq i \leq k$. Also denote by $x^{(-i)}$ for the vector produced by excluding the i_{th} component from the vector $\boldsymbol{x}$.

$$x^{(-i)} = \left(x^{(1)}, \ldots, x^{(i-1)}, x^{(i+1)}, \ldots, x^{(k)}\right)$$

We also follow the same notational conventions for the random variable $X \sim \pi$. The conditional distribution $X^{(i)}|X^{(-i)} = x^{(-i)}$ for all $i = 1, \ldots, k$ is denote by

$$\pi_i\left(\cdot|x^{(-i)}\right).$$

The Gibbs sampler which samples from π is implemented as shown below.

The Deterministic Scan Gibbs Sampler

1. `Choose` X_0;

2. `Set` $n = 0$;

3. `Repeat the following steps:`

   `Set` $i = 1$;

   `While` $i < k + 1$

   {

   `Sample` $X_{n+1}^{(i)} \sim \pi_i\left(\cdot|x^{(-i)}\right)$, `where`

   $$x^{(-i)} = \left(X_{n+1}^{(1)}, \ldots, X_{n+1}^{(i-1)}, X_n^{(i+1)}, \ldots, X_n^{(k)}\right)$$

   $i = i + 1$

   }

   $n = n + 1$

The above scheme is also referred to as the *deterministic scan* (DS) Gibbs sampler because of the way the algorithms visits each of the k components. It creates a Markov chain on $\mathcal{X}$ with transition kernel P which is the composition of k kernels, $P^{(i)}$, $i = 1, \ldots, k$. In particular, if $z, w \in \mathcal{X}$ we define

$$P^{(i)}(z, dw) = \begin{cases} \pi_i\left(dw^{(i)}|x^{(-i)}\right), & \text{for } w^{(-i)} = x^{(-i)} \\ 0, & \text{otherwise} \end{cases}$$

and $P_{DS} = P^{(k)}P^{(k-1)} \cdots P^{(1)}$. There are alternative updating schemes (see for example, Roberts and Sahu, 1997) which we describe below.

The *random scan* (RS) Gibbs sampler at each iteration chooses one of the k components to update. Therefore its transition kernel can be written as

$$P_{RS} = \frac{P^{(1)} + \cdots + P^{(k)}}{k}.$$

The RS Gibbs sampler can be implemented as follows:

The Random Scan Gibbs Sampler

1. Choose X_0;
2. Set $n = 0$;
3. Repeat the following steps:

   Sample I from $U(\{1, 2, \ldots, k\})$;

   Sample $X_{n+1}^{(I)} \sim \pi_i\left(\cdot | x^{(-I)}\right)$

   Set $X_{n+1}^{(j)} = X_n^{(j)}$, for $j \neq I$;

   $n = n + 1$

It can be checked that each $P^{(i)}$ is reversible with respect to π, from which easily follows that π invariant for either the composition, as in the DS or the mixture as in the RS Gibbs sampler of the $P^{(i)}$'s; see for example Theorem 3.4.2 and Proposition 3.3.3. of Roberts and Tweedie (2006).

Apart from the RS and DS, there exist some other variation of the Gibbs sampler; the *random permutation* Gibbs sampler chooses at each iteration a permutation of the components, and updates the components according to that permutation. Note that this preserves reversibility. Another natural way to make the Gibbs sampler reversible is to carry out two iterations of the Gibbs sampler, the second one being implemented with the order of the other components reversed. Note that by applying this algorithm, the k_{th} component is update twice successively, at the

end of the first iteration and the beginning of the second. The resulting algorithm is called *reversible* Gibbs sampler. The implementation of this kind of the latter algorithms is straightforward; see for example Roberts and Tweedie (Section 2.2.2, 2006).

1.6.2 The Two-Component Gibbs Sampler (Data Augmentation)

The data augmentation was originally developed by Tanner and Wong (1987) for finding fixed point solutions to integral equations which appear in statistical inference and it can be viewed as the stochastic analogue to EM algorithm (see Dempster et al., 1977). It is most often used to obtain samples from the joint distribution of $X = \left(X^{(1)}, X^{(2)}\right)$ say, by sampling from the conditional distributions. Such a scheme has a similar structure with the Gibbs sampler with Gelfand and Smith (1990) showing that the latter is at least as efficient as the former. Following the standard practice in the literature (see for example, Liu et al., 1994, Meng and van Dyk, 2001), we will identify in this thesis the data augmentation with the two-component Gibbs sampler.

Data augmentation is by far the most widely adopted computational method for performing modern Bayesian analysis of missing data problems. The target distribution is the joint posterior of the missing data $\boldsymbol{X}$ and the parameters $\boldsymbol{\theta}$. By construction, simulation from the conditional distributions $\pi(\boldsymbol{\theta}|\boldsymbol{X}, Y)$ and $\pi(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{Y})$ are tractable and more feasible than simulation from the marginal distribution of the parameters given the observed data, $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$. Note that there are many cases where the latter is not even available in closed form due to the integration in (1.5). Therefore we use the two-component Gibbs sampler which update $\boldsymbol{X}$ and $\boldsymbol{\theta}$, to obtain samples from $\pi(\boldsymbol{\theta}, \boldsymbol{X}|\boldsymbol{Y})$.

1.6.3 The Metropolis-Hastings Algorithm

The Metropolis algorithm (Metropolis et al., 1953) manages to sample π, at least approximately, in a way which does not require the knowledge of its normalisation constant. In this section we will describe the more general Metropolis−Hastings algorithm introduced by Hastings (1970). It is generally believed that most of the MCMC algorithms can be considered as a special case of this algorithm. We denote by π_u the un-normalised density on $\mathbb{R}^d$ with respect to d-Lebegue measure, μ_d^{Leb}. Also assume that is possible to carry out simulations of a Markov chain with transition density $q(X, \cdot)$ with respect to the same measure. Such a transition density, called *proposal density* does not need to have any connection with π_u, although its choice is important since it can actually influence the efficiency of the resultant Markov chain.

The Metropolis-Hastings algorithms proceeds as follows. An initial starting value X_0 is chosen; then given the current state of the chain, $X_n = x$, a candidate value $Y_{n+1} = y$ is generated according to the proposal density $q(X_n, \cdot)$. The generated values is then accepted with probability $\alpha(x, y)$, given by:

$$\alpha(x,y) = \begin{cases} \min\left(\frac{\pi_u(y)}{\pi_u(x)}\frac{q(y,x)}{q(x,y)}, 1\right), & \text{if } \pi_u(x)q(x,y) > 0 \\ 0, & \text{if } \pi_u(x)q(x,y) = 0 \end{cases}$$

If the candidate value is accepted, then we set $X_{n+1} = y$, otherwise if it is not accepted, we set $X_{n+1} = x$. It easy to see that the Markov chain induced by such an algorithm has transition law P with densities

$$p(x,y) = q(x,y)\alpha(x,y), \quad x \neq y$$

with respect to μ_d^{Leb} and with probability of remaining at the same value equal to

$$r(x) = \int q(x,y)(1-\alpha(x,y))\ \mathrm{d}y.$$

The algorithm is implemented as follows:

The Metropolis Hastings Algorithm

1. `Choose` X_0;
2. `Set` $n = 0$;
3. `Repeat the following steps:`

    `Sample` $Y_{n+1} \sim q(X_n, \cdot)$;

    `Sample` $U_{n+1} \sim U(0, 1)$;

    `If` $U_{n+1} \leq \alpha(X_n, Y_{n+1})$ `then`

        `Set` $X_{n+1} = Y_{n+1}$;

    `Else`

        `Set` $X_{n+1} = X_n$;

    $n = n + 1$

It can be easily proven (see for example, the Lemma 2.4.1. of Roberts and Tweedie, 2006) that the algorithm ensures reversibility of the chain with respect to π, i.e. satisfies the detailed balance

$$\pi(x)p(x, y) = p(y)p(y, x).$$

We should note that any $\alpha(\cdot, \cdot)$ which satisfies the following equation

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x)$$

can be used. A class of algorithms which have other accept/reject rules can be found in Peskun (1973). However, it turns out that the accept/rule of the Metropolis-Hastings algorithm optimises the proportion of ultimately accepted

moves. Therefore, it is also optimal in the sense of minimising the asymptotic variance of any ergodic average moment estimator (see for example Peskun, 1973, Tierney, 1998, Roberts and Tweedie, 2006).

The framework of the Metropolis-Hastings algorithm is very general since it does not impose any restriction on the choice of $q(\cdot,\cdot)$. Therefore, we will proceed by describing some special cases of this algorithm which have draw much attention in the literature. The simplest possible choice of for the proposal distribution chooses $q(\cdot,\cdot)$ to be independent of its first argument:

$$q(x,y) = q(y)$$

and therefore we can write the accept/reject ratio as

$$\alpha(x,y) = \min\left(\frac{\pi_u(y)}{\pi_u(x)}\frac{q(x)}{q(y)}, 1\right).$$

This is algorithm is called *Independence Sampler* and it is clear that by taking $q(\cdot)$ to be proportional to $\pi_u(\dot{)}$ the algorithm reduces to i.i.d. sampling from π.

The algorithm which was essentially introduced in Metropolis et al. (1953) is known as *Symmetric Random walk Metropolis*. The proposal distribution is of the following form

$$q(x,y) = q\left(|x-y|\right)$$

and reveals states that is a function of the distance between x and y. In this case the accept/reject ratio reduces to

$$\alpha(x,y) = \min\left(\frac{\pi_u(y)}{\pi_u(x)}, 1\right)$$

The accept/reject mechanism can be interpreted as follows. We accept all moves which increase π_u but reject moves which decrease π. Thus, the algorithm biases the random walk by moving towards modes of π more often that moving away

from them (Roberts and Tweedie, 2006). This algorithm became one of the most widely used MCMC methods due to the fact that is extremely easy to implement. In the accept/reject ratio, only $\pi_u(\cdot)$ is involved while the proposal densities do not take any part at all. Therefore many calculations can be avoided. Possibly, the most popular proposal for performing a RWM is typically of this form:

$$q(x, y) \equiv N(x, \sigma^2)$$

where σ is considered as a scaling factor chosen by the user to optimise algorithm performance; see for example Roberts et al. (1997).

Finally, the-so-called *Multiplicative Random walk Metropolis* offers an attractive alternative to the RWM when the state space is in the positive half line. Such an algorithm can be considered as a logarithmic random walk algorithm, in the sense that is equivalent to the RWM with a $N(0, \sigma^2)$ proposal distribution and target distribution obtained by a logarithmic transformation of the original target. The proposed move is to a random multiple of the current state. Thus, from the current state, x, we propose a candidate value $y = z \exp(U)$ where, $U \sim N(0, \sigma^2)$. The accept/reject ratio turns out to be:

$$\alpha(x, y) = \min\left(\frac{\pi_u(y)}{\pi_u(x)}\frac{y}{x}, 1\right).$$

It can be illustrated via simulations that such an algorithm can behave much more efficiently by having frequent short excursions into the tail of the target density especially in comparison of the RWM which has rare but lengthy excursions.

1.6.4 Metropolis within Gibbs

The Metropolis within Gibbs, also known as componentwise updating algorithm, is a hybrid of the Gibbs sampler and the Metropolis-Hastings algorithm and is used extensively in this thesis. Suppose that the state space is factorised as $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$

and we would like to use Gibbs sampler to obtain samples from π. Nevertheless, it is often the case that either or both of the conditional distributions $\pi_i\left(\cdot|x^{(-i)}\right)$ are of standard form so as to easily simulate from. The Metropolis within Gibbs algorithm replaces the direct simulation by a Metropolis-Hastings step which has $\pi_i\left(\cdot|x^{(-i)}\right)$ as the invariant distribution.

It is reasonable to assume that the ease in the implementation of the Metropolis within Gibbs over the Gibbs sampler comes at the expense of speed of convergence. Introduction of the Metropolis steps can have severe negative impact on the convergence rate of the algorithm (see for example, Sections 4.3 and 6.12.2 of Papaspiliopoulos, 2003). Nevertheless, there are Metropolis within Gibbs algorithms which perform better than the "pure" Gibbs; see examples and references in Section 2.7 of Roberts and Tweedie (2006).

The Metropolis-Hastings algorithm becomes very relevant when considering missing data problems where the space is factorised in terms of the parameters $\boldsymbol{\theta}$ and the missing data $\boldsymbol{X}$. In many complex models it is hard to design a Metropolis-Hastings algorithm for the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{\theta}$. On the other hand, the full conditional distribution of $\pi(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y})$ is often available in closed form and Gibbs sampler can be used straightforward to draw samples from it, while the conditional of $\pi(\boldsymbol{X}|\boldsymbol{\theta},Y)$ is not and therefore a Metropolis-Hastings algorithm is necessary. Thus, we resort to the Metropolis within Gibbs sampler which can be generalised to the case where $\mathcal{X} = \mathcal{X}_1,\ldots,\mathcal{X}_k$, with $k > 2$ and implemented as follows:

The Metropolis within Gibbs Algorithm

1. Choose X_0;

2. Set $n = 0$;

3. Repeat the following steps:

   Set $i = 1$;

   While $i < k + 1$;

   {

   Update $X_{n+1}^{(i)}$ according to $\pi_i\left(\cdot, x^{(-i)}\right)$, where

   $$x^{(-i)} = \left(X_{n+1}^{(1)}, \ldots, X_{n+1}^{(i-1)}, X_n^{(i+1)}, \ldots, X_n^{(k)}\right)$$

   $i = i + 1$

   }

$n = n + 1$

1.7 Hierarchical Models and Parameterisations

All Bayesian models can be viewed as hierarchical models, since we typically assume that the distribution of the observed data $\boldsymbol{Y}$ depends on some unobserved random quantities $\boldsymbol{X}$ whose distribution depends on other random quantities $\boldsymbol{\theta}$. The distribution of $\boldsymbol{\theta}$ depends on other quantities which can be assumed either random or known. An important property of this kind of model, as described above, is the conditional independence between $\boldsymbol{Y}$ and $\boldsymbol{\theta}$ given $\boldsymbol{X}$.

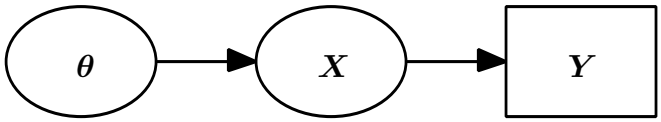

Figure 1.1: The graphical model of the centered reparameterisation

We term the parameterisation in terms of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ as the *centered parameterisation* (CP), due to the fact that the missing data are centered between the observed data and the parameters. Suppose instead, that we can find $\tilde{\boldsymbol{X}}$ and some function $h(\cdot,\cdot)$ such that $X = h(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})$ and $\tilde{\boldsymbol{X}}$ is *a priori* independent of $\boldsymbol{\theta}$. We term $(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})$ the *non−centered* parameterisation (NCP) and its graphical model is given in Figure 1.2

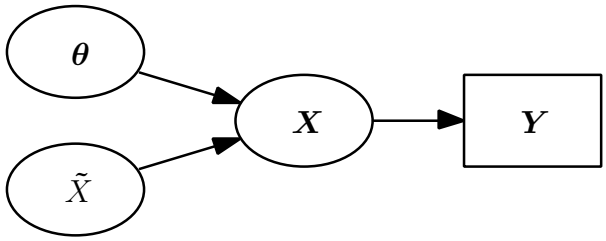

Figure 1.2: The graphical model of the non− centered reparameterisation

In both parts of this thesis, we are concerned with constructing NCP for missing data problems which share the aforementioned structure. Our goal is to find a reparameterisation to improve the performance of the Metropolis within Gibbs algorithm when it is slow under a CP.

1.8 Basics of Lèvy Processes

Lévy processes play an important role in the second part of this thesis. Therefore it is convenient to introduce, informally, some basic concepts and definitions at this early stage. A stochastic process $x(t)$, $t \geq 0$ where $x(0) := 0$ almost surely, is called a Lèvy process if it has independent and stationary increments, i.e. $x(t +$

$s) - x(t), t, s > 0$, is independent of the history of the process up to time t and its distribution depends only on the separation s (see for example Sato, 1999).

A simple Lévy process is the *Poisson process*, a stochastic process which finds applications in diverse areas of science such as physics, teletraffic modelling and biology. A counter is introduced which counts the number of occurrences from a starting point, and set $x(t)$ to be the number of occurrences in the interval $(0, t]$. We assume that occurrences in disjunct intervals are independent of each other. In addition, the distribution of the increments does not change in time, i.e. the process $x(t)$ is said to have *stationary increments*. Finally, the number of occurrences after time t follows the probability function

$$P\left(x(t) = x\right) = \exp\{-\lambda t\}\frac{(\lambda t)^x}{x!}$$

where λ is the *intensity* of the occurrences. In other words, the number of occurrences at time t, $x(t)$ is Poisson distributed with rate λt.

Another example of a Lèvy process is the *Brownian motion*. In its standard form, $x(1) \sim N(0, 1)$, but more generally we can have $x(1) \sim N(0, \sigma^2)$. The increments of this process are Gaussian

$$x(t + s) - x(t) \sim N(0, s\sigma^2)$$

a property which can be used to simulate values from this process; for instance, Figure 1.3 shows a standard Brownian motion path on $[0, 1]$ which has been simulated by splitting time in small intervals and simulating from the corresponding increments. It can be shown that the Brownian motion is the only Lèvy process with almost sure continuous sample path (see for example, Feller, 1971).

Finally, a *Gamma process* which is specified by $x(1) \sim Ga(\alpha, \beta)$ is another example of Lévy process. The increments are also Gamma distributed

$$x(t + s) - x(t) \sim Ga(\alpha s, \beta).$$

This is a pure jump process, a feature shared by all Lèvy processes with positive increments. The Gamma process has an infinite number of jumps in any bounded interval of time, but only a finite number of them are non−negligible size; see Section 5.8 of Papaspiliopoulos (2003) for more details.

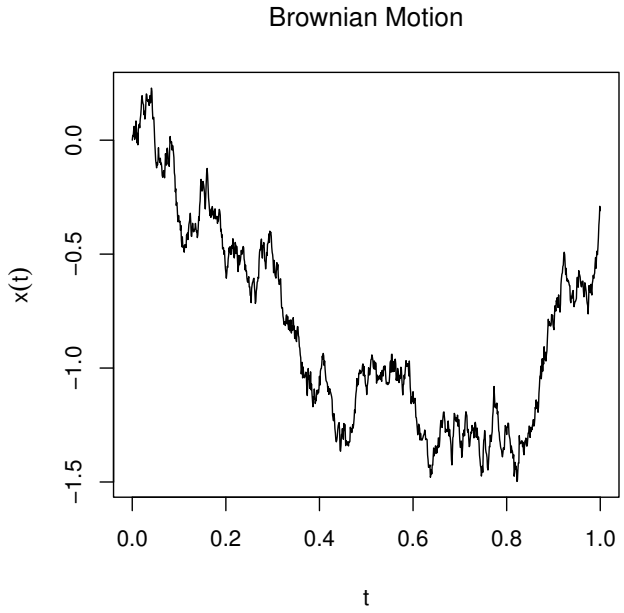

Figure 1.3: A path in $[0, 1]$ of a standard Brownian motion. It has been simulated by discretising time in intervals of length 0.001 and simulating from the corresponding increments of the process

1.9 Non−Centered Parameterisations for Bayesian Hierarchical Models

In the first part of this thesis, we are mainly concerned with developing and constructing a framework for applying NCP parameterisation for partially observed stochastic epidemic models so as to improve the efficiency of the existing centered algorithms. An extensive account of the second part refers to methods of drawing inference via MCMC methods. We will show that in some cases a NCP can significantly perform better than the corresponding centered. Therefore, in this section we will review the basic concepts of the non−centered methodology; we refer to

Papaspiliopoulos (2003) and Papaspiliopoulos et al. (2003) for more details.

1.9.1 Motivation

Convergence of the MCMC algorithms, particularly when using Gibbs sampler or related techniques, depends crucially on the parameterisation adopted for the unknown quantities. A centered parameterisation is a very natural framework for both a modelling and interpretation perspective; that is to use $\boldsymbol{\theta}, \boldsymbol{X}$. Thus an algorithm for sampling from the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{X}$ which we will consider them as parameters and missing data respectively, given the observed data $\boldsymbol{Y}$ can be implemented as follows:

Centered Algorithm

1. Update $\boldsymbol{\theta}$ by drawing samples from the conditional distribution $\pi(\boldsymbol{\theta}|\boldsymbol{X}, Y)$;
2. Update $\boldsymbol{X}$ by drawing samples from the conditional distribution $\pi(\boldsymbol{X}|\boldsymbol{\theta}, Y)$.

In many complex hierarchical models, the full conditional distribution of the parameters given the missing data is of a standard form and Gibbs sampler can be applied. On the other hand, the conditional distribution of the missing data given the parameters it is not of an easy form and therefore a Metropolis-Hastings algorithm is essential; that is the Metropolis within Gibbs algorithm.

Figure 1.1 reveals the *a priori* dependence between $\boldsymbol{X}$ and $\boldsymbol{\theta}$ and in many contexts this dependence is very strong. The presence of data tends to reduce the effect of that dependence, but the efficiency of the centered algorithm will depend crucially on this. The motivation behind non−centering is to find an alternative parameterisation $(\boldsymbol{X}, \boldsymbol{\theta}) \rightarrow (\tilde{\boldsymbol{X}}, \boldsymbol{\theta})$ where the new missing data $\tilde{X}$ is *a priori* independent

of $\boldsymbol{\theta}$. The corresponding MCMC algorithm can be implemented then as follows:

Non−Centered Algorithm

1. `Update` $\boldsymbol{\theta}$ `by drawing samples from the conditional distribution` $\pi(\boldsymbol{\theta}|\tilde{\boldsymbol{X}}, Y)$;
2. `Update` $\boldsymbol{X}$ `by drawing samples from the conditional distribution` $\pi(\tilde{\boldsymbol{X}}|\boldsymbol{\theta}, Y)$.

Although a Gibbs step may be feasible for drawing samples from $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ under a CP, this might be not the case under a NCP. In other words, the conditional distribution of the parameters $\boldsymbol{\theta}$ could be not of a standard form and therefore a Metropolis-Hastings step is needed. This leads to a significant computational edge in favour of CP. Nevertheless, as Papaspiliopoulos et al. (2003), we also believe that there is an important role of the NCP in many contexts especially in hierarchical models where the latent process is relatively weakly identified by the data. In addition NCP have much to offer when there exists high *a priori* dependence between the missing data and the model's parameters.

1.9.2 Rates of Convergence of the Gibbs Sampler

In this section we will focus on the rate of convergence of the Gibbs sampler for the two different parameterisations within the Gaussian context. Following Papaspiliopoulos et al. (2003), let $Z = (Z_1, Z_2)$ denote a random variable with density π, partitioned into two components, Z_1, Z_2 of arbitrary dimension. A two-component Gibbs sampler on π under the parameterisation (Z_1, Z_2) iterates the following procedure.

1. `Sample` Z_1 `from the conditional distribution of` $Z_1|Z_2$

2. `Sample` Z_2 `from the conditional distribution of` $Z_2|Z_1$

It is beyond the scope of this thesis to discuss rates of convergence of algorithms; see Roberts and Tweedie (2006) for a recent summary. Nevertheless, when the two-component Gibbs sampler can be implemented, there exist a complete theory which we will very briefly describe. Denote by $\mathcal{L}^2$ the set of all real functions f, $f : \mathcal{Z} \to \mathcal{R}$, which are square-integrable with respect to π, i.e.

$$\mathcal{L}^2 := \left\{ f : \int_{\mathcal{Z}} (f(z))^2 \pi(z) \, \mathrm{d}z < \infty \right\} \tag{1.9}$$

Similarly, we define:

$$\mathcal{L}_0^2 = \left\{ f \in \mathcal{L}^2 : \int_{\mathcal{Z}} f(z)\pi(z) \, \mathrm{d}z = 0 \right\} \tag{1.10}$$

Let $P^n(x, \cdot)$ denote the distribution of the two-component Gibbs sampler after n iterations, where x denotes an arbitrary starting value for the (Z_1, Z_2) pair. The $\mathcal{L}^2$ rate of convergence, denoted by ρ, is understood as the rate at which expectations of arbitrary square-integrable functions $f \in \mathcal{L}_0^2$ converge to their stationary values as $n \to \infty$ according to the $\mathcal{L}^2$ norm. The $\mathcal{L}^2$ norm for any signed measure μ non-singular with respect to π is defined as

$$||\mu||_{\mathcal{L}^2}^2 = \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right)^2 \mathrm{d}\pi. \tag{1.11}$$

Amit (1991) observed the $\mathcal{L}_2$ distance from stationarity decays as $A(x)b(n)\rho^n$ for some function $b(n)$ which varies slower than an exponential function. The rate $\rho \leq 1$ is defined as

$$\rho^{1/2} = \sup \, \mathrm{corr} \left(f(Z_1), g(Z_2) \right) \tag{1.12}$$

where the supremum is taken with respect to all real-valued non-constant functions f and g which have finite variances under π. Amit (1991) also showed that other norms, such as total variation distance, have this rate at least for a large class of

plausible target distributions.

As Papaspiliopoulos et al. (2003) point out, it has been long recognised that the correlation structure of the target distribution determines the convergence behavior of the Gibbs sampler; see Hills and Smith (1992) and Gelfand et al. (1995).

Equation 1.12 is of little practical use, since in general it is not possible to evaluate the supremum; nevertheless, an important exception is for Gaussian target π where supremum of the kind appearing in 1.12 are achieved exclusively by linear functions. In addition, for Gibbs samplers with larger number of components, it is impossible to find an explicit statement similar to Equation 1.12 which relates the rate of convergence of the algorithm to the target distribution correlation structure.

1.9.3 Rates of Convergence for CP and NCP for a Normal Hierarchical Model.

In this section, we refer to the results obtained by Roberts and Sahu (1997) where in the case of a Gaussian target distribution explicit formulae are available for rates of convergence of the sampler, in terms of target distribution correlation matrix. Following Papaspiliopoulos et al. (2003) we will consider the following Normal Hierarchical model written as

$$\begin{aligned} Y_i &= X_i + \sigma_y \epsilon_i, \\ X_i &= \theta + \sigma_x z_i, \qquad i = 1, \ldots, m \end{aligned} \tag{1.13}$$

Here, ϵ_i and z_i are standard Normal random variables, θ is assigned a uniform improper prior and the variances are considered to be known. The parameterisation $(\theta, \boldsymbol{X})$, where $\boldsymbol{X} = (X_1, \ldots, X_m)$ is known as centered parameterisation; see Gelfand et al. (1995).

The name non−centered parameterisation was originally used for the NHM in

Gelfand et al. (1995). In this context the NCP writes the model as

$$\begin{aligned} Y_i &= \tilde{X}_i + \theta + \sigma_y \epsilon_i, \\ \tilde{X}_i &= \sigma_x z_i, \qquad i = 1, \ldots, m \end{aligned} \tag{1.14}$$

Note that $\tilde{\boldsymbol{X}} = (\tilde{X}_i, \ldots, \tilde{X}_m)$ and θ are *a priori* independent but conditionally on the data, they are dependent. In this example, Gibbs sampling can be applied very easily in this context using either the CP or the NCP and therefore we are interested in assessing the performance of the sampler under these two different parameterisations.

We would like to consider sampling from the joint posterior distribution of $\boldsymbol{X}$ and θ of the model which appears in Equation 1.13 using a Gibbs sampler. Since this is a multivariate Gaussian distribution we can explicitly evaluate the rate of $\mathcal{L}^2$ convergence, denoted by ρ_{C}, using the results from Roberts and Sahu (1997); see also Roberts and Tweedie (2006),

$$\rho_{\mathrm{C}} = 1 - \kappa$$

where

$$\kappa = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2}.$$

Note that since $(\sigma_x^2 + \sigma_y^2)^{-1} = 1/\mathrm{var}(\theta|\boldsymbol{Y})$ and $1/\mathrm{var}(\theta|\boldsymbol{X}, \boldsymbol{Y}) = 1/\mathrm{var}(\theta|\boldsymbol{X})$ the expression for κ can be also written as

$$\kappa = \frac{(\sigma_x^2 + \sigma_y^2)^{-1}}{(\sigma_x^2)^{-1}}$$

which is the ratio of observed by augmented information for θ under the CP. Within this context, $1 - \kappa$ is the Bayesian fraction of missing information in the sense defined by Rubin (1987). The relationship between observed and augmented information and rates of convergence of algorithms was noted first in a very general

framework for the EM algorithm, (see for example Meng and van Dyk, 1997) but can be translated to the data augmentation methodology in this specialised linear model context (Sahu and Roberts, 1999). Therefore, the CP will perform well when $\kappa \to 1$, i.e. when the data are relatively very informative in the sense that the observed data contain almost as much information about the parameter as the augmented.

In the case of a NCP and when a Gibbs sampler is used, the $\mathcal{L}^2$ rate of convergence, denoted by ρ_{nc}, turns out to be:

$$\rho_{\text{nc}} = \kappa.$$

When the one parameterisation produces very slow mixing for the Gibbs sampler the other will be performing very well. For this model with flat priors assumed the relative performance of the CP and the NCP can be derived explicitly since $\rho_{\text{nc}} = 1 - \rho_{\text{C}}$. Not that this relation does not hold when proper priors are used.

1.9.4 General Framework for Non-Centered Parameterisations

The NCP for the NHM enable us to formulate a general framework for constructing non−centered parameterisations to a much more general context. Specifically, we find $\tilde{X}_i$ which is *a priori* independent of θ and from which X_i can be constructed via a deterministic function:

$$X_i = h(\tilde{X}_i, \theta).$$

Within the Gaussian context under a NCP, it is easy to identify $h(\cdot, \cdot)$ as $h(X_i, \theta) = \theta + \tilde{X}_i$. For the general model presented in the graphical model in Figure 1.1 although such a function $h(\cdot, \cdot)$ always exists, it is not unique. However, it can be difficult to identify such a function h which is analytically sufficiently tractable to use. A commonly used technique to find an appropriate function h is via the

inverse CDF method.

From the experience in the NHM context, Papaspiliopoulos et al. (2003) argue that we would expect an NCP to be more effective than its CP rival when $\boldsymbol{X}$ is poorly identified by the data and remains highly correlated with θ.

1.9.5 Partially Non−Centered Algorithms

Motivated by the results of Section 1.9.2, we know that the CP is the optimal algorithm where the relative observation error σ_y/σ_x tends to zero. On the other hand, the NCP is optimal when this error tends to infinity, i.e. the absence of any data (Papaspiliopoulos et al., 2003). Therefore, we would like to construct an algorithm that will take into account the quantity of observation present in the observed the data. Consider the following parameterisation for the NHM:

$$\begin{aligned} Y_i &= \omega\theta + \tilde{X}_i^{\omega} + \sigma_y \epsilon_i \\ \tilde{X}_i^{\omega} &= (1-\omega)\theta + \sigma_x z_i \end{aligned}$$

where $i = 1, \ldots, m$ and $\omega \in [0, 1]$. This parameterisation is called *partially non−centered* (PNCP). It can be easily seen that

$$\tilde{X}_i^{\omega} = (1-\omega)X_i + \omega\tilde{X}_i$$

where X_i and $\tilde{X}_i$ as defined (1.13). Obviously, if $\omega = 0$ ($\omega = 1$), then the above reparameterisation is just the CP (NCP). Since the joint posterior distribution of $\tilde{\boldsymbol{X}}^{\omega} = \left(\tilde{X}_1^{\omega}, \ldots, \tilde{X}_m^{\omega}\right)$ and θ remains Gaussian, the rate of convergence of the corresponding Gibbs sampler under this parameterisation can be derived and taken from Papaspiliopoulos (2003) is equal to

$$\rho_{\text{pnc}}^{\omega} = \frac{\omega - (1-\kappa)^2}{\omega^2\kappa + (1-\omega)^2(1-\kappa)} \tag{1.15}$$

Note that $\rho^{\omega}_{\text{pnc}} = 0$ for $\omega = 1 - \kappa$ which suggests that the PNCP algorithm can be tuned appropriately to produce IID samples by setting $\omega = 1 - \kappa$.

Partial non−centering can be used for many models outside the Gaussian context. Papaspiliopoulos et al. (2003) indicate that there is no unique way of defining a continuum of partial non−centering strategies. Nevertheless, they note that often there will be a natural one suggested by the model structure.

Concluding, we should bring to attention that outside the Gaussian context it is rare that pure Gibbs sampling can be applied in conjunction with the PNCP and therefore appropriate Metropolis within Gibbs strategies will be necessary.

1.10 Quantification of the Algorithm's Efficiency

Suppose, we have a probability density $\pi(x)$, corresponding to some random variable, X and a function f of interest. We have already explained in Section 1.1 that often in Bayesian statistics we are interested in evaluating expectations of the following form:

$$\mathbb{E}_{\pi}(f) = \int_{x} f(x)\pi(x)\ \mathrm{d}x. \tag{1.16}$$

It has been described in Section 1.6 how various MCMC algorithms allow us to draw a sequence $\{X_n, n = 1, 2, \ldots\}$ which although neither independent nor identically distributed still satisfies the following:

$$\widehat{f}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx \mathbb{E}_{\pi}(f) \tag{1.17}$$

Nevertheless, as Sokal (1996) points out that the key difficulty, is that the successive draws $X_1, X_2, \ldots$ might be very strongly correlated. Therefore the variance of the estimates produced from Monte Carlo simulation based on these samples may be much higher than "static" Monte Carlo, i.e. the case of i.i.d. samples. In this section we describe some useful measures in order to assess the efficiency of MCMC algorithms. Following Sokal (1996) we are interested in deriving measures

which could quantify the efficiency of an MCMC algorithm.

For a stationary chain, X_1 is sampled from $\pi(\cdot)$ and therefore for all $k > 0$ and $m \geq 0$:

$$\text{cov}(f(X_k), f(X_{k+m})) = \text{cov}(f(X_1), f(X_m))$$

which is the autocovariance at lag m. Also, because of stationarity,

$$\text{var}(f(X_i)) = \sigma^2(f), i = 1, \ldots, n.$$

The variance of the estimator $\widehat{f}_n$ can be calculated as follows:

$$\begin{aligned}
\text{var}(\widehat{f}_n) &= \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} f(X_i)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{var}(f(X_i)) + 2\sum_{i=1}^{n-1}\text{cov}(f(X_i), f(X_{i+1}))\right. \\
&+ \left. 2\sum_{i=1}^{n-2}\text{cov}(f(X_i), f(X_{i+2})) + \ldots\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{var}(f(X_i)) + 2(n-1)\text{cov}(f(X_1), f(X_2))\right. \\
&+ \left. 2(n-2)\text{cov}(f(X_1), f(X_3)) + \ldots\right) \\
&= \frac{\sigma^2(f)}{n}\left(1 + 2\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)\frac{\text{cov}(f(X_1), f(X_i))}{\text{var}(f(X_1))}\right) \\
&= \frac{\sigma^2(f)}{n}\left(1 + 2\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)\text{corr}(X_1, X_{1+i})\right) \qquad (1.18)
\end{aligned}$$

Equation 1.18 should be compared with the corresponding equation for uncorrelated random variables which turns out to be $\sigma^2(\widehat{f}_n) = \sigma^2(f)/n$. The difference is the factor in the bracket of (1.18) which is defined as *the integrated autocorrelation time* (IAT):

$$\tau_{int} = 1 + 2\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)\frac{\text{cov}(f(X_1), f(X_i))}{\text{var}(f(X_1))} \qquad (1.19)$$

Often, we are interested in the limit as $n \to \infty$ and therefore IAT becomes:

$$\widetilde{\tau}_{int} = 1 + 2\sum_{i=2}^{\infty} \text{corr}(X_1, X_i). \tag{1.20}$$

IAT represents the effective number of dependent samples that is equivalent to a single independent sample. On the other hand, the quantity $n/\widetilde{\tau}_{int}$ may be considered as the effective equivalent sample size if the elements of the chain had been independent.

In order to estimate IAT in practice, we could examine the chain from the point at which is seemed to have converged and estimated the correlation at lag i as follows:

$$\widehat{\gamma}_i = \frac{1}{n-i}\sum_{j=1}^{n-i}\left(f(X_j) - \widehat{f}_n\right)\left(f(X_{j+i}), \widehat{f}_n\right).$$

Then substituting the estimated autocorrelations in (1.20) gives an estimate of the IAT. Nevertheless, it is difficult to obtain precise estimation of the autocorrelation function at lag i when i is typically large, since when i is large, $\text{corr}(f(X_1), f(X_i))$ adds a constant amount of noise. Therefore, an accurate estimation of $\widetilde{\tau}_{int}$ is typically a hard task.

An other practically important issue is how can we be make sure that convergence has been reached. Note, that the theory says for large n the resulting values of the chain, say $X_n, X_{n+1}, \ldots$, is an approximate sample from the target distribution. In practice, the problem is to determine what a "large" n means. There are a number of diagnostic tests proposed in the literature (see for example Brooks and Gelman, 1998, Brooks and Roberts, 1999, Cowles and Carlin, 1996, and the refererences therein) that provide us with different indicators on the stationarity of the chain. However, none of these tests can actually guarantee convergence. Throughout this thesis, we investigate the "trace", a plot of the history, of the chain for long (typically a few hundreds of iterations) runs. All the results reported in this thesis are based on chains that appear to have converged.

Part I

Efficient Bayesian Inference for Partially Observed Stochastic Epidemics

Chapter 2

Bayesian Inference for Stochastic Epidemic Models

2.1 Introduction

Understanding the spread of an infectious disease is a highly crucial issue in order to prevent major outbreaks of an epidemic. Human infections such as influenza, malaria and HIV are still major causes of morbidity and mortality worldwide. In 2001, the UK experienced a range of severe economic and social effects of a Foot-and-Mouth (FMD) epidemic. It is also remarkable the threat that governed humanity when a new infection, *Severe Acute Respiratory Syndrome* (SARS), was spreading speedily across the world in the spring of 2003. More recently, many European countries have suffered from the *Highly Pathogenic Avian Influenza* (HPAI) disease which affected their poultry industries. For both FMD and SARS, considerable transmission of the disease had already taken place even before the danger had been noticed. Therefore, the available control strategies need to be imposed rapidly so as to effectively stop the spread of the infection. A detailed and careful understanding of the basic theory of epidemic models is essential so as to enable us to develop successful policies.

This chapter is concerned with methods for drawing inference for epidemic models using Markov Chain Monte Carlo (MCMC) methodology. First, we will explain why models are important in epidemic theory (2.1.1). Then we will provide a review of the history of epidemic modelling by discussing simple deterministic and stochastic models, focusing on their differences and similarities (Section 2.1.2). We will refer to previous work on epidemic modelling and concentrate on the stochastic model which has drawn a considerable attention within the literature, the *general stochastic epidemic* (GSE) (Section 2.1.4). The GSE will be the basis of the model we introduce in Section 2.2. Furthermore, we demonstrate that standard MCMC algorithms often lead to inadequately-mixing Markov chains (Section 2.6) and that more efficient algorithms are required. These algorithms are based on non−centered parameterisations and are presented in Sections 2.4 and 2.5.

2.1.1 The Need for Epidemic Models

The analysis of outbreak data can be more effective when it is based on a model for the actual process which generates the data. Models could be used to provide a better understanding of the transmission dynamics, the infection process, and the epidemiologically quantities of interest. The suitability of the model and the validity of the assumptions on which it is mainly based, depend on the purpose for which it was constructed.

A number of reasons exists for using epidemic models of historical incidence data. Such an analysis can be useful for diseases which occur due to re-emerging pathogens as described in a review by Ferguson et al. (2003). This is of particular interest at the moment because the world recently has faced the danger of a Highly Pathogenic Avian Influenza disease. Ducatez et al. (2006) and Enserink (2006) discuss multiple introduction of the disease in Nigeria and how could this result to transmission of the disease to Europe. The world also experienced a SARS outbreak in 2003, see for example Riley et al. (2003) and Lipsitch et al.

(2003), with significant impacts to the public health (Anderson et al., 2004). Two years earlier, in 2001, the UK suffered a Foot−and−Mouth epidemic which had a rather significant economic impact on the areas which were affected (Bennett et al., 2001).

When an epidemic model is fitted to a data set and is found to provide an adequate description of the mechanism which has generated the data, we can make use of the fitted model in several ways. In general, apart from providing estimates for the parameters of interest which are largely responsible for driving the dynamics of the disease, models also have the ability to answer questions which refer to the progress of the disease based on the current state of the outbreak. The FMD in 2001 is one of the cases which illustrates that models could also be applied for real time use (see for example, Ferguson et al., 2001a,b, Keeling et al., 2001, Morris et al., 2001).

In addition, epidemic models play an important role in determining the effect of different control strategies. One of the major strengths of epidemic models is their capability to predict where the disease is likely to spread next. This can guide the conduction of effective control policies to prevent a major spread. It can also suggest optimal plans for controlling a future outbreak by adopting vaccination strategies; see for example in the context of Food and Mouth (Keeling et al., 2003, Tildesley et al., 2006). It is therefore important to construct a model for which we can draw inference regarding its unknown quantities so as to be able to give answers to important scientific questions regarding the underlying processes of an outbreak.

2.1.2 Historical Background

Mathematical modelling of infectious diseases has a long history; see for example Bailey (1975). The first approach to epidemic modelling is generally taken to be a paper by Daniel Bernoulli on the prevention of an infectious disease, namely small-

pox, by inoculation. The analysis which was performed then by Bernoulli can be found in Daley and Gani (1999, Sec. 1.1). Nevertheless, as Bailey (1975) points out, it was another hundred years before the physical basis for the cause of the infectious disease became well-established. One of the earliest studies of epidemic modelling was introduced in a paper by Hamer (1906). The author, assumed that the probability of a new infection in the next discrete time step is proportional to the product of the number of susceptibles and the number of infectives. A few years later, Ross (1916, 1917a,b) translated this "mass action principle" or "homogeneous mixing" to the continuous time setup. The first complete mathematical model for the spread of an infectious disease which received attention in the literature was a deterministic one, introduced by Kermack and McKendrick (1927). We shall briefly describe the main features of this model in Section 2.1.2.1.

2.1.2.1 Deterministic Models

First, consider a closed population (i.e. there are neither births nor deaths nor immigration) of size $\mathcal{N} + a$ and assume that at time $t = 0$ there are a initially infected individuals. Such an assumption of a closed population is reasonable for epidemics which occur in a time relative to the change in the population. At any given point time each individual i is in one of the three states: i) **S**usceptible ii) **I**nfected iii) **R**emoved.

The only transitions which we allow, are the following: from susceptible to infected and from infected to removed. Therefore an individual is called *susceptible* if they do not have the disease but are susceptible to infection, *infected* if they have got the disease and able to infect other (susceptibles). We assume that at the end of their infectious period, they become *removed* either by death or immunity, i.e. cannot infect any other susceptibles. In general they do not take part in the epidemic any longer.

Denote by X_t, Y_t and Z_t the number of susceptibles, infected and removed individ-

Figure 2.1: The three transition states of an individual.

uals respectively at time $t \geq 0$. It is sufficient for describing the epidemic to keep track of (X_t, Y_t, Z_t) since for all t the following equality holds: $X_t + Y_t + Z_t = \mathcal{N} + a$. The model is then defined by the following set of differential equations:

$$\begin{aligned} \frac{dX_t}{dt} &= -\beta X_t Y_t \\ \frac{dY_t}{dt} &= \beta X_t Y_t - \gamma Y_t \\ \frac{dZ_t}{dt} &= \gamma Y_t \end{aligned} \tag{2.1}$$

with initial state $(X_0, Y_0, Z_0) = (x_0, y_0, 0)$. The factor $\beta X_t Y_t$ is a crucial non−term indicating that infections occur at high rate only when there are many susceptibles and infectives. It follows from the above equation that $dX/dZ = -(\beta/\gamma)X$. So,

$$X_t = x_0 \exp\{-\theta Z_t\}$$

and hence

$$Y_t = \mathcal{N} - Z_t - X_t = \mathcal{N} - Z_t - x_0 \exp\{-\theta Z_t\}$$

where $\theta = \beta/\gamma$. Kermack and McKendrick (1927) showed that the number of infectives Y_t is increasing unless $x_0 > 1/\theta$. That is, there will be a growing epidemic. This observation is known as the threshold result, i.e. different behavior of the epidemic will occur depending on whether $x_0 > 1/\theta$ or not. Another important observation is that as $t \to \infty$ then $Z_t \to Z_\infty < \mathcal{N}$ where Z_∞ is the solution of $Z = \mathcal{N} - x_0 \exp\{-\theta Z\}$. In other words, this a very important property which states that not everyone becomes infected. Summarizing, we should note that many of the epidemic models used today have this general epidemic model as

their basis.

2.1.2.2 Stochastic Models

Stochastic epidemic models were also being developed early in the 20^{th} century along deterministic ones. McKendrick (1926) was the first to propose a stochastic version of the general epidemic model. However, at that time, there was more interest in discrete−time models and this model did not receive much attention. A model which attracted more attention that time was the chain-binomial model proposed of Reed and Frost in lectures in 1928 (Wilson and Burke, 1942, 1943).

In the standard Reed−Frost model, given the numbers X_t, Y_t of the susceptibles and infectives at time t, Y_{t+1} has a binomial distribution with index X_t and mean $X_t(1-p)^{Y_t}$ and $Y_{t+1} = X_t - X_{t+1}$. In consequence, individuals are assumed to be infective for a single time unit and in that time they can make an infectious contact, independently and with probability p, with any member of the population who is susceptible. This means that the number of potentially infectious contacts scales with the population size. Since the Reed−Frost model is only usually applied to small populations, this is not a problem. However, there have been various modifications to the Reed−Frost model which refer to the number of contacts and the probability that a susceptible escapes the infection by a single infective (see for example, Dietz and Schenzle, 1985).

The stochastic models began to draw more attention and be analyzed more extensively in the late 1940's. Then, Bartlett (1949) studied the stochastic version of the model introduced by Kermack and McKendrick (1927) and since then, the amount of effort put into modelling infectious disease has blown out.

2.1.2.3 Deterministic or Stochastic?

Disease spread is an inherently stochastic phenomenon and there are a number of arguments why a stochastic model should be preferable to a deterministic one.

Real life epidemics, can either go extinct with a small number of individuals who became infected during the outbreak, or end up with a significant proportion of the population having contracted the disease. It is therefore, only stochastic models that can capture this behavior and the probability of each event occurred. Moreover, stochastic models allow us to intuitively define them since they can naturally capture the infection process between different individuals.

Isham (2005) claims that the general view in the past seems to have been that a deterministic model gives an average behaviour of a corresponding stochastic system at least asymptotically and that for large populations using a stochastic model, which is more difficult to analyse than a deterministic one, there is little to be gained. However, it is now widely accepted that both deterministic and stochastic models have their strengths and can accommodate good understanding of the underlying process (see for example Isham, 2005). We should note that it is often the case to observe a disease outbreak with an atypical behavior even in the case for large populations. Nevertheless, even if they show an average behaviour care needs to be taken when we are interested in prediction (Isham, 1991, 1993).

Isham (2005) also indicates that one of the most noticeable changes of the last fifteen years has been the increased acceptance by biologists of the important role that mathematical modelling has to play in providing solutions of many of their most difficult problems. Moreover, they noticed that such models need to incorporate *intrinsic* stochasticity in many ways.

Similarly, the stochastic effects become more important when we are interested in determining effective control strategies or answering questions regarding recurrence and extinction of infections. It is known (see for example, Isham, 2005) that with a deterministic epidemic model with open population (i.e. allow for births, deaths, or/and immigration) if in the beginning of the epidemic $R_0 > 1$, the infection never completely dies out. In contrast, a stochastic epidemic model may fade out completely when it reaches a state where there is a single infective and moreover it is in any case to die out eventually unless there is an external source of infection

(Isham, 2005). Taking into account the above arguments, in this thesis, we will only focus on stochastic epidemic models.

2.1.3 Previous Work on Epidemic Modelling and Inference

There exists a comprehensive literature on deterministic and stochastic epidemic modelling and in this section we will only mention some of the main books on epidemic modelling and a series of papers which have drawn attention over the years. We have already mentioned in Section 2.1.2 the work in epidemic modelling in its early stages. Most such work prior to 1975 is contained in Bailey (1975) where the author presents an account of both deterministic and stochastic models. He also also illustrates the use of a variety of model using simulated data but also applications to real data.

Becker (1989) is mainly concerned with the statistical analysis of infectious disease data. The author deals with chain-binomial models with and without random effects as well as with other stochastic models in continuous time incorporating observable and latent infectious periods. He also makes use of the theory of stochastic processes and in particular, the theory of martingales to provide non−parametric methods of inference. Anderson and May (1991) model the spread of the disease for several situations and give many practical applications, but unlike Bailey (1975), they only focus on deterministic models. The six− months epidemics workshop which took place in 1993 in the Isaac Newton Institute in Cambridge, resulted in three collections of papers edited by Grenfel and Dobson (1995), Mollison (1995), Isham and Medley (1996).

In addition, the book by Daley and Gani (1999) offers an introduction to stochastic epidemic modelling as well as several historical remarks for both deterministic and stochastic models. Two recent books which have received considerable attention are i) the monograph by Andersson and Britton (2000) and ii) the book by Diekmann and Heesterbeek (2000). The former provides in the first part an

introduction to stochastic modelling while in the second, the authors discuss some basic statistical analysis for stochastic epidemic models. In the latter the authors focus their interest only in deterministic models which they apply in real data.

Although there is an extensive list of monographs on the modelling side of epidemics, however, there does not exist a monograph concerned with the progress over the years on the inference of stochastic epidemics. Becker and Britton (1999) present a nice review of statistical methodology for the analysis of outbreak data prior 1999. They also indicate that due to the increase of computing power the last two decades modern statistical methods offered a suitable framework for analysing effectively outbreak data using realistic models. There also exists a variety of review papers of epidemic models for particular diseases such as smallpox and Foot-and-Mouth; see for example Ferguson et al. (2003) and Keeling (2005). Since 1999 there have been many papers concerned with more complicated models than those described in Sections 2.1.2.1 and 2.1.2.2. However, we postpone the discussion of the work on such kind of models for Section 2.2.

2.1.4 The General Stochastic Epidemic Model (GSE)

In this section we describe the principles and the basic assumptions of the most well studied stochastic model, the−so−called *general stochastic epidemic* (GSE). We adopt a similar notation as the one adopted for the deterministic SIR model. A closed population (i.e. no births/deaths/imigration) of size $\mathcal{N}+a$ is considered and we assume that at time $t=0$ there are α initially infected individuals. The infectious periods of different individuals are independent and identically distributed according to some random variable D, which can have any arbitrary but specified distribution. In addition, we assume that the epidemic is observed up to a certain time, say T. Denote by $n_I \leq \mathcal{N}$ and $n_R \leq \mathcal{N}$, the number of individuals who got infected and removed by time T respectively. In general, $n_I \leq n_R \leq \mathcal{N}$.

The epidemic process (X_t, Y_t) is Markov if and only if the infectious period has the

lack-of-memory property. This is the special (Markovian) case where the infectious periods follow an Exponential distribution. Such a model is known as the *general stochastic epidemic* (GSE). Then, the process (X_t, Y_t) can be fully described in terms of continuous time Markov chains with the following transition rates:

$$\begin{aligned}(i,j) \to (i-1, j+1) &: \beta X_t Y_t \\ (i,j) \to (i, j-1) &: \gamma Y_t\end{aligned}$$

while the transition probabilities turn out to be:

$$\begin{aligned}\mathbb{P}[X_{t+\delta t} - X_t = -1,\ Y_{t+\delta t} - Y_t = 1 \mid \mathcal{H}_t] &= \beta \cdot X_t \cdot Y_t \cdot \delta t + o(\delta t) \\ \mathbb{P}[X_{t+\delta t} - X_t = 0,\ Y_{t+\delta t} - Y_t = -1 \mid \mathcal{H}_t] &= \gamma \cdot Y_t \cdot \delta t + o(\delta t) \\ \mathbb{P}[X_{t+\delta t} - X_t = 0,\ Y_{t+\delta t} - Y_t = 0 \mid \mathcal{H}_t] &= 1 - \beta \cdot X_t \cdot Y_t \cdot \delta t - \gamma \cdot Y_t \cdot \delta t + o(\delta t)\end{aligned}$$

where $\mathcal{H}_t$ is the sigma-algebra generated by the history of the process up to time t, i.e. $\mathcal{H}_t = \sigma\{(X_s, Y_s) : 0 \leq s \leq t\}$, with $\mathcal{H}_0 = \sigma\{X_0 = \mathcal{N}, Y_0 = \alpha)\}$ specifying the initial conditions. Therefore, the probability of an infection or a removal at the time interval $[t + \delta t)$ are $\beta X_t Y_t + o(\delta t)$ and $\gamma Y_t + o(\delta t)$ respectively. The correction term $o(\delta t)$ becomes negligible for small δt, i.e. $\frac{o(\delta t)}{\delta t} \to 0$ as $\delta t \to 0$.

The form of the transition probabilities show that the probability of infection at time t is proportional to the total number of infectives and susceptibles at time t. The constant of proportionality, β, is referred to as the *infection* rate. The transition probability of a removal shows that the length of the infectious periods are independent, identically distributed exponential random variables with mean $1/\gamma$, and therefore γ is referred as the *removal* rate for each individual. The epidemic continues until there are no more infected individuals left circulating in the population which will happen almost surely in finite time (Ball, 1983).

An extensive discussion of the properties of deterministic and stochastic versions of the SIR model is given by Bailey (1975); update for many variations of the

standard SIR model can be found for example in Andersson and Britton (2000) and Diekmann and Heesterbeek (2000). Furthermore, the dynamics of deterministic and stochastic SIR models in discrete time are analysed and compared in Allen and Burgin (2000).

2.1.5 Final Size of the Epidemic and The Basic Reproduction Number $[R_0]$

Before presenting any statistical issues which refer to the general stochastic epidemic model, we concentrate to the most important measures in stochastic epidemic modelling; the *final size* of epidemic and the *basic reproduction number* R_0. In this section we will briefly describe these useful epidemiological quantities.

2.1.5.1 Final Size Distribution

The final size the epidemic, say Z, is simply defined as the number of initially susceptible individuals that ultimately become infected. For $\theta \geq 0$, let $\phi(\theta) = \mathbb{E}[\exp\{-\theta D\}]$ be the moment generating function of the infectious period D and let p_k be the probability that the final size of the epidemic is equal to k, $0 \leq k \leq n$. Ball (1986) proved that

$$\sum_{i=1}^{l} \frac{\left(\binom{\mathcal{N}-k}{l-k}\right) p_k}{\left(\phi\left(\frac{\lambda(\mathcal{N}-l)}{\mathcal{N}}\right)\right)^{k+m}} = \binom{\mathcal{N}}{l}, \quad 0 \leq k \leq n \tag{2.2}$$

Note that an alternative definition of GSE assumes that during their infectious period, an individual makes contacts with each of the susceptibles at times given by the points of a homogeneous Poisson process with intensity $\lambda/\mathcal{N}$.

The system of equations in 2.2 is triangular in the p_k's and hence, in principle, it is easy to calculate the final size probabilities recursively, i.e. p_0, then p_1, p_2 and so on. Nevertheless, problems often occur in some specific circumstances such as extreme values for the parameters, even for small populations. When an Exponential

infectious period is assumed, Bailey (1975) has derived a different set of equations for the final size of probability that has better numerical stability. However, the Laplace transformation methods which are applied to the forward equations of the Markovian epidemic process do not generalize for a non-Markovian setup. Recently, Demiris and O'Neill (2006) employed multiple precision arithmetic to surmount this numerical problems. They also concluded that the branching process approximations as used to calculate the probability of an epidemic taking off was found to be effective, even for small numbers of initial susceptibles.

2.1.5.2 R_0 and the Threshold Result

The following definition is taken from Heesterbeek and Dietz (1996):

> *R_0 is the expected number of secondary infections produced by a typical infected individual during its entire infectious period in a population consisting of susceptibles only.*

In the GSE model, a *typical* individual can be any of the infectives since the model assumes homogeneous mixing and will, on average, be infectious for time $1/\gamma$. Then, the number of susceptibles infected by one infective per unit time is $\beta\mathcal{N}$. Hence the total number of infections produced by one infective, is equal to $\beta\mathcal{N}/\gamma$. In the case of the deterministic SIR model, the parameter γ can be interred as the reciprocal of the infectious period. In general, for an arbitrary infectious number, D, the basic reproduction ratio is defined as follows:

$$R_0 = \beta\mathcal{N} \cdot \mathbb{E}[D].$$

In more complicated models, the definition of R_0 is not straightforward and care is required to define an appropriate measure.

We shall describe why R_0 is such a significantly important measure in epidemics. First, recall the deterministic SIR model as presented in Section 2.1.2.1 and that

the number of infectives Y_t increases as long as the number of initial susceptibles in the population $x_0 = \mathcal{N}$ is greater than the quantity γ/β (Kermack and McKendrick, 1927). In other words, this is equivalent to the inequality that $R_0 > 1$. This reveals the significance of R_0. If $R_0 \leq 1$ then the latter condition cannot be met and therefore only a minor outbreak can result and R_0 is considered as *threshold parameter* . With an infection for which $R_0 > 1$ a population will be protected from epidemic outbreaks as long as the number of susceptibles is kept below the threshold by vaccination.

The threshold behavior of the stochastic SIR model for large populations (Whittle, 1955, Williams, 1971, Bailey, 1975, Andersson and Britton, 2000) is generally speaking analogous to that of the deterministic model. Intuitively, if the initial number of infectives, α, is small then during the early stages of the epidemic in a large population, essentially all the contacts of infectives are with susceptibles and a branching process approximation is appropriate (see also for example, Ball, 1983). We should make clear that such results are exactly valid only asymptotically, typically as the population size becomes infinite. Although the branching approximation idea has a long history, Ball and Donnelly (1995) used a coupling argument to investigate how the approximation improves as the population tends to infinity.

Specifically, in a population of infinitely many susceptibles, if $R_0 \leq 1$ then, with probability one, only a finite number of susceptibles will become infected (i.e. minor outbreak). If $R_0 > 1$ there is a positive probability that infinitely many susceptibles will become infected (i.e. major outbreak). We should bring to attention that, for finite populations, corresponding definitions of major and minor outbreaks are more difficult to define. Nevertheless, it is broadly true, an epidemic is either very likely will die out with minor impact or else might end up with a large proportion of susceptibles getting infected. Depending on the value of R_0 and whether is greater or smaller than one, then its value, approximately, will indicate which of the two situations is more likely. Therefore, it now becomes clear why R_0

is so important in epidemic theory since also implies the amount of effort needed to prevent an epidemic.

2.1.6 The Likelihood of GSE for Different Model Setups

As it has been already mentioned, one reason for modelling epidemics is to draw conclusions about particular diseases. In this section, we examine the important area of drawing inference about the model parameters. A formal statistical analysis has an essential role to play in bridging the gap between the mathematical theory and public health. Statistical inference uses the *likelihood function*.

In this section, we describe in detail the various model setups for the GSE which have been used throughout the literature. For each of the presented model setups we derive the likelihood of the observed data given the (unknown) parameters β and γ. First, we refer to the setup adopted by many researchers (see, for example, Bailey and Thomas, 1971, O'Neill and Roberts, 1999). Then, we review the the considerable amount of work on epidemics related with martingales which has been presented over the last two decades (Becker, 1989). Finally, we present an alternative setup which has been recently proposed in the literature (Britton and O'Neill, 2002, Neal and Roberts, 2005).

2.1.6.1 Bailey and Thomas' Setup

We adopt the following notation by letting $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_{n_R})$, where $\tau_1 = 0$, to denote the (ordered) successive removal times observed during $[0, T]$. In other words, τ_i refers to the ith removal time. Denote by ϕ_1 the initial infection time and $\boldsymbol{\phi} = (\phi_2, \ldots, \phi_{n_I})$ the remain successive infection times during (ϕ_1, T). If the initial infective does not manage to infect any other susceptible by the time is removed (i.e. τ_1) then the epidemic is ceased. Therefore, in order to obtain

epidemics with $n_I \geq 2$ the following constraints are imposed:

$$\phi_{i-1} < \phi_i < \tau_{i-2} \quad \text{for} \quad i = 3, \ldots, n_I. \tag{2.3}$$

Because of the homogeneous mixing, the epidemic can be fully described by tracking the number of infected and removed individuals at each time point without the need of knowing which individual got infected or removed.

If we consider the order of the successive events which occur during the epidemic (infections or/and removals) and take into account the memoryless property of the exponential distribution, we can easily derive the likelihood as follows:

$$f(\boldsymbol{\tau}, \boldsymbol{\phi}|\beta, \gamma, \phi_1) \propto \prod_{j=1}^{n_R} \gamma Y_{\tau_j^-} \cdot \prod_{j=2}^{n_I} \beta X_{\phi_j^-} Y_{\phi_j^-} \cdot \exp\left\{ - \int_{\phi_1}^{T} (\beta X_t Y_t + \gamma Y_t)\, dt \right\} \tag{2.4}$$

where the notation ϕ_j^- denotes the left hand limit, so for example $Y_{\phi_j^-}$ denotes the $\lim_{\uparrow \phi_j}(Y_s)$, or in other words the time immediately prior to ϕ_j. Note that the products $\prod_{j=1}^{n_R} \gamma Y_{\tau_j^-}$ and $\prod_{j=2}^{n_I} \beta X_{\phi_j^-} Y_{\phi_j^-}$ depend on the infection times $\phi_1, \ldots, n_I$ due to the terms $X_{\phi_j -}$ and $Y_{\phi_j^-}$. This form of the likelihood in (2.4) is also given in Britton and O'Neill (2002), O'Neill and Becker (2001), O'Neill and Roberts (1999).

2.1.6.2 A Setup Based on Martingales

Becker (1989), Rida (1991) and Andersson and Britton (2000) have used a setup based on counting processes using the theory of martingales. Their approach is based on constructing a suitable martingale from the epidemic process $\{(X_t, Y_t) : t \geq 0\}$ and use the powerful tools from martingale theory to draw inference for the parameters of interest.

Suppose that a random process is followed continuously over time beginning at time $t = 0$. Denote by $\mathcal{H}_t$ the history of the process up to time t. A *martingale* is

a random process $M = \{M_t, t \geq 0\}$ such that for every $t \geq 0$:

1. the value of M_t is determined by $\mathcal{H}_t$;
2. $\mathbb{E}[|M_t|] < \infty$;
3. $\mathbb{E}[(M(u) \mid \mathcal{H}_t] = M_t$, for all $u \geq t$.

The first and the second property of a martingale are always satisfied. The former because of the model specification and the latter because we deal with the epidemics in finite populations and with finite infection rates (Becker, 1989). We refer to the third one as the martingale property which requires that the expected change, over time, in the value of a martingale is always unbiased. It is easy to see that $\mathbb{E}[M_t \mid H_0] = M_0$ for all $t \geq 0$. When $M_0 \equiv 0$ we have $\mathbb{E}[M_t] = 0$ for every $t \geq 0$ and we refer to M as a zero mean martingale.

We are interested in martingales which arise from counting processes. A counting process $N = \{N_t, t \geq 0\}$ is a random process which counts the occurrence of certain events over time, N_t being the number of events occurring in the time interval $(0, t]$. We set $N_0 = 0$ and take N to be continuous on the right as its jump point. Denote by $\mathrm{d}N_t$ the number of events occurring in the time interval $(t + \mathrm{d}t)$ and by $\mathcal{H}_t$ the history of N and other processes up to time t. In stochastic epidemic models we are often concerned with those counting processes:

$$\begin{aligned} \mathbb{P}[\mathrm{d}N_t = 1 \mid \mathcal{H}_t] &= A_t \, \mathrm{d}t \\ \mathbb{P}[\mathrm{d}N_t = 0 \mid \mathcal{H}_t] &= 1 - A_t \, \mathrm{d}t \end{aligned}$$

where A_t is the intensity process of N which is often a random process. Denote by $N_t = X_0 - X_t$ the number of individuals infected in $[0, t)$ and $R_t = N - X_t - Y_t$ which holds for any t. At this stage, we assume that the initial condition is known, i.e. the total number of initially infected and susceptibles, $X_0 = x_0$, $Y_0 = y_0$, $R_0 = r_0$.

Then, the intensity process of N is given by:

$$A_t = \beta_t(y_0 + N_t)(x_0 - N_t).$$

Proposition 1 *Define the process M, by $M_t := N_t - \int_0^t A_x dx$. This is a zero-mean martingale with respect to the history $\mathcal{H}$.*

Proof :

First note that $M_0 = 0$. Let x be a value such that $t < x < u$, then

$$\begin{aligned} \mathbb{E}[\mathrm{d}N_x \mid \mathcal{H}_t] &= \mathbb{E}[\mathbb{E}[\mathrm{d}N_x \mid \mathcal{H}_x] \mid \mathcal{H}_t] \\ &= \mathbb{E}[A_x \mathrm{d}x \mid \mathcal{H}_t] \end{aligned}$$

Divide by $\mathrm{d}x$ and letting $\mathrm{d}x \to 0$ we get:

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}[N_x \mid \mathcal{H}_t] = \mathbb{E}[A_x|\mathcal{H}_t]$$

If we integrate both sides with respect to x from t to u:

$$\int_t^u \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}[N_x \mid \mathcal{H}_t]\mathrm{dx} = \int_t^u \mathbb{E}[A_x \mid \mathcal{H}_t]\mathrm{d}x$$

$$\mathbb{E}[N_u|\mathcal{H}_t] - N_t = \mathbb{E}\left[\int_t^u A_x \mathrm{d}x|\mathcal{H}_t\right]$$

which gives the desire result

$$\mathbb{E}[M_u \mid \mathcal{H}_t] = M_t.$$

□

Becker (1989) and Rida (1991) express the infection rate as $\beta = \lambda/\mathcal{N}$ for some $\lambda > 0$. In other words we could look at the proportion of susceptibles at time t, $\widetilde{X_t} = X_t/\mathcal{N}$. This allows us to interpret the infections occurring, as points

of a homogeneous Poisson process with rate β. The transition probabilities with respect to β can be written as follows:

$$\begin{aligned}
\mathbb{P}(\mathrm{d}N_t = 1,\ \mathrm{d}R_t = 0 \mid \mathcal{H}_t) &= \beta \cdot \widetilde{X_t} \cdot Y_t\ \mathrm{d}t + o(\mathrm{d}t) \\
\mathbb{P}(\ \mathrm{d}N_t = 0,\ \mathrm{d}R_t = 1 \mid \mathcal{H}_t) &= \gamma \cdot Y_t\ \mathrm{d}t + o(\ \mathrm{d}t) \\
\mathbb{P}(\ \mathrm{d}N_t = 0,\ \mathrm{d}R_t = 0 \mid \mathcal{H}_t) &= 1 - \beta \cdot \widetilde{X_t} \cdot Y_t\ \mathrm{d}t - \gamma \cdot Y_t\ \mathrm{d}t + o(\ \mathrm{d}t)
\end{aligned}$$

Assume that a realization of the general epidemic is completely and continuously observed up to its end. Then by using counting process theory (eg Andersen et al., 1993) we obtain the log−likelihood, also given in Becker and Britton (1999, Section 2.1.2), as follows:

$$\log L(\beta, \gamma) \propto \int_0^T \left(\log \{\beta \widetilde{X_u} Y_u\}\ \mathrm{d}N_u - \beta \widetilde{X_u} Y_u + \log \{\gamma Y_u\}\ \mathrm{d}R_u - \gamma Y_u\ \mathrm{d}u \right) \quad (2.5)$$

2.1.6.3 An Alternative Setup

Britton and O'Neill (2002) and Neal and Roberts (2005) have adopted a rather different setup which we shall describe in this section. We label the individuals who got infected during the epidemic as $i = 1, \ldots, n_I$ and those we did not as $i = n_I + 1, \ldots, \mathcal{N}$. We assign to each of them their infection (I_i) and removal (R_i) time respectively, assuming that $I_i = \infty$, for $i = n_I + 1, \ldots, \mathcal{N}$, for the individuals who did not get infected during the epidemic. We label the initial infective k, such that $I_k < I_j$ for all $j \neq k$. We proceed with the following definition of the *infectious pressure* that a susceptible individual gets from the current infectives:

Definition 1 *A susceptible individual j when it becomes infected gets individual-specific infectious pressure β from (an infected) individual i if and only if*

$$I_i < I_j < R_i$$

Therefore the **total** *infectious pressure which is subjected to individual j when it becomes infected is equal to P_j:*

$$P_j = \sum_{i \in \mathcal{Y}_j} \beta$$

where $\mathcal{Y}_j = \{i : I_i < I_j < R_i\}$

The likelihood can be broken into two independent parts: i) the infectious L_1 and the ii) removal L_2 part.

Infectious part: Denote by S the total person-to-person infectious pressure during the course of the epidemic:

$$\begin{aligned} S &= \sum_{i=1}^{n_I} \sum_{j=1}^{\mathcal{N}} \beta \left((R_i \wedge I_j) - (I_j \wedge I_i) \right) \\ &= \beta \sum_{i=1}^{n_I} \sum_{j=1}^{\mathcal{N}} \left((R_i \wedge I_j) - (I_j \wedge I_i) \right) \\ &= \beta \cdot A \end{aligned}$$

where $A = \sum_{i=1}^{n_I} \sum_{j=1}^{\mathcal{N}} (R_i \wedge I_j - I_i \wedge I_j)$. The infection component can be then written:

$$L_1 = \prod_{i=1, i \neq k}^{n_I} \left(\sum_{j \in \mathcal{Y}_i} \beta \right) \times \exp\{-S\} \tag{2.6}$$

Removal part: The infectious period for an individual, i say, is $R_i - I_i$. The contribution to the likelihood is given by this exponential: $\gamma \exp\{-\gamma(R_i - I_i)\}$. So, if we consider every individual who got infected, we then get L_2:

$$L_2 = \prod_{i=1}^{n_R} \gamma \exp\{-\gamma(R_i - I_i)\} \tag{2.7}$$

Combining (2.6) and (2.7) we get the likelihood of the data given the model pa-

rameters:

$$\begin{aligned} f\left(\boldsymbol{I},\boldsymbol{R}|\beta,\gamma\right) &= L_1 \times L_2 \\ &= \prod_{i=1,i\neq k}^{n_I}\left(\sum_{j\in\mathcal{Y}_i}\beta\right)\times\exp\left\{-\beta A\right\} \\ &\times \gamma^{n_R}\exp\left\{-\sum_{i=1}^{n_R}\gamma(R_i-I_i)\right\} \end{aligned} \tag{2.8}$$

It is easy to check that (2.4) and (2.8) are identical. This alternative setup was used in Britton and O'Neill (2002) and Neal and Roberts (2005). The great advantage of it, is that it allows us to write the integral $\int_{I_k}^{T} X_t Y_t \, dt$ as $\sum_{i=1}^{n_I}\sum_{j=1}^{\mathcal{N}}(R_i \wedge I_j - I_j \wedge I_i)$. From a practical point of view, in order to evaluate that integral, it needs to be discretised by a transformation into a sum over the successive events of the epidemic whilst such a discretisation is substituted by the double sum S using the alternative setup.

Comparisons to the Bailey and Thomas' setup

There exist many differences between the alternative and the Bailey and Thomas' setup. We will refer to the latter as the 'original' setup.

First, the main difference relies on the fact that within the alternative setup each individual in the population are labeled and are also associated with them are the corresponding infection and removal times. In other words, each pair (I_i, R_i) refers to the infection and removal time of the individual labeled as i. On the other hand, within the context of the original setup, the individuals are not labeled at all and each pair (ϕ_i, τ_i) refers to the i_{th}, in sequential order, infection and removal time respectively.

Moreover, the alternative setup is necessary when modelling heterogeneously mixing populations, while the original setup is only suitable for homogenously mixing populations. Finally, the alternative setup allows us to implement non−centered reparameterisations which are introduced in Sections 2.4 and 2.5.

2.1.7 Likelihood−Based Inference for Complete Data

Having obtained the likelihood of the observed data given the model parameters, considering three different model setups, we first show how the frequentist approach provides estimates for the infection and removal rate of the GSE as well as with their corresponding standard errors. Then, we also adopt a Bayesian approach since it allows coherent incorporation of prior information which can be either experts' opinion or information on previous disease outbreaks. Inference for the basic reproduction number, R_0, is also presented in this section within the classical and Bayesian framework. Note, that throughout this section we make the unrealistic assumption that we fully observe the epidemic, i.e. the actual infection and removal times of the individual as well as the order of these events.

2.1.7.1 The Classical Approach

Given a model setup, by differentiating the likelihood (either 2.4 or 2.5 or 2.8) it is straightforward to derive *maximum likelihood estimates* (MLE) for the parameters of interest, β and γ (see also Becker, 1989, Chapter 7.3)

$$\widehat{\beta} = \frac{n_I - 1}{\int_{I_1}^{T} X_t Y_t \, dt} \quad \text{or} \quad \widehat{\beta} = \frac{n_I - 1}{\sum_{i=1}^{n_I} \sum_{j=1}^{\mathcal{N}} (R_i \wedge I_j - I_j \wedge I_i)} \tag{2.9}$$

$$\widehat{\gamma} = \frac{n_R}{\int_{I_1}^{T} Y_t \, dt} \quad \text{or} \quad \widehat{\gamma} = \frac{n_R}{\sum_{i=1}^{n_R} (R_i - I_i)} \tag{2.10}$$

Rida (1991) derives asymptotic normality and consistency for the estimators given in (2.9) and (2.10) as the size of the total population N tends to ∞. Note that these estimators also correspond to the results from Andersson and Britton (2000, Section 9.2). By differentiating the first derivative of the likelihood with respect to β and γ, the corresponding standard errors turn out to be:

$$\sigma_{\hat{\beta}} = \frac{\hat{\beta}}{\sqrt{n_I - 1}} \tag{2.11}$$

$$\sigma_{\hat{\gamma}} = \frac{\hat{\gamma}}{\sqrt{n_R}}. \tag{2.12}$$

Having obtained estimates and standard errors of the MLE estimates we are also able to obtain (approximate) confidence intervals by Normal approximation.

Estimating The Basic Reproduction Ratio $[R_0]$

We have already seen how crucial the basic reproduction number R_0 for an epidemic is (see Section 2.1.5.2) and therefore we are interested in providing inference for it. Becker (1989) proposed a martingale method of moments to derive a formula to estimate R_0 where only final size data are available in the case of homogeneous mixing, i.e. common infection rate β as in the SIR model:

$$\hat{R_0} = \frac{1}{n_R} \sum_{i=\mathcal{N}-1}^{\mathcal{N}-n_I+1} \frac{1}{i} \tag{2.13}$$

and the corresponding error turns out to be:

$$\hat{\sigma}_{R_0} = \frac{1}{n_R} \sqrt{\left(\sum_{i=\mathcal{N}-1}^{\mathcal{N}-n_I+1} \frac{1}{i^2} \right) - n_R \cdot \hat{R_0}^2} \tag{2.14}$$

Nevertheless, Becker (1989) does not deal in much detail with the extreme case that everybody gets infected during the epidemic, i.e. $n_I = N$. However, Hoehle (2003) proposed to insert a correction term due to infectivity wasted as nobody is susceptible anymore.

It is interesting to see that the estimation of R_0 relies only on final size data and does not depend on the infection times. So even, if the infections are not observed which usually the case then inference for R_0 can be still drawn. Intuitively, (see also the definition of R_0 in Section 2.1.5), if there is a large outbreak, in the sense that

a large proportion of the initially susceptible individuals contracted the disease, then even with the absence of temporal data (i.e. infection/removal times), we could guess that R_0 should be quite large. On the other hand, if there is a minor outbreak, then R_0 will be quite small.

2.1.7.2 The Bayesian Approach

A Bayesian approach can be adopted to incorporate such available information before seeing the data. We adopt the setup by Bailey and Thomas (1971) (see Section 2.1.6.1) and assign (conjugate) Gamma prior distributions for β and γ with the following probability density functions:

$$\begin{aligned} \beta &\sim Ga(\lambda_\beta, \nu_\beta) \\ \pi(\beta) &\propto \beta^{\lambda_\beta - 1} \exp\{-\lambda_\beta \beta\} \end{aligned}$$

and

$$\begin{aligned} \gamma &\sim Ga(\lambda_\gamma, \nu_\gamma) \\ \pi(\gamma) &\propto \gamma^{\lambda_\gamma - 1} \exp\{-\lambda_\gamma \gamma\} \end{aligned}$$

We apply Bayes theorem by multiplying the priors and the likelihood and get the posterior distribution of β and γ given the data (infection and removal times):

$$\begin{aligned} \pi(\beta, \gamma | \boldsymbol{I}, \boldsymbol{R}) &\propto \beta^{\lambda_\beta + n_I - 2} \exp\left\{-\beta \left(\int_{\phi_1}^{T} X_t Y_t \, \mathrm{d}t + \nu_\beta\right)\right\} \\ &\times \gamma^{\lambda_\gamma + n_R - 1} \exp\left\{-\gamma \left(\int_{\phi_1}^{T} Y_t \, \mathrm{d}t + \nu_\gamma\right)\right\} \end{aligned} \tag{2.15}$$

The two parameters are *a posteriori* conditionally independent and therefore their posterior distributions are given by:

$$\pi(\beta | \boldsymbol{I}, \boldsymbol{R}) \equiv Ga\left(\lambda_\beta + n_I - 1, \nu_\beta + \int_{\phi_1}^{T} X_t Y_t dt\right) \tag{2.16}$$

$$\pi(\gamma|\boldsymbol{I},\boldsymbol{R}) \equiv Ga\left(\lambda_\gamma + n_R, \nu_\gamma + \int_{\phi_1}^{T} Y_t dt\right) \tag{2.17}$$

Having obtained $\pi(\beta|\boldsymbol{I},\boldsymbol{R})$ and $\pi(\gamma|\boldsymbol{I},\boldsymbol{R})$ it is very straightforward to construct credibility intervals or get point estimates such as medians and means. Identical results will be obtained by adopting a different model setup such as those in Sections 2.1.6.2 and 2.1.6.3.

Drawing Inference for The Basic Reproduction Ratio $[R_0]$

The basic reproduction ratio R_0 is defined as $\beta\mathcal{N}/\gamma$. Therefore we could obtain its posterior distribution by transformation. On the other hand, an easier approach, is to draw samples from each of the posteriors (2.16) and (2.17) and by dividing the two samples and then we get posterior samples of $\pi(R_0|\boldsymbol{I},\boldsymbol{R})$.

2.1.8 Inference for Partially Observed Epidemics

In general, inference problems for disease outbreak data are complicated and often their statistical analysis requires the development of problem-specific methodology. There are many reasons that make such an analysis awkward. First, there are often various levels of inherent dependence that we should take into account. Naturally, the more realistic the model is, the more complex becomes and its analysis gets harder. For instance, if we assume an epidemiologically plausible distribution for the infectious periods, such as Gamma or Weibull instead of the mathematically convenient Exponential, an additional level of dependence is induced. Furthermore, although it is relatively easy to define a stochastic epidemic model, there is often a very large number of ways that can result in the same outcome. The above statements are in contrast with the usual independence assumption that underlies many standard statistical methods.

One of the most difficult problems that needs to be overcome when analysing

disease outbreak data, is the fact that such data are incomplete in many different ways. In general, the epidemic process is rarely fully observed. From an inference point of view, it would be desirable to observe the times at which an individual gets infected (infection time), who infected them, as well as the times at which the individual ended its infectious period (removal time). In practice, only the times at which an individual is detected are observed and seldom the infection times are known. Moreover, it is often the case, that neither the infection nor the removal times are available and only the number of the individuals who contracted the disease out of the total size of the initial susceptible population is known. Such data, are usually routinely collected surveillance data.

This section reviews the existing methods of how the classical and the Bayesian approach tackle the lack of detailed observed data to make statistical inference about the infection and the removal rate of the GSE feasible.

2.1.8.1 The Classical Approach Based on Martingale Methods

We have already showed that in terms of a fully observed epidemic and using counting process theory (e.g. Andersen et al., 1993), we can derive MLEs (see Section 2.1.7). However, in this section we focus on estimating the infection (β or λ) and the removal rate (γ) in the absence of the infection times. When the epidemic is partially observed, then the likelihood cannot be written in closed form (Becker and Britton, 1999). Bailey (1975, pg.118) provided methods of estimation based on approximations of certain recursive formulae defining the likelihood or rely on large population approximation.

On the other hand, tools from martingale theory can be used with the method of moments to obtain estimates with explicit expression (Becker, 1979). By making use of the optional sampling theorem, Becker (1979) provides an estimator for the infection rate β:

$$\widehat{\beta}_T = \frac{N \sum_{i=1}^{N_T} 1/(N+1-i)}{\int_0^T \mathbb{I}\{X_s^- > 0\} Y_{s^-} \, \mathrm{d}s} \tag{2.18}$$

where X_s and Y_s denote the number of susceptibles and infectives at time s.

Rida (1991) provides an asymptotically equivalent estimator of (2.18). The numerator can be approximated by $-N \log 1 - N_T/N$ for large N. In addition, if at least one susceptible remains at the end of the epidemic, i.e. $X_T > 0$ then

$$\int_0^T \mathbb{I}\{X_s^- > 0\} Y_{s^-} \, \mathrm{d}s = \sum_{i=1}^{N_T} V_{N(i)} + \sum_{j=1}^{\mathrm{a}} V_{-j}$$

where $V_{N(i)}$ are the lengths of the infectious periods of the N_T infected individuals and the V_{-j} are the infectious periods of the a initial infectives and get

$$\widetilde{\beta} = \frac{-N \cdot \log\left\{1 - \frac{N_T}{N}\right\}}{\sum_{i=1}^{N_T} V_{N(i)} + \sum_{j=1}^{\mathrm{a}} V_{-j}} \tag{2.19}$$

As Rida (1991) points out, when only the removal times are observed then neither (2.18) nor (2.19) can be used and the inference for β (or λ) becomes very hard (if not impossible). However, if we let

$$\widehat{E(V)} = \frac{\sum_{i=1}^{N_T} V_{N(i)} + \sum_{j=1}^{\mathrm{a}} V_{-j}}{N(T) + \mathrm{a}}$$

be an estimate of the mean duration of the infectious period, then:

$$\widehat{\mu}(\tau) = \widehat{\lambda}(\tau)\widehat{E(V)} = \frac{-\log\left(1 - N(T)/\mathcal{N}\right)}{(N(T) + \mathrm{a})/\mathcal{N}} \tag{2.20}$$

An approximate variance of this estimator is given as follows:

$$\mathrm{var}(\widehat{\mu(T)}) \approx \frac{\mathcal{N}}{N(T)(\mathcal{N} - N(T))} + \frac{(\widehat{\mu}(T))^2}{N(T) + \mathrm{a}} \tag{2.21}$$

Note that $\widehat{\mu(T)}$ provides an estimate for the basic reproduction number R_0 while $\mathrm{var}(\widehat{\mu(T)})$ represents its approximate variance. Such an estimate is a similar but an alternative to the one described in Section 2.1.7.1.

Although, estimation of R_0 is important, inference for the infection or the re-

moval rate is also of interest. Becker and Britton (1999) present the idea of back-projection which is now widely used and known as data augmentation technique. They observe that, if there is a way from the available data (removal process / removal times) to construct the unobserved realization infection process, we can then plug in the reconstructed process into the explicit expressions (2.9) and (2.10) and get the maximum likelihood estimates.

We assume that infections occur according to a non-homogeneous Poisson process with intensity λ_t at time t. Then

$$\mathbb{E}(N_t) = \int_0^t \lambda_x \, \mathrm{d}x = \Lambda_t.$$

Hence, if we consider that at first stage N_t can be approximated by its expectation, $\mathbb{E}(N_t)$, then an estimate $\widehat{\Lambda}_t$ of Λ_t for all t, will allows us to (approximately) estimate the unobserved process N_t. Then, X_t, Y_t can be reconstructed by $X_t = X_0 - \widehat{\Lambda}_t$ and $Y_t = Y_0 + \widehat{\Lambda}_t - R_t$. The removal process is independent of the infection process and has intensity

$$\mu_t = \int_0^t \lambda_{t-u} \, \mathrm{d}F_D(u) \tag{2.22}$$

where D is the duration of the infectious period and F_D is the corresponding distribution function. In general, is assumed that $\{R_t\}$ is an observed Poisson process with intensity function μ_t and $F_D(u)$ is assumed to be known from past studies. Becker and Britton (1999, Section 7.2) indicate that (2.22) is the basis of the method of the back-projection and refer to the several approaches which have been used to obtain an estimate of λ_t. Some of them are basically based on assumptions of some of the quantities of interest to be known, as in Becker and Hasofer (1997) where they reconstruct the infection process Y_t assuming the removal rate γ is given.

Concluding, the review paper by Becker and Britton (1999) also makes the point that the proposal to reconstruct the infection process and the use of the "complete" data to draw inference for the parameters, is in the spirit of data augmentation

methods which are appropriate for missing data problems. The Bayesian analysis and the recent advances of Markov chain Monte Carlo methodology offer a natural framework to analyze data which fall in this context (eg Tanner and Wong, 1987) without the need of making any kind of approximation or unrealistic assumptions.

2.1.8.2 The Bayesian Approach using MCMC methods

The first approaches on MCMC methods for stochastic epidemic models are in O'Neill and Roberts (1999) and Gibson and Renshaw (1998). In this section we fully describe how we can apply Bayesian inference and Markov Chain Monte Carlo algorithms to draw inference for the parameters of interest. By the time T when we observe the epidemic, either it might have ceased or it will be still in progress. Therefore, if the epidemic is still in progress there might have been infected individuals which we haven't observed by that time T.

We adopt the setup by Bailey and Thomas (1971) (see Section 2.1.6.1) and therefore the infection times $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{n_I})$ are treated as unknown parameters which need to be imputed. Then, a prior for the initial infection time needs to be specified, as well as for the β and γ. Following O'Neill and Roberts (1999) we assign (conjugate) Gamma priors with parameters $(\lambda_\beta, \nu_\beta)$ and $(\lambda_\gamma, \nu_\gamma)$ respectively. Note that *a-priori* the model parameters (β and γ) are independent. Recall that we set the first removal time $\tau_1 = 0$, and therefore first, for the initial infection time $-\phi_1$ an Exponential prior with mean $1/\delta$ is assumed and secondly $\phi_2 < 0$ otherwise the epidemic will cease.

$$\begin{aligned} \pi(\beta) &\sim Ga(\lambda_\beta, \nu_\beta) \\ \pi(\gamma) &\sim Ga(\lambda_\gamma, \nu_\gamma) \\ -\phi_1 &\sim \text{Exp}(\delta) \end{aligned}$$

By adopting the original setup described in Section 2.1.6.3 and multiplying the likelihood in (2.4) and the priors we get the posterior distribution of the parameters

and the missing data given the removal times:

$$\begin{aligned}
\pi(\beta,\gamma,\boldsymbol{\phi}|\boldsymbol{\tau}) &\propto \pi(\beta)\cdot\pi(\gamma)\cdot\pi(-\phi_1)\cdot L(\boldsymbol{\phi},\boldsymbol{\tau}|\beta,\gamma) \\
&\propto \prod_{i=1}^{n_R}\gamma Y_{\tau_i^-}\cdot\prod_{j=2}^{n_I}\beta X_{\phi_j^-}Y_{\phi_j^-}\cdot\exp\left\{-\int_{\phi_1}^{T}(\beta X_t Y_t+\gamma Y_t)\,dt\right\} \\
&\times \beta^{\lambda_\beta-1}\exp\{-\beta\nu_\beta\}\times\gamma^{\lambda_\gamma-1}\exp\{-\gamma\nu_\gamma\} \qquad (2.23)
\end{aligned}$$

It is straightforward to derive the full conditional posterior distributions for each of the model parameters and the initial infection time:

$$\begin{aligned}
\pi(\beta|\gamma,\boldsymbol{\phi},\boldsymbol{\tau}) &\equiv Ga\left(\lambda_\beta+n_I-1,\nu_\beta+\int_{\phi_1}^{T}X_tY_tdt\right) \\
\pi(\gamma|\beta,\boldsymbol{\phi},\boldsymbol{\tau}) &\equiv Ga\left(\lambda_\gamma+n_R,\nu_\gamma+\int_{\phi_1}^{T}Y_tdt\right) \\
\pi(-\phi_1|\beta,\gamma,\boldsymbol{\phi}_{-1},\boldsymbol{\tau}) &\equiv \mathrm{Exp}\left(\beta\mathcal{N}+\gamma+\delta\right) \\
\pi(\phi_j|\boldsymbol{\phi}_{-j},\beta,\gamma,\phi_1,\boldsymbol{\tau}) &\propto \prod_{i=1}^{n}Y_{\tau_j^-}\prod_{j=2}^{m}X_{\phi_j^-}Y_{\phi_j^-}\exp\left\{-\int_{\phi_1}^{T}(\beta X_tY_t+\gamma Y_t)\,dt\right\}
\end{aligned}$$

O'Neill and Roberts (1999) pointed out that sampling directly from $\pi(\phi_j|\boldsymbol{\phi}_{-j},\boldsymbol{\tau},\beta,\gamma)$ for $j=2,\ldots,n_I$ is problematical and therefore they proposed a Metropolis Hastings step instead. In general, the following MCMC scheme can be applied.

Metropolis within Gibbs Sampling Scheme

(Repeat the following steps)

1. Start the chain with initial values: $\beta^0, \gamma^0, \boldsymbol{\phi}^0$;
2. Update β by using Gibbs Sampler and drawing from $\pi(\beta|\gamma, \boldsymbol{\phi})$;
3. Update γ by using Gibbs Sampler and drawing from $\pi(\gamma|\beta, \boldsymbol{\phi})$;
4. Update ϕ_1 by using Gibbs sampler and drawing from $\pi(\phi_1|\beta, \gamma, \boldsymbol{\phi}_{-1})$;
5. Choose one of the three moves with equal probability:
   - Move an infection time, or
   - Remove an infection time, or
   - Add a new infection time.

Steps 2,3 and 4 are standard Gibbs sampler updates while step 5 needs more discussion. If we assume that the epidemic is still in progress then we can:

- *Add a new infection time*:

  We propose to add a new infection time, say ϕ_s by proposing from $U(\phi_1, T)$. Add this infection time to the set of infections $(\boldsymbol{\phi})$ with probability:

$$\left\{ \frac{\pi(\boldsymbol{\phi}, \phi_s|\phi_1, \beta, \gamma)}{\pi(\boldsymbol{\phi}|\phi_1, \beta, \gamma)} \frac{(T - \phi_1)}{n_I + 1} \wedge 1 \right\}$$

- *Remove a infection time*:

  We choose uniformly an infection time ϕ_s from the current list

of infection times and remove it from the set of infections ($\boldsymbol{\phi}$) with probability:

$$\left\{\frac{\pi(\boldsymbol{\phi}_{-s},|\phi_1,\beta,\gamma)}{\pi(\boldsymbol{\phi}|\phi_1,\beta,\gamma)}\frac{n_R}{(T-\phi_1)}\wedge 1\right\}$$

- *Move an infection time*:

  We choose uniformly one of the existing infection times, ϕ_s and we propose a replacement candidate, ϕ_s' sampled uniformly on (ϕ_1, T) and is accepted with probability:

$$\left\{\frac{\pi(\boldsymbol{\phi}_{-s},\phi_s'|\phi_1,\beta,\gamma)}{\pi(\boldsymbol{\phi}|\phi_1,\beta,\gamma)}\wedge 1\right\}$$

Note that if the epidemic is known to be complete then the only move that is allowed is the last one, since the number of the infections must be always n_I. In addition, the same proposal in order to update (move) an infection time has also been in used in Britton and O'Neill (2002).

Applying the above MCMC algorithm, we are in a position to draw samples from the marginal posterior distributions of the model parameters $\pi(\beta|\boldsymbol{\tau})$ and $\pi(\gamma|\boldsymbol{\tau})$ and obtain point estimates or/and credibility intervals. In the following section we are considering extensions of the GSE model and Bayesian methods of drawing inference the parameter of interest.

In an ideal world, we would have complete observation of a single epidemic or even better, multiple replication of it. As we have already discussed, more often, the available information is incomplete. The classical framework provides estimates for the basic reproduction number R_0 rather for the infection and the removal rate. On the other hand, the idea of "back−projection (Becker and Britton, 1999) also recently known as *data−augmentation*, fit more naturally within the Bayesian framework without the need of any impractical assumptions.

2.1.9 Discussion

So far we have described in detail the GSE model, the most well studied epidemic model. Apart from discussing its properties, researchers have also concentrated on its limitations and assumptions in order to derive more realistic models. The characterisation as "general", which was given to the GSE model, seems now inappropriate since the model has, over the years, been generalized in many ways. In this section we refer to some of the work which is related with extensions of the GSE.

A significant amount of effort has been put for understanding the spread of childhood infections, especially measles. In a basic measles model, an extra state, the *latent* period is usually included in the standard SIR to obtain the SEIR model. The individuals in the latent state (i.e. exposed) are infected but not yet infectious such that they can infect other susceptibles. Such a model seems more appropriate than the SIR when the incubation period of an individual is very long.

A crucial assumption of the standard SIR model is the one which refers to a closed population. However, it often becomes improper to assume such a population when an epidemic lasts for long period and changes in the population occur. The properties of the SIR model with demography have been studied much in the literature (see for example, Bailey, 1975, Andersson and Britton, 2000). However, results for such model are more difficult to obtain than in the SIS model (Isham, 2005).

The GSE model and other variations of the SIR theme have a simple and relatively tractable mathematical structure. The assumption of an exponentially distributed infectious period is not epidemiologically motivated, although it makes the statistical and the probabilistic analysis simpler. For instance, assuming an Exponential infectious period, using Markov process theory we can obtain deterministic and diffusion approximations for the whole trajectory, which are valid for large population sizes. Andersson and Britton (2000, Chapter 5) present such results which can

be mainly found in Ethier and Kurtz (1986, Chapter 11). In principle, any distribution for modelling the infectious period of an individual which can be described by its Laplace transform can be used (Ball, 1986); see applications on diseases with Gamma (O'Neill and Becker, 2001) and Weibull (Streftaris and Gibson, 2004) distributed infectious periods. More specifically, researchers have also discussed the effect of a non−Exponential distribution on the persistence of measles (Keeling and Grenfell, 1998, 2000, Lloyd, 2001)

Concluding, the main characteristic assumption of GSE refers to homogeneously mixing models. We argue in the following section why such models are not always appropriate and the need of extending to non−homogeneously mixing models is essential.

2.2 Heterogeneously Mixing Stochastic Epidemic Models (HMSE)

We have mostly concentrated so far on models for populations of homogeneously mixing hosts. However, such an assumption is often not realistic for a variety of applications. Therefore, it is crucial for applied purposes to incorporate sources of heterogeneity, without dismissing the aim of building a parsimonious model. Often, we distinguish intrinsic heterogeneity of the individuals, for example a variation in the susceptibility due to genetics, from heterogeneity of mixing where the infection rate between individuals depends on their distance in the sense that an infected individual is more likely to infect those susceptibles who are close to it rather those who are further away.

Isham (2005, Sec 4.2) describes the different sources of variation between hosts which could be relevant to transmission of infection. For instance, the period from infection with HIV to diagnosis of full AIDS is known to vary with the individual's age Billard et al. (1990). Therefore, in such cases it could be important to model

the age structure population. A large number of models discussed in the literature which allow heterogeneity between the hosts are disease specific and usually focus on the complexities of a particular infection by adopting a framework similar to the SIR model but with a much larger state space.

Apart from assigning individuals particular covariates such as age, another approach to heterogeneity is to divide the population into groups where it is assumed that the individual mix homogeneously within in each group. Contacts between groups are modelled by using a *mixing matrix* whose elements r_{ij} specify the probability that an individual in group i will have a potentially infectious contact where the individuals at each different group are chosen at random. More complex structure of the mixing matrices have been considered, eg. Koopman et al. (1989). In addition, cases demography in the model such that the group sizes change over time have also been studied (see for example, Morris, 1991, 1996).

Many researchers are concerned with the epidemics in structured populations. Longini and Koopman (1982) have studied models in which individuals live in households and may be infected from an infective which either belongs to the same or to a different household and it is assumed that the disease within the household progresses independently of the dynamics of the community. There exists a comprehensive literature for models of this kind and we shall briefly mention a few key references. Addy et al. (1991) extend the work by Ball (1986) and Britton and Becker (2000) use the model Longini and Koopman (1982) to estimate the critical vaccination coverage required to prevent epidemics in a population which is partitioned into households. MCMC methods have been applied to analyse temporal and final size from households outbreaks, eg. O'Neill et al. (2000). Work on vaccination has also be done, see for example Ball and Lyne (2002) and Becker et al. (2003).

Epidemics with two levels of mixing have recently been introduced by Ball et al. (1997). Such a model assumes two different kind of contacts; a *local* and a *global*. Apart from describing the infection process of such a model, the authors also briefly

consider statistical inference for their model discuss various vaccination strategies. Classical inference using (pseudo)likelihood methods is available (Ball and Lyne, 2006). See also the recent work by Demiris and O'Neill (2005) on how to draw Bayesian inference for such type of models.

It is easy to realize that in real life application the individuals interact with a number of different environments and it is practically impossible to capture every aspect of the population structure. Lately, researchers have been concerned with modelling the population structure through a *random network* structure. Britton and O'Neill (2002) use MCMC methods to conduct Bayesian inference for a model where individuals have social contacts according to a Bernoulli random graph. There has been an intrinsic interest to extend such simple models to more complicated to social structures, and also other networks, such as internet networks including the-so-called scale free networks; see a review by Albert and Barabási (2002).

Another important assumption of the models which have been discussed so far is that they are temporally homogeneous. Nevertheless, such an assumption for endemic diseases or epidemics which last for long periods does not seem appropriate and it might be necessary to allow for time−dependent contact rates (Becker, 1989). Moreover, Hayakawa et al. (2003) extend the basic GSE in two key directions. Apart from assuming a multi-type model they also assume that the number of susceptibles is unobserved. Then, they derive statistical inference for both the infection rate and the size of the population.

Summarizing, although all the models outlined in this section play a significant role in the epidemic theory and modelling, we will not attempt to explore their properties in any more detail. This thesis will attempt to focus on drawing Bayesian inference for a general heterogeneously mixing stochastic epidemic model using MCMC methods. Such a model is described and analysed in the remaining sections of this chapter. In this section, first, we shall describe the model and its assumptions. Then, we will show how to apply modern statistical methods (MCMC) by

extending the current available methodology of the homogeneously mixing model to draw inference for the model parameters, such as the infection and the removal rate.

2.2.1 Model Construction

A natural approach is to extend to a non-homogeneously mixing by assuming that individual i makes an infectious contact with a susceptible individual j at rate β_{ij} and remains infectious for some time which is distributed according to a random variable D. If $D \sim \text{Exp}(\gamma)$, then such a model is equivalent to an extended GSE with individual-specific infection rates β_{ij}. For mathematical convenience, we will assume that

$$\beta_{ij} = \beta_0 \cdot h_{ij} \tag{2.24}$$

where h_{ij} is a deterministic function which not only can involve any individual-specific characteristics, such as age and sex, but also a measure of the distance between them. The distance is incorporated in order to allow the infection rates to decrease as this distance between the individual increases. Throughout this section, we will assume that the function h_{ij} is fully known. In the next chapter we consider the case where h_{ij} is associated with some unknown parameters for which we would interested in drawing inference for them.

2.2.2 Bayesian Inference

First we should note the difference between the HMSE and the GSE regarding the data required. The latter requires data which refer not just to the removal time (as Bailey and Thomas' setup, see Section 2.1.6.1), but also to which actual individual has been removed. Therefore, we adopt the alternative setup (see Section 2.1.6.3) where each individual is associated with their infection and removal time. In addition, for simplicity in the calculations we assume a closed population and that there is one initial infective, $a = 1$. First, we derive the likelihood which can

be broken into two independent parts: i) the infectious (L_1) and the ii) removal part (L_2).

Infectious part: Denote by S the total person-to-person infectious pressure during the course of the epidemic:

$$\begin{aligned} S &= \sum_{i=1}^{n_I}\sum_{j=1}^{N}\beta_{ij}\left((R_i \wedge I_j) - (I_i \wedge I_j)\right) \\ &= \beta_0 \sum_{i=1}^{n_I}\sum_{j=1}^{N} h_{ij}\left((R_i \wedge I_j) - (I_i \wedge I_j)\right) \\ &= \beta_0 \cdot A \end{aligned}$$

where $A = \sum_{i=1}^{n_I}\sum_{j=1}^{N} h_{ij}\left((R_i \wedge I_j) - (I_i \wedge I_j)\right)$. The infection component can be then written:

$$L_1 = \prod_{i=1, i\neq k}^{n_I}\left(\sum_{j\in\mathcal{Y}_i}\beta_{ji}\right) \times \exp\{-S\} \tag{2.25}$$

where $\mathcal{Y}_i = \{j : I_j < I_i < R_j\}$.

Removal part: The contribution to the likelihood of the infectious period of each of the infected individuals depends on its chosen distribution. Suppose, that each individual i remains infectious for some time $D_i = R_i - I_i$. Let Q be an arbitrary but specified non-negative distribution. Let $g_Q(\cdot)$ denote the probability density function of Q and let ω denote the parameters governing Q. For general infectious period Q, the contribution to the likelihood is:

$$L_2 = \prod_{i=1}^{n_R} g_Q(R_i - I_i; \omega) \tag{2.26}$$

Combining (2.25) and (2.26) we get the likelihood of the data given the model parameters:

$$f\left(\boldsymbol{I}, \boldsymbol{R} | \beta_0, \gamma\right) = L_1 \times L_2$$

$$= \beta_0^{n_I} \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in \mathcal{Y}_i} h_{ji} \right) \times \exp\{-\beta_0 A\} \times \prod_{i=1}^{n_R} g_Q(R_i - I_i; \omega)$$

In particular, we can consider $Q \sim \text{Exp}(\gamma)$, that is, $g_Q(x) = \gamma \exp\{-\gamma x\}$ and L_2 is given below:

$$L_2 = \gamma^{n_R} \exp\left\{ -\sum_{i=1}^{n_R} \gamma (R_i - I_i) \right\}.$$

Alternatively, we can consider $Q \sim \text{Gamma}(\alpha, \gamma)$, that is, $g_Q(x) = \frac{\gamma^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\gamma x\}$ and L_2 turns out to be:

$$L_2 = \gamma^{\alpha n_R} \exp\left\{ -\gamma \sum_{i=1}^{n_R} (R_i - I_i) \right\} \prod_{i=1}^{n_R} \frac{(R_i - I_i)^{\alpha-1}}{\Gamma(\alpha)}$$

For illustration purposes, we shall restrict our attention to the Gamma distribution where the shape parameter α is assumed to be known and we are focusing on drawing inference for the scale parameter, γ. Note, that for $\alpha = 1$, $Q \sim \text{Exp}(\gamma)$. We adopt a Bayesian approach and therefore we assign (independent) conjugate priors for β_0 and γ: $\pi(\beta_0) \sim Ga(\lambda_\beta, \nu_\beta)$, $\pi(\gamma) \sim Ga(\lambda_\gamma, \nu_\gamma)$. The full posterior distribution turns out to be:

$$\begin{aligned} \pi(\beta_0, \gamma, \boldsymbol{I} | \boldsymbol{R}) \quad \propto \quad & \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in \mathcal{Y}_i} h_{ji} \right) \times \beta_0^{\lambda_{\beta_0} + n_I - 1 - 1} \exp\{-\beta_0 (A + \nu_{\beta_0})\} \\ \times \quad & \gamma^{\alpha n_R + \lambda_\gamma - 1} \times \exp\left\{ -\gamma \left(\sum_{i=1}^{n_R} (R_i - I_i) + \nu_\gamma \right) \right\} \\ \times \quad & \prod_{i=1}^{n_R} (R_i - I_i)^{\alpha-1} \end{aligned} \tag{2.27}$$

The full conditional distributions the model parameters are:

$$\pi(\beta_0 | \gamma, \boldsymbol{I}, \boldsymbol{R}) \equiv Ga\left(n_I + \lambda_{\beta_0} - 1, \nu_{\beta_0} + A\right) \tag{2.28}$$

$$\pi(\gamma | \beta, \boldsymbol{I}, \boldsymbol{R}) \equiv Ga\left(\alpha n_R + \lambda_\gamma, \nu_\gamma + \sum_{i=1}^{n_R} (R_i - I_i)\right) \tag{2.29}$$

If the infection times ($\boldsymbol{I}$) are known, then inference for β_0 and γ is (again) straight-

forward without the need of MCMC since the parameters are conditionally independent. In practice, we do not observe the infection times and we need to sample from their conditional distribution as well.

2.2.3 MCMC implementation

The techniques discussed in Section 2.1.8.2 regarding the GSE model can be easily adopted. This section presents the various MCMC algorithms which have been proposed in the literature (O'Neill and Roberts, 1999, Neal and Roberts, 2005) for stochastic epidemic models and we also suggest some further modifications to this methodology.

First, we concentrate on the choice of the target distribution. A natural choice is the $\pi(\beta_0, \gamma, \boldsymbol{I}|\boldsymbol{R})$. Neal and Roberts (2005) suggested to integrate the infection rate out from the full posterior distribution $\pi(\beta_0, \gamma, \boldsymbol{I}|\boldsymbol{R})$ and construct an MCMC algorithm on the joint distribution of $(\gamma, \boldsymbol{I})$ given the observed data $\boldsymbol{R}$:

$$\pi(\gamma, \boldsymbol{I}|\boldsymbol{R}) = \int_{\beta_0} \pi(\beta_0, \gamma, \boldsymbol{I}|\boldsymbol{R}) \, \mathrm{d}\beta_0$$

$$\begin{aligned} \pi(\gamma, \boldsymbol{I}|\boldsymbol{R}) \quad \propto \quad & \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in \mathcal{Y}_i} h_{ji} \right) \times (\nu_{\beta_0} + A)^{-(\lambda_{\beta_0} + n_R - 1)} \\ \times \quad & \gamma^{n_R + \lambda_\gamma - 1} \exp\left\{ -\gamma \left(\sum_{i=1}^{n_R} (R_i - I_i) + \nu_\gamma \right) \right\} \\ \times \quad & \prod_{i=1}^{n_R} (R_i - I_i)^{\alpha - 1} \end{aligned} \tag{2.30}$$

Other available options is to either integrate γ out

$$\pi(\beta_0, \boldsymbol{I}|\boldsymbol{R}) = \int_{\gamma} \pi(\beta_0, \gamma, \boldsymbol{I}|\boldsymbol{R}) \, \mathrm{d}\gamma$$

$$\pi(\beta_0, \boldsymbol{I}|\boldsymbol{R}) \propto \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in \mathcal{Y}_i} d_{ji} \right) \times \beta_0^{\lambda_{\beta_0}+n_I-1-1} \exp\left\{-\beta_0(A+\nu_{\beta_0})\right\} \times \left(\nu_\gamma + \sum_{i=1}^{n_R} (R_i - I_i) \right)^{-(\lambda_\gamma + n_R)} \times \prod_{i=1}^{n_R} (R_i - I_i)^{\alpha-1} \tag{2.31}$$

or even both β_0 and γ and then obtain the marginal distribution of the missing data ($\boldsymbol{I}$) given the observed ($\boldsymbol{R}$):

$$\pi(\boldsymbol{I}|\boldsymbol{R}) \propto \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in \mathcal{Y}_i} d_{ji} \right) \times (\nu_{\beta_0} + A)^{-(\lambda_{\beta_0}+n_I-1)} \times \left(\nu_\gamma + \sum_{i=1}^{n_R} (R_i - I_i) \right)^{-(\lambda_\gamma + n_R)} \times \prod_{i=1}^{n_R} (R_i - I_i)^{\alpha-1} \tag{2.32}$$

$$\pi(\boldsymbol{I}|\boldsymbol{R}) = \int_{\beta_0} \int_{\gamma} \pi(\beta_0, \gamma, \boldsymbol{I}|\boldsymbol{R}) \, \mathrm{d}\beta_0 \, \mathrm{d}\gamma$$

If we construct an MCMC algorithm based on one of these target distributions, then samples from the posterior distribution of the parameters which were integrated out can be drawn from the resultant samples. For instance if we integrate β_0 out, then β_0 values can be generated from the sample of $(\boldsymbol{R}, \boldsymbol{I}, \gamma)$ since $\pi(\beta_0|\boldsymbol{I}, \gamma, \boldsymbol{R}) \sim Ga(n_I + \lambda_{\beta_0} - 1, \nu_{\beta_0} + A)$. Having chosen the target distribution, a variety of algorithms can be implemented.

We first consider the standard MCMC algorithm for stochastic epidemic models. This similar to the algorithm which was used to draw samples from the parameters of the GSE model.

Centered algorithm I: $[C]$

(Repeat the following steps)

1. Start the chain with initial values: $\beta_0^0, \gamma^0, \boldsymbol{I}^0$;
2. Update β_0 by using Gibbs Sampler and drawing from $\pi(\beta_0|\gamma, \boldsymbol{I}, \boldsymbol{R})$;
3. Update γ by using Gibbs Sampler and drawing from $\pi(\gamma|\beta_0, \boldsymbol{I}, \boldsymbol{R})$;
4. Choose uniformly one (or more) infection times $I_j, j = 1, \ldots, n_I$ and update it (them) using Metropolis Hastings algorithm;

Sampling directly the infection times is problematical and therefore a Metropolis step is used instead. Step 4 refers to the available options of updating the infection times. We can either choose at random to update one infection time having the other fixed (*random scan*), or a random subset of the infection times (eg. say 10%), or update each of the infection times individually (*deterministic scan*). Moreover, we might also want to perform a block update. Our choice needs to be made on the basis of computational time and the mixing properties of the Markov chain. We discuss the properties in Section 2.6 by performing a simulation study.

Once the choice of the number of infection times to be updated is made, then we need to decide how to implement the Metropolis step which consists of the proposing a new infection time I_j' from a distribution, say $q(\cdot)$. Apart from a standard random walk Metropolis algorithm (Neal and Roberts, 2004):

$$q(I_j, I_j') \equiv N(I_j, \sigma^2),$$

an independence sampler which makes use of the likelihood equations can also be

applied:

$$q(R_j - I_j, R_j - I'_j) \equiv \text{Gamma}(\alpha, \gamma).$$

We focus now, on the different target distributions which we are interested in drawing samples from. Due to the conditional independence of the model parameters the algorithm used for the target distribution having β_0 integrated out ($[C - \beta_0]$) is very similar to the $[C]$ algorithm.

Centered algorithm II: $[C - \beta_0]$

(Integrate β_0 out)

(Repeat the following steps)

1. `Start the chain with initial values:`$\gamma^0, \boldsymbol{I}^0$;
2. `Update` γ `by using Gibbs Sampler and drawing from` $\pi(\gamma|\boldsymbol{I}, \boldsymbol{R})$;
3. `Choose uniformly one (or more) infection times` I_j, $j = 1, \ldots, n_I$ `and update it (them) using Metropolis Hastings algorithm;`

[Get β_0 values from the resultant sample of $(\boldsymbol{I}, \boldsymbol{R})$]

On the other hand, if we integrate γ out then the independence sampler presented proposed in $[C]$ and $[C - \beta_0]$ algorithms, i.e $q(R_j - I_j, R_j - I'_j) \equiv \text{Gamma}(\alpha, \gamma)$, becomes inappropriate. This is due to the fact that γ does not exist in the parameter space any longer and hence cannot be used within the proposal of the infection times.

In order to take advantage of a proposal which makes use of the likelihood equations

we could propose an infection time from a similar distribution,

$$R_i - I_i' \sim \text{Exp}\left(\gamma^f\right)$$

where γ^f can be a fixed value obtained from a pilot study. Alternatively, γ^f can be the MLE of γ given the current value of the infection times, $I_i^c, i = 1, \ldots, n_I$ at each MCMC step:

$$\gamma^f = \frac{\alpha n_R}{\sum_{i=1}^{n_R}(R_i - I_i^c)}$$

Note that unlike the former, the latter is not an independence sampler because the proposed value I_i' depends on the current value I_i^c, indirectly, through γ^f. Note, that extra care is required while evaluating the ratio of the proposal densities $q(\cdot)$ in order to accept or reject I_i'.

Centered algorithm III: $[C - \gamma]$

(Integrate γ out)

(Repeat the following steps)

1. `Start the chain with initial values:` $\beta^0, \boldsymbol{I}^0$`;`
2. `Update` β_0 `by using Gibbs Sampler and drawing from` $\pi(\beta_0|\boldsymbol{I}, \boldsymbol{R})$`;`
3. `Choose uniformly one (or more) infection times` I_j`,` $j = 1, \ldots, n_I$ `and update it (them) using Metropolis Hastings algorithm;`

[Get γ values from the resultant sample of $(\boldsymbol{I}, \boldsymbol{R})$]

If integrate both the model parameters out, we update the infection times in a similar way to the $[C - \gamma]$ algorithm since the proposal in the $[C]$ and $[C - \beta_0]$ is also inappropriate for this algorithm.

Centered algorithm IV: $[C - \beta_0 - \gamma]$

(Integrate β_0 and γ out)

1. `Start the chain with initial values:` $\boldsymbol{I}^0$;

2. `Choose uniformly one (or more) infection times` I_j,

   $j = 1, \ldots, n_I$

   `and update it (them) using Metropolis Hastings algorithm;`

[Get β_0 and γ values from the resultant sample of $(\boldsymbol{I}, \boldsymbol{R})$]

We can summarize the available centered algorithms for the spatial stochastic epidemic model as shown in Table 2.1.

Table 2.1: Nomenclature for the centered MCMC algorithms

Algorithm	**Nomenclature**
Centered	$[C]$
Centered with β_0 integrated out	$[C - \beta_0]$
Centered with γ integrated out (γ^f is assigned a fixed value)	$[C - \gamma]$
Centered with β_0 and γ integrated out (γ^f is assigned a fixed value)	$[C - \beta_0 - \gamma]$
Centered with γ integrated out (γ^f is assigned its MLE)	$[C_2 - \gamma]$
Centered with β_0 and γ integrated out (γ^f is assigned its MLE)	$[C_2 - \beta_0 - \gamma]$

2.3 On Centered Reparameterisations

2.3.1 Motivation

It has been already stated that the estimation of R_0 does not rely on observing the infection times (see Section 2.1.7.1). Taking into account that in practice infection times are not observed and that missing data can cause problems of deriving efficient MCMC algorithms (see Section 1.9.2), this section considers a centered reparameterisation for stochastic epidemic models which makes use of ψ. We bring to attention that we have a change in variables from $(\beta_0, \gamma, \boldsymbol{I}, \boldsymbol{R})$ to $(\psi, \gamma, \boldsymbol{I}, \boldsymbol{R})$ rather than reconstructing the model having model parameters ψ and γ.

The main difference between this and the standard parameterisation, is that the two model parameters are not conditionally independent on the missing data. The Jacobian is equal to $1/\gamma$ and the posterior distribution of the parameters (by change of variable theorem) becomes:

$$\begin{aligned}\pi(\psi|\gamma, \boldsymbol{I}, \boldsymbol{R}) \propto & \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j\in Y_i} d_{ji}\right) \times \psi^{n_I+\lambda_{\beta_0}-1-1} \exp\left\{-\psi\cdot\gamma(A+\nu_{\beta_0})\right\} \\ \times & \ \gamma^{n_I+\alpha n_R+\lambda_{\beta_0}+\lambda_\gamma-1-1} \exp\left\{-\gamma\left(\sum_{i=1}^{n_R}(R_i-I_i)+\nu_\gamma\right)\right\} \\ \times & \prod_{i=1}^{n_I}\left(\frac{1}{\Gamma(\alpha)}(R_i-I_i)^{\alpha-1}\right) \end{aligned} \tag{2.33}$$

The full conditional distributions of $\psi|\gamma, \boldsymbol{I}, \boldsymbol{R}$ and $\gamma|\psi, \boldsymbol{I}, \boldsymbol{R}$ still remain of a standard form:

$$\pi(\psi|\gamma, \boldsymbol{I}) \equiv Ga\left(n_I + \lambda_{\beta_0} - 1, \gamma(A+\nu_{\beta_0}) + \nu_{\beta_0}\right) \tag{2.34}$$

$$\pi(\gamma|\psi, \boldsymbol{I}) \equiv Ga\left(\alpha n_R + n_I + \lambda_{\beta_0} + \lambda_\gamma - 1, \nu_\gamma + \sum_{i=1}^{n_R}(R_i - I_i) + \psi(A + \nu_{\beta_0})\right) \tag{2.35}$$

Unlike ψ and γ, samples from the conditional distribution of the infection times cannot be drawn such easily and therefore a Metropolis step should be applied in a pretty much similar way like the other centered algorithms by proposing from the Exponential distribution $R_i - I_i' \sim \text{Ga}(\alpha, \gamma)$. We then can implement the $[C_\psi]$ algorithm.

Centered Reparameterisation I: $[C_\psi]$

(Repeat the following steps)

1. `Start the chain with initial values:` $\psi^0, \gamma^0, \boldsymbol{I}^0$`;`
2. `Update` ψ `by using Gibbs Sampler and drawing from` $\pi(\psi|\gamma, \boldsymbol{I}, \boldsymbol{R})$`;`
3. `Update` γ `by using Gibbs Sampler and drawing from` $\pi(\gamma|\psi, \boldsymbol{I}, \boldsymbol{R}$`;`
4. `Choose uniformly one (or more) infection times` I_j`,`$j = 1, \ldots, n_I$ `and update it (them) using Metropolis Hastings algorithm;`

2.3.2 Integrate ψ out

Because of the standard form of $\pi(\psi|\gamma, \boldsymbol{I}, \boldsymbol{R})$ we can integrate ψ out and obtain the marginal distribution of γ and $\boldsymbol{I}$ given the observed removal times $\boldsymbol{R}$. Unlike the $[C - \beta_0]$ parameterisation, γ is not Gamma distributed any more. Therefore, apart from using Metropolis to update the infection times, another Metropolis step

is needed to draw samples from the full conditional posterior distribution of the γ. A natural approach is to use a (multiplicative) random walk since it has always to be strictly positive.

$$\begin{aligned}\pi(\gamma|\boldsymbol{I},\boldsymbol{R}) &\propto \prod_{i=1,i\neq k}^{n_I}\left(\sum_{j\in Y_i} d_{ji}\right)\times\left(\gamma\cdot(\nu_{\beta_0}+A)\right)^{-(n_I+\lambda_{\beta_0}-1)}\\ &\times \gamma^{n_I+\alpha n_R+\lambda_{\beta_0}+\lambda_\gamma-1-1}\exp\left\{-\gamma\left(\sum_{i=1}^{n_R}(R_i-I_i)+\nu_\gamma\right)\right\}\\ &\times \prod_{i=1}^{n_I}\left(\frac{1}{\Gamma(\alpha)}(R_i-I_i)^{\alpha-1}\right)\end{aligned} \tag{2.36}$$

Nevertheless, there is also a choice of various independence samplers. First, we could make use of conditional distribution of $\gamma|r_0,\boldsymbol{I},\boldsymbol{R}$ and propose γ' from:

$$q(\gamma,\gamma')\equiv Ga\left(n_I+\alpha n_R+\lambda_{r_0}+\lambda_\gamma-1,\sum_{i=1}^{n_R}(R_i-I_i)+\nu_\gamma+r_0\left(A+\nu_{r_0}\right)\right) \tag{2.37}$$

Obviously, this proposal cannot be used explicitly because it involves ψ which does not exist in the parameter space since it has been integrated out. However a similar proposal could be used by assigning ψ in (2.37) a fixed value which may be obtained via pilot studies.

In addition, we could also propose γ from:

$$q(\gamma,\gamma')\equiv f(\cdot).$$

One way to make this proposal efficient is to choose $f(\cdot)$ to be an approximation to the posterior distribution of $\pi(\gamma|\boldsymbol{I},\boldsymbol{R})$. This can be done by performing a pilot study first, i.e. run the $[C]$ algorithm for an adequate number of iterations and approximate the obtained distribution via the method of moments (for instance by matching the mean and the variance) with a distribution of a standard form.

It is preferable to choose $f(\cdot)$ having heavy tails (for example, a t distribution) to avoid regular problems of an independent sampler not visiting the tails of the target distribution often.

Centered Reparameterisation II: $[C_\psi - \psi]$

(Repeat the following steps)

1. Start the chain with initial values:$\gamma^0, \boldsymbol{I}^0$;
2. Update γ by using Metropolis Hastings algorithm;
3. Choose uniformly one (or more) infection times $I_j, j = 1, \ldots, n_I$ and update it (them) using Metropolis Hastings algorithm;

2.3.3 Integrate γ out

Instead of ψ, we could integrate γ out. We then get the following conditional posterior density:

$$\begin{aligned}\pi(\psi|\boldsymbol{I},\boldsymbol{R}) \;\propto\; & \prod_{i=1,i\neq k}^{n_I}\left(\sum_{j\in Y_i} d_{ji}\right)\times\psi^{n_I+\lambda_{\beta_0}-1-1} \\ & \times\;\left(\nu_\gamma+\sum_{i=1}^{n_R}(R_i-I_i)+\psi(A+\nu_{\beta_0})\right)^{-(\alpha n_R+n_I+\lambda_{\beta_0}+\lambda_\gamma-1)} \\ & \times\;\prod_{i=1}^{n_I}\left(\frac{1}{\Gamma(\alpha)}(R_i-I_i)^{\alpha-1}\right) \qquad (2.38)\end{aligned}$$

Once γ is integrated out, the independence sampler used in $[C_\psi]$ must be slightly modified for the same reasons as in the $[C-\gamma]$ and $[C-\beta_0-\gamma]$ (see Section 2.2.3 for details). Moreover, $\psi|\boldsymbol{I},\boldsymbol{R}$ is not Gamma distributed any longer and therefore a Metropolis step has to be applied. Apart from a (multiplicative) random walk,

we could also use an independence sampler as we did to update γ in $[C_\psi - \psi]$. We could propose ψ' from:

$$q(\psi, \psi') \equiv Ga(n_I + \lambda_\psi - 1, \gamma \cdot (A + \nu_\psi))$$

where γ is substituted by a fixed value. We might also propose

$$q(\psi, \psi') \equiv h(\cdot)$$

where $h(\cdot)$ approximates $\pi(\psi|\boldsymbol{R})$ via the method of moments.

Centered Reparameterisation III: $[C_\psi - \gamma]$

(Repeat the following steps)

1. Start the chain with initial values: $\psi^0, \boldsymbol{I}^0$;
2. Update ψ by using Metropolis Hastings algorithm;
3. Choose uniformly one (or more) infection times $I_j, j = 1, \ldots, n_I$ and update it (them) using Metropolis Hastings algorithm;

Concluding, Table 2.2 represents the nomenclature for the algorithms which have been considered in that section referring to some centered reparameterisations. Note that unlike $[C - \beta_0 - \gamma]$, we are not able to integrate ψ and γ both at the same time since they are not conditionally independent on the infection time. Therefore an algorithm $[C_\psi - \psi - \gamma]$ is very difficult (if possible) to be implemented.

Table 2.2: Nomenclature for the centered reparameterized MCMC algorithms

Algorithm	**Nomenclature**
Centered $(\psi, \gamma, \boldsymbol{I})$	$[C_\psi]$
Centered with ψ integrated out	$[C_\psi - \psi]$
Centered with γ integrated out	$[C_\psi - \gamma]$

2.4 On Non−Centered Parameterisations

2.4.1 Introduction

In this section we first review the existing non−centered (NC) and partially non−centered algorithms for stochastic epidemic models as introduced in Neal and Roberts (2005). Such reparameterisations will allow us to break the strong correlation between γ and $\boldsymbol{I}$. Then then we show how we can apply the same algorithms for the HMSE model (see Section 2.2).

2.4.2 Non−Centered Parameterisations

Neal and Roberts (2005) were the first to introduce a γ−non−centered parameterisation (γNCP) for stochastic epidemic models. Apart from introducing such reparameterisations for the GSE, the authors also applied NCP for models which extend the GSE, such as by assuming a Gamma or Weibull infectious period instead of the Exponential.

The centered parameterisations described in Section 2.2.3 alternate between updating the model parameters (β, γ) and the missing data $(\boldsymbol{I})$. On the other hand, the NC parameterisation update the model parameters and the missing data together. Let

$$U_i = \gamma \cdot (R_i - I_i), \text{ for } i = 1, \ldots, n_I.$$

It is trivial to show that *a priori* $U_i \sim Gamma(\alpha, 1)$ and given (U_i, R_i, γ) the infection times can be easily derived because $I_i = R_i - \frac{1}{\gamma}U_i$.

Papaspiliopoulos (2003) implemented NC algorithms for a variety of different models that he looked by reconstructing the likelihood equations with respect to the new parameters. This is not an easy task to do within the context of stochastic epidemic models. On the other hand, another way to view the NCP is to see it as a change of variables from $(\boldsymbol{I}, \beta_0, \gamma, \boldsymbol{R})$ to $(\boldsymbol{U}, \beta_0, \gamma, \boldsymbol{R})$ where we need to take into account the Jacobian which is equal to $1/\gamma^{n_R}$. The posterior distribution $\pi(\beta_0, \gamma, \boldsymbol{U}|\boldsymbol{R})$ for the spatial stochastic model can be derived as follows:

$$\begin{aligned}\pi(\beta_0, \gamma, \boldsymbol{U}|\boldsymbol{R}) \;\propto\; & \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in Y_i^{\boldsymbol{U}}} d_{ji} \right) \times \beta_0^{\lambda_{\beta_0} + n_I - 1 - 1} \exp\left\{-\beta_0 (A_{\boldsymbol{U}} + \nu_{\beta_0})\right\} \\ \times \;& \gamma^{\lambda_\gamma - 1} \exp\left\{-\gamma \nu_\gamma\right\} \prod_{i=1}^{n_R} \left\{ \frac{1}{\Gamma(\alpha)} U_i^{\alpha - 1} \exp\left\{-U_i\right\} \right\} \end{aligned} \tag{2.39}$$

where $A_{\boldsymbol{U}}$ and $Y_i^{\boldsymbol{U}}$ are functions of $\boldsymbol{U} = (U_1, \ldots, U_{n_I})^T$. Computationally, they can be easily calculated as follows:

$$A_{\boldsymbol{U}} = \sum_{i=1}^{n_I} \sum_{j=1}^{N} h_{ij} \left(\left(R_i \wedge I_j^U\right) - \left(I_i^U \wedge I_j^U\right) \right)$$

and

$$Y_i^U := \left\{ j : I_j^U < I_i^U < R_j \right\}$$

where $I_i^U = R_i - \frac{1}{\gamma}U_i$ for any $i, j = 1, \ldots, n_I$.

The form of the posterior distribution as shown in (2.39) indicates that β_0 and γ are conditionally independent and therefore the infection parameter can be easily integrated out to obtain the marginal distribution of γ and $\boldsymbol{U}$:

$$\begin{aligned}\pi(\gamma, \boldsymbol{U}|\boldsymbol{R}) \;\propto\; & \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in Y_i^{\boldsymbol{U}}} d_{ji} \right) \times \left(A_{\boldsymbol{U}} + \nu_{\beta_0}\right)^{-(\lambda_{\beta_0} + n_I - 1)} \\ \times \;& \gamma^{\lambda_\gamma - 1} \exp\left\{-\gamma \nu_\gamma\right\} \prod_{i=1}^{n_R} \left\{ \frac{1}{\Gamma(\alpha)} U_i^{\alpha - 1} \exp\left\{-U_i\right\} \right\} \end{aligned} \tag{2.40}$$

Unlike, the centered algorithm $[C - \beta_0 - \gamma]$, within the NC reparameterisation is not possible to integrate γ out since it appears in the functions $A_{\boldsymbol{U}}$ and $Y_i^{\boldsymbol{U}}$. Adopting the alternative setup as described in Section 2.1.6.3 we associate each individual with their infection (I_i) and removal (R_i) time. This allow us to very naturally introduce the NCP by the transformation $U_i = \gamma(R_i - I_i)$, $i = 1, \ldots, n_I$. If the original setup by Bailey and Thomas (1971) had been adopted it, the implementation of the NCP would have been harder.

In principle, the following MCMC algorithm could have been implemented:

Non−Centered MCMC Algorithm

(Repeat the following steps)

1. `Start the chain with initial values:` β_0^0, $\gamma^0, \boldsymbol{U}^0$;
2. `Update` β_0 `using Gibbs Sampler;`
3. `Update` γ `using a Metropolis step by proposing` $\gamma' \sim q_1(\cdot)$;
4. `Choose uniformly one (or more) of the` U_j`'s` $, j = 2, \ldots, n_I$ `and update it (them) by proposing` $U_j' \sim \texttt{Ga}(\alpha, 1)$ $(1 \leq j \leq n_I)$.

However, such an algorithm is not very easily implemented. This is due to the difficulty of not being able to easily derive the likelihood with respect to parameters γ, β_0 and $\boldsymbol{U}$. Therefore we propose an alternative way of implementation which makes use of the existing computer code used for the centered algorithms (see also, p. 100 Papaspiliopoulos, 2003).

Non−Centered MCMC Algorithm (*Alternative Format*): [NC]

(Repeat the following steps)

1. Start the chain with initial values: β_0^0, $\gamma^0, \boldsymbol{I}^0$;
2. Update β_0 using Gibbs Sampler;
3. Choose uniformly one (or more) of the I_j's, $i = 1, \ldots, n_I$ and update it (them) by proposing I_j^1 from $R_j - I_j' \sim \text{Ga}(\alpha, \gamma)$;
4. Set $U_i^c = \gamma^0(R_i - I_i^1)$;
5. Propose $\gamma' \sim h(\cdot, \cdot)$;
6. Set $I_i^{'} = R_j - \frac{1}{\gamma'} U_j^c$, for $1 \leq i \leq n_I$.
7. Accept γ' with probability

$$1 \wedge \frac{\pi(\gamma' | \boldsymbol{I}^{'}, \boldsymbol{R}, \beta_0)}{\pi(\gamma | \boldsymbol{I}^1, \boldsymbol{R}, \beta_0)} \cdot \frac{h(\gamma', \gamma)}{h(\gamma, \gamma')};$$

The key difference between a centered and a non−centered approach is shown clearly in Step 5.3 where by updating γ we update jointly (as a block) the missing data at the same time. The same algorithm can be applied if we wish to integrate β_0 out and have as target distribution, the one obtained in equation (2.40).

Step 5.1 states the Metropolis Hastings algorithm for the update of γ. Neal and Roberts (2005) suggest a random Walk Metropolis but we also consider and propose other proposals in Section 2.5.

2.4.3 Partially Non−Centered Parameterisations

Following the approach in Papaspiliopoulos et al. (2003), Neal and Roberts (2005) introduced partially non−centered parameterisations (PNCP) for stochastic epi-

demic models. We adopt this methodology for the HMSE.

The set of the infected individuals in the epidemic, is partitioned into two groups, $\mathcal{C}$ and $\mathcal{U}$. Let $\boldsymbol{I}^{\mathcal{C}}$ and $\boldsymbol{I}^{\mathcal{U}}$ denote the infection times of the individuals in groups $\mathcal{C}$ and $\mathcal{U}$ respectively. For those individuals in the $\mathcal{U}$, let

$$U_i = \gamma(R_i - I_i) \;\; (i \in \mathcal{U}),$$

i.e. we propose a change in variable from $(\boldsymbol{I}^{\mathcal{C}}, \boldsymbol{I}^{\mathcal{U}}, \beta_0, \gamma, \boldsymbol{R})$ to $(\boldsymbol{I}^{\mathcal{C}}, \boldsymbol{U}^{\mathcal{U}}, \beta_0, \gamma, \boldsymbol{R})$. If $\mathcal{U} = \emptyset$, then we get the centered parameterisation. The Jacobian for the transformation is $\gamma^{-\omega}$, where ω is the number of individuals in the set $\mathcal{U}$. The posterior distribution then becomes:

$$\begin{aligned}
\pi(\beta_0, \gamma, \boldsymbol{I}^{\mathcal{C}}, \boldsymbol{U}^{\mathcal{U}} | \boldsymbol{R}) \;\; \propto \;\; & \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in Y_i^{\boldsymbol{U}}} d_{ji} \right) \times \beta_0^{\lambda_{\beta_0} + n_I - 1 - 1} \exp\left\{ -\beta_0 (A_{\boldsymbol{U}} + \nu_{\beta_0}) \right\} \\
\times \;\; & \gamma^{\lambda_\gamma - 1} \exp\left\{ -\gamma \nu_\gamma \right\} \\
\times \;\; & \gamma^{\alpha\omega} \exp\left\{ -\gamma \sum_{i \in \mathcal{C}} (R_i - I_i) \right\} \prod_{i \in \mathcal{C}} \left\{ \frac{1}{\Gamma(\alpha)} (R_i - I_i)^{\alpha} \right\} \\
\times \;\; & \prod_{i \in \mathcal{U}} \left\{ \frac{1}{\Gamma(\alpha)} U_i^{\alpha - 1} \exp\left\{ -U_i \right\} \right\}
\end{aligned} \tag{2.41}$$

The infection rate β_0 can be integrated out from (2.41) and get then the marginal distribution of γ and the infection times (centered and non−centered, $\boldsymbol{I}^{\mathcal{C}}, \boldsymbol{I}^{\mathcal{U}}$ respectively).

$$\begin{aligned}
\pi(\gamma, \boldsymbol{I}^{\mathcal{C}}, \boldsymbol{U}^{\mathcal{U}} | \boldsymbol{R}) \;\; \propto \;\; & \prod_{i=1, i\neq k}^{n_I} \left(\sum_{j \in Y_i^{\boldsymbol{U}}} d_{ji} \right) \times (A_{\boldsymbol{U}} + \nu_{\beta_0})^{-(\lambda_{\beta_0} + n_I - 1)} \\
\times \;\; & \gamma^{\lambda_\gamma - 1} \exp\left\{ -\gamma \nu_\gamma \right\} \\
\times \;\; & \gamma^{\alpha\omega} \exp\left\{ -\gamma \sum_{i \in \mathcal{C}} (R_i - I_i) \right\} \prod_{i \in \mathcal{C}} \left\{ \frac{1}{\Gamma(\alpha)} (R_i - I_i)^{\alpha} \right\} \\
\times \;\; & \prod_{i \in \mathcal{U}} \left\{ \frac{1}{\Gamma(\alpha)} U_i^{\alpha - 1} \exp\left\{ -U_i \right\} \right\}
\end{aligned} \tag{2.42}$$

If for $1 \leq i \leq n_I$ we let:

$$Z_i = \begin{cases} 1 & \text{with probability } \mu_i \\ 0 & \text{with probability } 1 - \mu_i \end{cases} \tag{2.43}$$

Then set $\mathcal{C} = \{i : Z_1 = 1\}$ and $\mathcal{U} = \{i : Z_i = 0\}$. Then, the following γPNC algorithm could be adopted:

Partially Non−Centered MCMC Algorithm

(Repeat the following steps)

1. Start the chain with initial values: β_0^0, $\gamma^0, \boldsymbol{I}^{\mathcal{C}^0}, \boldsymbol{U}^{\mathcal{U}^0}$;
2. Update β_0 using Gibbs Sampler;
3. Update $\boldsymbol{Z}$ and hence $\mathcal{C}$ and $\mathcal{U}$;
4. Update γ by proposing $\gamma^{'} \sim h(\cdot)$;
5. Draw j uniformly at random from $1, 2, \ldots, n_I$.
   If $j \in \mathcal{C}$ $(j \in \mathcal{U})$ then update I_j (U_j) using a Metropolis-Hastings step by proposing
   $R_j - I_j^{'} \sim \texttt{Ga}(\alpha, \gamma)$ $(U_j \sim \texttt{Ga}(\alpha, 1))$.

This algorithm cannot be very easily implemented for the similar reasons explained in Section 2.4 and therefore we present an alternative and easier way of implementing the PNC reparameterisation.

Partially Non−Centered MCMC Algorithm (*Altern. Format*) [PNC]

(Repeat the following steps)

1. Start the chain with initial values: β_0^0, $\gamma^0, \boldsymbol{U}^0$;
2. Update β_0 using Gibbs Sampler;
3. Choose uniformly one (or more) of the I_j's, $i = 1, \ldots, n_I$ and update it (them) by proposing I_j^1 from $R_j - I_j^1 \sim \text{Ga}(\alpha, \gamma)$;
4. Update Z and hence $\mathcal{C}$ and $\mathcal{U}$;
5. Set $U_i = \gamma(R_i - I_i^1)$ for $i \in \mathcal{U}$;
6. Propose γ', say $\gamma' \sim h(\cdot, \cdot)$;
7. Set $I_i' = R_i - \frac{1}{\gamma'} U_i^c$ for $i \in \mathcal{U}$ and $I_i' = I_i^1$ for $i \in \mathcal{C}$;
8. Accept γ' with probability

$$1 \wedge \frac{\pi(\gamma' | \boldsymbol{I}', \boldsymbol{R}, \beta_0)}{\pi(\gamma | \boldsymbol{I}^1, \boldsymbol{R}, \beta_0)} \cdot \frac{h(\gamma', \gamma)}{h(\gamma, \gamma')};$$

Following Papaspiliopoulos et al. (2003), we could apply an alternative form of a γPNCP, by partially non−centering each infectious period. Neal and Roberts (2005) argued that such an algorithm did not perform as well as the non−centered algorithm outlined in Section 2.4.3 and therefore we decide not implement such a PNCP.

The NC and PNC algorithms can be slightly modified to be appropriated when the infection rate is integrated out and get $[NC-\beta_0]$ and $[PNC-\beta_0]$ respectively. The alternative implementations of the NC and PNC algorithms can be relatively easy applied by making use of the existing computer codes for the centered algorithms.

In general the approach we used to construct a PNC algorithm can be summarized as follows:

1. `Get a sample from` $\pi(\boldsymbol{I}|\boldsymbol{R})$ `and` $\pi(\gamma|\boldsymbol{R})$ `via a centered algorithm;`
2. `Transform the` $\boldsymbol{I}$`'s to` $\boldsymbol{U}$`'s and update` γ `using Metropolis Hastings algorithm;`

The nomenclature for the PNC algorithms described in this section are shown in Table 2.3.

Table 2.3: Nomenclature for the PNC algorithms

Algorithm	**Nomenclature**
$\delta\%$ PNC $(\beta_0, \gamma, \boldsymbol{I}^{\mathcal{C}}, \boldsymbol{I}^{\mathcal{U}})$	$[\delta\%PNC]$
$\delta\%$ PNC with β_0 integrated out	$[\delta\%PNC - \beta_0]]$

2.5 On Efficient Partially Non−Centered Parameterisations

In this section we focus on deriving more efficient ways of implementing a non−centered and a partially non−centered parameterisation as described in Section 2.4. Without loss of generality, we concentrate only on the partially non−centered algorithms, since a 100% PNC is equivalent to a fully NC.

2.5.1 Draw samples of γ and $\boldsymbol{I}$

As stated in Section 2.4, in general, the PNC algorithms can be summarized in two steps. First, we get a sample from of $(\gamma, \boldsymbol{I})$ via a centered parameterisation having

as a target distribution $\pi(\gamma, \boldsymbol{I}|\boldsymbol{R})$ and then update γ using the non-centered reparameterisation: Neal and Roberts (2005) have proposed to get posterior samples from $(\gamma, \boldsymbol{I})$ via the following approach:

Neal and Roberts Approach [NR]

1. Choose one (or more) infection times I_j, $j = 1, \ldots, n_I$ and update it (them) using Metropolis Hastings by proposing from $R_j - I'_j \sim \text{Exp}(\gamma)$ and accept it with probability:

$$1 \wedge \frac{\pi(\boldsymbol{I}'|\gamma, \boldsymbol{R})}{\pi(\boldsymbol{I}|\gamma, \boldsymbol{R})} \frac{q(\boldsymbol{I}', \boldsymbol{I})}{q(\boldsymbol{I}, \boldsymbol{I}')}$$

2. Set $U_i = \gamma(R_i - I'_i)$ for $i = 1, \ldots, n_I$;

3. Update γ using Random Walk Metropolis by proposing $\gamma' \sim N(\gamma, \sigma_\gamma)$ and accept it with probability:

$$1 \wedge \frac{\pi(\gamma'|\boldsymbol{U})}{\pi(\gamma|\boldsymbol{U})};$$

We propose another way of implementing the PNC algorithm:

Our Approach

1. Obtain a sample of $(\boldsymbol{I}', \gamma')$ via the centered algorithm which has the best performance;
2. Set $U_i = \gamma'(R_i - I'_i)$ for $i = 1, \ldots, n_I$;
3. Update γ' using Metropolis algorithm by proposing $\gamma'' \sim h(\gamma, \gamma')$ and accept it with probability:

$$1 \wedge \frac{\pi(\gamma''|\boldsymbol{U})}{\pi(\gamma'|\boldsymbol{U})} \frac{h(\gamma'', \gamma\prime)}{h(\gamma', \gamma'')};$$

It easy to see that the main difference between our and Neal and Roberts (2005) approach is the way that samples of $(\boldsymbol{I}, \gamma)$ are drawn. They draw them using the conditional distribution of the infection times given the removal rates, $\pi(\boldsymbol{I}|\gamma, \boldsymbol{R})$ (in a centered framework), which actually is equivalent to the $[C - \beta_0]$ algorithm. However, the simulation study in Section 2.6 indicates the $[C-\beta_0]$ does not always perform better than the other variations of the centered algorithm. Therefore, the 1st Step represents the algorithm via which we can draw the most efficiently samples of $(\boldsymbol{I}, \gamma)$ among the centered algorithms.

As it has been stated already, overall, the relative performance of the centered algorithms shown in Table 2.1 mainly depends on how informative about the parameter γ, the infection times are. We should note that if we choose to obtain a sample of $(\boldsymbol{I}, \gamma)$ via a centered algorithm where γ is integrated out, such as the algorithms $[C_1 - \gamma]$, $[C_1 - \beta_0 - \gamma]$, $[C_2 - \gamma]$, $[C_2 - \beta_0 - \gamma]$ then in order to preserve the invariance of the Markov chain we need a further modification, i.e. update γ within the centered framework before applying the NCP. Suppose that the $[C_1 - \beta_0 - \gamma]$ algorithm is chosen for the Step 1:

1i. `Start with initial values` $\boldsymbol{I}^0$;

1ii. `Choose one (or more) infection times` I_j, $j = 1, \ldots, n_I$ `and update it (them) using Metropolis Hastings by proposing from an appropriate distribution and accept the new set of infection times` $\boldsymbol{I}'$ `with probability`:

$$1 \wedge \frac{\pi(\boldsymbol{I}'|\boldsymbol{R})}{\pi(\boldsymbol{I}|\boldsymbol{R})} \frac{q(\boldsymbol{I}', \boldsymbol{I})}{q(\boldsymbol{I}, \boldsymbol{I}')}$$

1iii. `Update` γ `within the centered framework by drawing from its conditional distribution:`

$$\pi(\gamma'|\boldsymbol{I}', \boldsymbol{R}) \equiv Ga\left(\alpha n_R + \lambda_\gamma, \nu_\gamma + \sum_{i=1}^{n_R}(R_i - I_i')\right)$$

Similar steps should be followed if a different target distribution is chosen which does not involve γ in its parameter space, eg. the algorithm $[C_1 - \gamma]$.

2.5.2 Update the removal rate γ

An important part of the a NC implementation is how to update the removal rate γ. Neal and Roberts (2005) proposed a random walk Metropolis to update γ. We also propose to choose an independence sampler:

- **Random Walk Metropolis**:

$$h(\gamma, \gamma') \equiv N(\gamma, \sigma^2)$$

Obviously, since $\gamma > 0$ we could also use a multiplicative random walk Metropolis.

- **"Pseudo-Independence" Sampler I** - *Pseudo*−Gibbs: We choose to update γ as if it was centered and propose from its full conditional distribution:

$$h(\gamma, \gamma') \equiv \pi(\gamma | \boldsymbol{I}^{\text{cur}}, \boldsymbol{R})$$

  where $\boldsymbol{I}^{\text{cur}}$ denotes the current infection times obtained via the centered algorithm (Step 1).

- **"Pseudo-Independece" Sampler II** - *Normal Approximation*: Having samples from the infection times at each iteration ($\boldsymbol{I}^{\text{cur}}$), we can make use of the maximum likelihood estimate and the corresponding standard error (Equations 2.10 and 2.12 respectively) and derive a Normal approximation to the distribution of $\widehat{\gamma}$:

$$h(\gamma, \gamma') \equiv N\left(\widehat{\gamma}, \epsilon \cdot \sigma^2_{\widehat{\gamma}}\right).$$

  where $\epsilon > 1$.

- **Independence Sampler** - *Adaptive Sampler*: Another option is to make use of the results obtained via the centered algorithms and propose γ from a distribution which approximates $\pi(\gamma|\boldsymbol{R})$. A natural approach is to use the *moment estimator* method and approximate $\pi(\gamma|\boldsymbol{R})$ with a Gamma distribution. The choice of Gamma relies on its property of having heavy tails (unlike Normal, for instance):

$$h(\gamma, \gamma') \equiv Ga(a, b).$$

The question which arises if we choose the last option, i.e. the *Adaptive Sampler* is how we will estimate a and b. We proposed to do it as follows:

1. `Run any of the` $[C]$ `algorithms for an adequate number of iterations`;

2. Get the posterior mean and variance from the obtained sample, say μ_γ and σ^2_γ;

3. Calculate a and b from:

$$a = \frac{\mu^2_\gamma}{\sigma^2_\gamma}, \qquad b = \frac{\mu_\gamma}{\sigma^2_\gamma}$$

Since we are interested in obtaining an approximation of $\pi(\gamma|\boldsymbol{R})$ the choice of the centered algorithm in Step 1 does not significantly affect the performance of this approach.

The main advantage of the NC parameterisations is that once we update γ, then the infection times ($\boldsymbol{I}$) are updated as well via the appropriate transformation. Therefore, if high acceptance rates for γ could be obtained via a well chosen independence sampler, then the infection times will be updated more often compared to a standard random walk Metropolis. Nevertheless, we have to be very careful with the independence samplers for reasons already explained in Section 2.3.3.

Another way of looking at the proposed approach for updating γ is to see it as a trick of a *joint block* update of γ and the infection times ($\boldsymbol{I}$). The more efficiently we draw samples of $(\gamma, \boldsymbol{I})$ from Steps $1i$ and $1ii$, then the faster the convergence of the MCMC algorithm will be. Concluding, any of the proposed samplers can be tried and the choice should depend on their relative efficiency.

Efficient Partially Non−Centered MCMC Algorithm [EPNC]

(Repeat the following steps)

1. `Start the chain with initial values` $\boldsymbol{I}^0$;
2. `Obtain a sample of` $(\boldsymbol{I}^1, \gamma^1)$ `via an appropriately chosen centered algorithm`;
3. `Set` $U_i = \gamma^1(R_i - I_i^1)$ `for` $i \in U$;
4. `Update` γ `using Metropolis Hastings algorithm by proposing from` $h(\gamma, \gamma')$;

Table 2.4 shows the nomenclature for the efficiently partially non−centered algorithms which were introduced in this section.

Table 2.4: Nomenclature for the EPNC MCMC algorithms

Algorithm	**Nomenclature**	**Update γ**
$\delta\%$ EPNC $(\beta_0, \gamma, \boldsymbol{I}^{\mathcal{C}}, \boldsymbol{I}^{\mathcal{U}})$	$[\delta\% \ EPNC]$	IS
$\delta\%$ EPNC with β_0 integrated out	$[\delta\% \ EPNC - \beta_0]$	IS
$\delta\%$ EPNC with β_0 and γ integrated out	$[\delta\% \ EPNC_1 - \beta_0 - \gamma]$	RWM
$\delta\%$ EPNC with β_0 and γ integrated out	$[\delta\% \ EPNC_2 - \beta_0 - \gamma]$	IS

2.6 An Extensive Simulation Study

When we are concerned with implementing an MCMC algorithm in order to analyse a real dataset, we have to make a range of various decisions to choose in advance the algorithm which offers the best performance.

First of all, a decision has to be made which class of algorithms (centered or non−centered) shall we focus on. Apart from this crucial initial choice, other decisions have to be taken. This is due to existence of a variety of MCMC algorithms (see for example, Table 2.1). Moreover, due to the availability of many variations of each of these algorithms (see for instance, Section 2.6.2.1), a careful choice of the most efficient algorithm should be made having taken into account the algorithm's performance in association with the cpu time needed.

Therefore, in this section we are interested in assessing the efficiency and drawing conclusions about the performance of each of the algorithms shown in Tables, 2.1, 2.2, 2.3, 2.4. The comparison is done through a simulation study which is presented in detail in this section. First, we describe how the data have been simulated (Section 2.6.1). Then, we focus on how should the infection times be updated (Section 2.6.2.1) and also provide results for each of the centered algorithms presented in Table 2.1. We give an explanation why the centered algorithms do not perform very well especially when the size of the data set increases.

Furthermore, we will show that the centered reparameterisations do not improve significantly the centered algorithms (and the variations), while on the other hand most of the (partially) non−centered parameterisations offer significantly better-mixing algorithms.

2.6.1 The Data

Throughout this section we consider 3 different simulated datasets. Each of them consists of 501 individuals, $\mathcal{N} = 500$ initially susceptibles and $a = 1$ initially

infective, uniformly located in a square $[0,1] \times [0,1]$ (see Figure 2.2). The datasets have been simulated from the following HMSE model (as defined in Section 2.2):

$$\beta_{ij} = \beta_0 \exp\{-\rho(i,j)\}$$
$$R_i - I_i \sim \text{Gamma}(\alpha, \gamma)$$

for $i, j = 1, \ldots, \mathcal{N}$ and let $\rho(i,j)$ denote the Euclidean distance between individuals i and j. It easy to see that for $\alpha = 1$ and $\beta_{ij} = \beta_0$, such a model is equivalent to the GSE. Table 2.5 presents the true values of the model parameters for the different datasets. Note that for all the simulated datasets, the (true) average infectious period (α/γ) of an individual is equal to 2, while the variances (α/γ^2) are equal to 8, 2, 0.8 for datasets D1, D2 and D3 respectively (see Figure 2.3 for the a graphical visualization of the corresponding distributions). Note that the datasets consist of fairly similar final epidemic sizes. For simplicity in the calculations we have assumed that α is known and we are only interested in drawing inference for β_0 and γ.

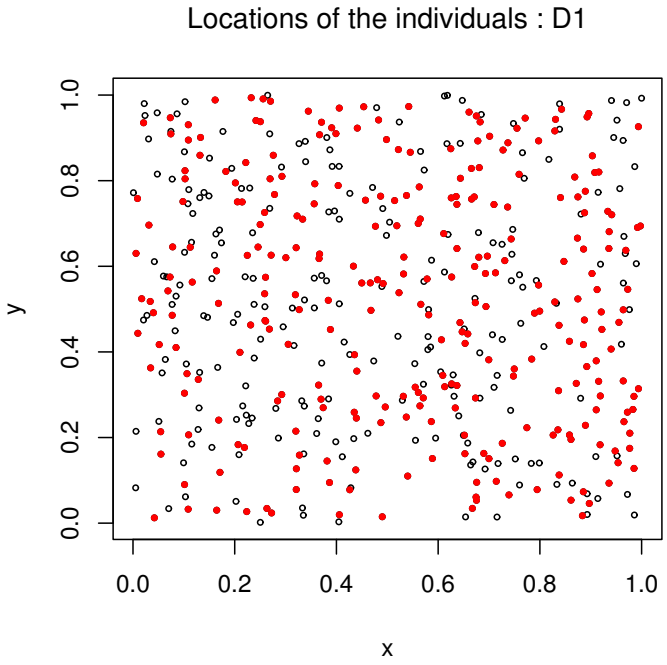

Figure 2.2: The locations of the 501 susceptibles individuals. Red dots denote the infected individuals of the dataset D1.

Table 2.5: Three simulated datasets with different infectious period

Simulated Dataset	**D1**	**D2**	**D3**
True β_0	0.0025	0.0025	0.0025
True α	0.5	2	5
True γ	0.25	1	2.5
$n_I = n_R$	284	275	286

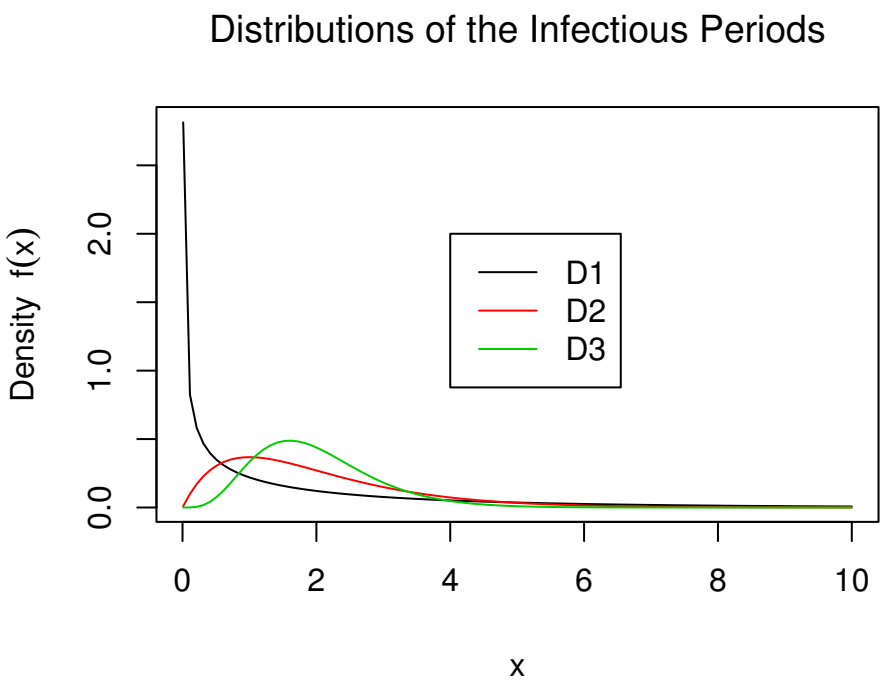

Figure 2.3: The distributions of the infectious periods for the simulated data sets 1 (black), 2 (red), 3 (green).

2.6.2 Centered Algorithms

In this section, we examine the performance of the centered algorithms and modifications of the latter. First, we focus on how the infection times can be updated and then we present results for each of them

2.6.2.1 Updating the Infection Times

Each of the centered algorithms in Table 2.1 involves a step which refers to an update of the infection times. Such a step can be simply performed by choosing at random one infection time (out of the n_I) and propose to update it by leaving the others fixed (*random scan*). Instead, we can repeat this step many times (*deterministic scan*) or we could also try to update them as block (*block update*).

Our decision should be mainly based on which approach provides a well-mixing

Markov chain. Nevertheless, the cpu time needed for the implementation any MCMC algorithm should also be considered before making any decision. In general, the required cpu time will mainly depend on the dimension of the missing data and the model parameters which makes any decision problem-specific. In this working example, we tried various sampling schemes regarding the update of the missing data: a *random scan*, a *j% deterministic scan* for $j = 10, 50, 100$ and a *block update*.

Once we have chosen the sampling scheme for the infection times, another decision about their proposal distribution should be made. A simple random walk Metropolis update for each of the infection times I_i, $i = 1, \ldots, n_I$ will require a careful tuning of many parameters which also in practice turns out to be very hard especially when n_I increases. On other the hand, such a problem of tuning is overcome by adopting the independence sampler, as described in Section 2.2.3:

$$q(R_j - I_j, R_j - I'_j) \equiv \text{Gamma}(\alpha, \gamma) \quad j = 1, \ldots, n_I.$$

As it has been already mentioned (Section 2.2.3), when the algorithms $[C - \gamma]$ and $[C - \beta_0 - \gamma]$ are applied the above proposal cannot be used explicitly and needs a further modification:

$$q(R_j - I_j, R_j - I'_j) \equiv \text{Gamma}(\alpha, \gamma^f) \quad j = 1, \ldots, n_I.$$

The proposed methods for sampling a new infection time as described in Section 2.2.3 were both used. First, a pilot study was ran to obtain a point estimate of the mode of the posterior distribution of γ, say γ^{mode}. Then, we set $\gamma^f = \gamma^{mode}$ (Algorithms $[C_1 - \gamma]$ and $[C_1 - \beta_0 - \gamma]$).

Alternatively, we assigned to γ^f the MLE of γ using the current values of the infection times, I_i^c, $i = 1, \ldots, n - 1$ at each step of the algorithm. Note that the latter proposal changes at every iteration whilst the former does not and also can

be characterised as "pseudo"-independence sampler because the proposed infection time depends, indirectly, to the current value (Algorithms $[C_2-\gamma]$ and $[C_2-\beta_0-\gamma]$).

2.6.3 Preliminary Findings

The results obtained for each of the presented algorithms in this chapter are obtained by running each of them for $200,000$ iterations and excluding the first $10,000$ which are considered as *burn in*. Since similar results were obtained for the parameters β_0 and γ we restrict our attention to γ only and where necessary to the average infection time ($\overline{I}$), as a summary statistic of the missing data instead of looking at any individual-specific infection time. Obviously, other measures of location such as medians or means, can also be used.

O'Neill and Roberts (1999) and Neal and Roberts (2005) used *random scan* to update the infection times. Nevertheless, they deal with relatively small populations ($n_I = 5$ and $n_I = 82$ infected individuals) and this does not cause any severe mixing problems. Within our example, the total population $\mathcal{N}$ and the actual population of infected individuals n_I are much larger (see Table 2.5). Note that in this case, by *random scan*, at each iteration we choose one infection time at random and propose to update it whilst the rest $(n_I - 1)$ remain fixed.

Figure 2.4 shows clearly that the size of the epidemic does not allow us to use *random scan*. Instead, if we choose at each iteration to update 10% of the missing via deterministic scan update, then the mixing is improving and is getting even better when we update half of them. Block updating of the infection performs really poorly. Out of $200,000$ moves of the Markov chain, only 47 were accepted. On the other hand, within a 100% deterministic scan scheme, on average the acceptance probability of the independence sampler was over 90%. However, we have to bring to attention that the more infection times we choose to update in each iteration then the more cpu time the algorithm needs to run. In addition, note that these results refer to only one of the simulated datasets (dataset 1).

Nevertheless, qualitatively similar results were obtained regarding the comparison of the different proportions of *deterministic scan* when the same algorithms were applied to the datasets 2 and 3 and therefore the details are omitted. Taking into account all these factors, we decide to choose a 10% *deterministic scan* update for all the centered algorithms.

Following Neal and Roberts (2005) we only focus on the parameter γ since the behavior of the parameters β_0 and $\overline{I}$ is pretty similar. Furthermore, to measure the efficiency of the different centered algorithms we compute the IAT (*integrated autocorrelation time*, see Section 1.10) for each of the algorithms in Table 2.1 and we also draw the corresponding ACFs.

Table 2.6 shows that when the variance of the infectious period decreases the performance of the algorithms deteriorates. Such a result is consistent for all the variations of the standard centered algorithm (looking at the Table horizontally). In other words, regardless the centered algorithm which is used, when a more informative infectious period is assumed (such as in dataset 3) the mixing of the Markov chains is worse than the case of a less informative infectious period (such as in dataset 1).

Apart from concentrating on the integrated autocorrelation time we also pay attention to the behavior of the ACFs which are shown in Figure 2.5, 2.6 and 2.7. These figures reveal that the more informative the infectious period is, the worse the performance of centered algorithms becomes. The same argument holds for all the variations of the standard (centered) algorithm.

The effect of integrating the model parameters (β_0 and γ) out can be determined by looking Table 2.6 vertically. Although, this marginalisation does not seem to hugely increase the efficiency, especially in datasets 2 and 3, nevertheless, it can be inferred that integrating both parameters out can provide, in some specific circumstances, a better-mixing Markov chain. The effect seems more significant in the case when the least informative infectious period is assumed, i.e. in the

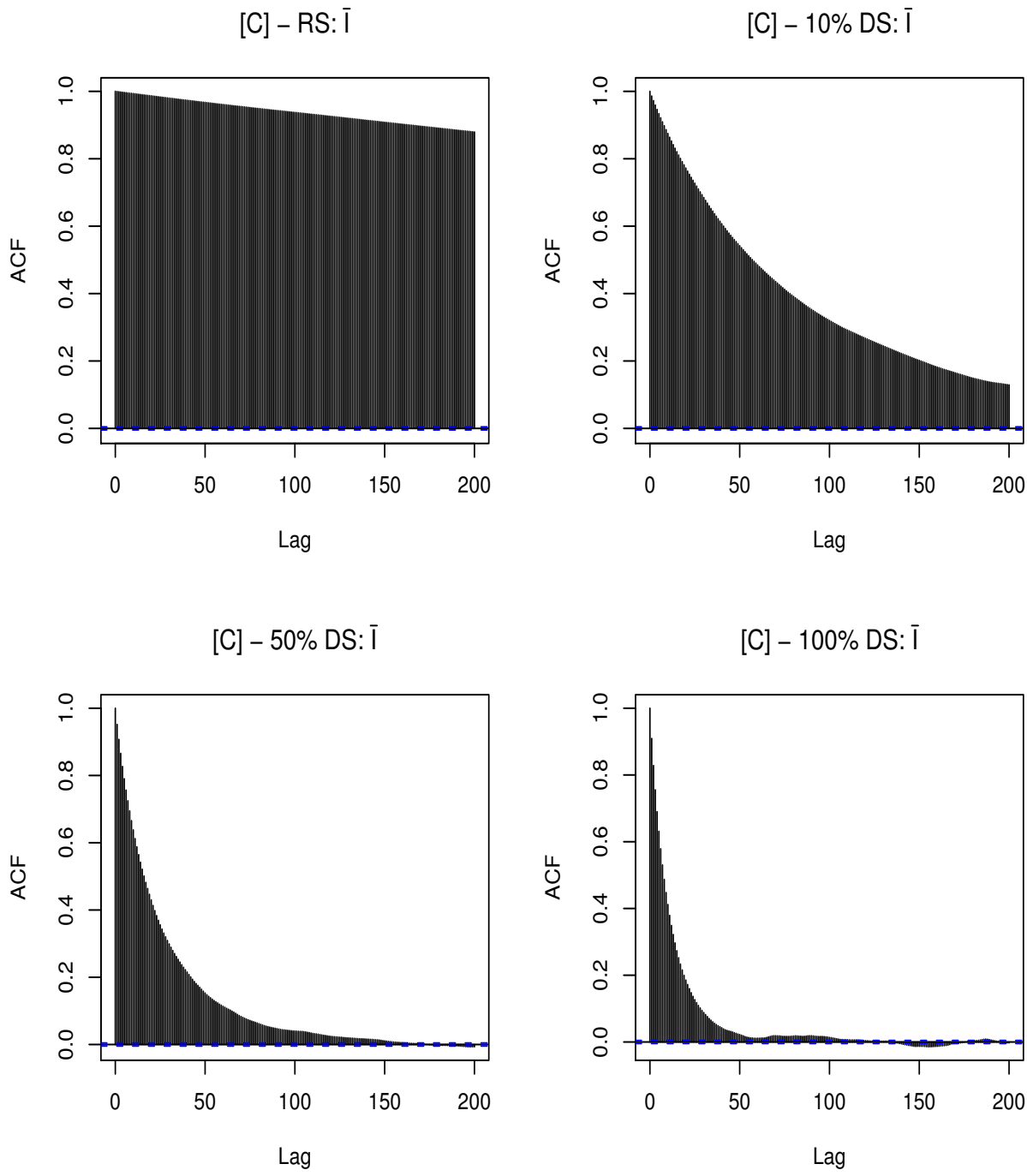

Figure 2.4: ACFs for the average infection time $\overline{I}$ using random scan (top left), 10%, 50% and 100% deterministic scan update (top right, bottom left and right, respectively) applying the standard $[C]$ algorithm to dataset 1.

dataset 1.

Although the performance of each of the centered algorithms regarding the model parameters depends on the distribution of the infectious period, on the other hand, the parameter $\psi = \beta_0/\gamma$ mixes very well regardless the dataset and the algorithm used. Note that this quantity, if $\alpha = 1$ is proportional to the R_0 (for the GSE) which can be estimated without the need of observing the infection times (missing data).

Concluding, it easy to see that although some variations of the centered algorithms perform better than the standard ones, we cannot claim that any of these can provide an adequately well-mixing Markov chain especially when the distribution of the infectious is very informative and therefore the development of algorithms with a faster rate of convergence is essential.

Table 2.6: Estimates of the integrated autocorrelation function of the parameter γ using the 10% *deterministic scan* centered algorithms for datasets D1, D2, D3

Algorithm	**D1**	**D2**	**D3**
$[C]$	130.00	275.01	338.99
$[C_1 - \beta_0]$	123.75	274.95	337.57
$[C_1 - \gamma]$	127.56	294.34	332.44
$[C_1 - \beta_0 - \gamma]$	119.99	283.12	328.82
$[C_2 - \gamma]$	125.93	280.81	339.07
$[C_2 - \beta_0 - \gamma]$	124.67	268.08	333.96

2.6.3.1 Reasons for Poor Mixing

In this section, we give an explanation why the standard centered algorithms do not provide us with good mixing Markov chains. Having obtained samples from the posterior distributions of γ and $\overline{I}$ using the standard (10% *deterministic scan*) $[C]$ algorithm, we can draw a correlation plot between these two parameters of interest for each of the three different datasets. Figure 2.9 firstly reveals that γ and $\overline{I}$ are

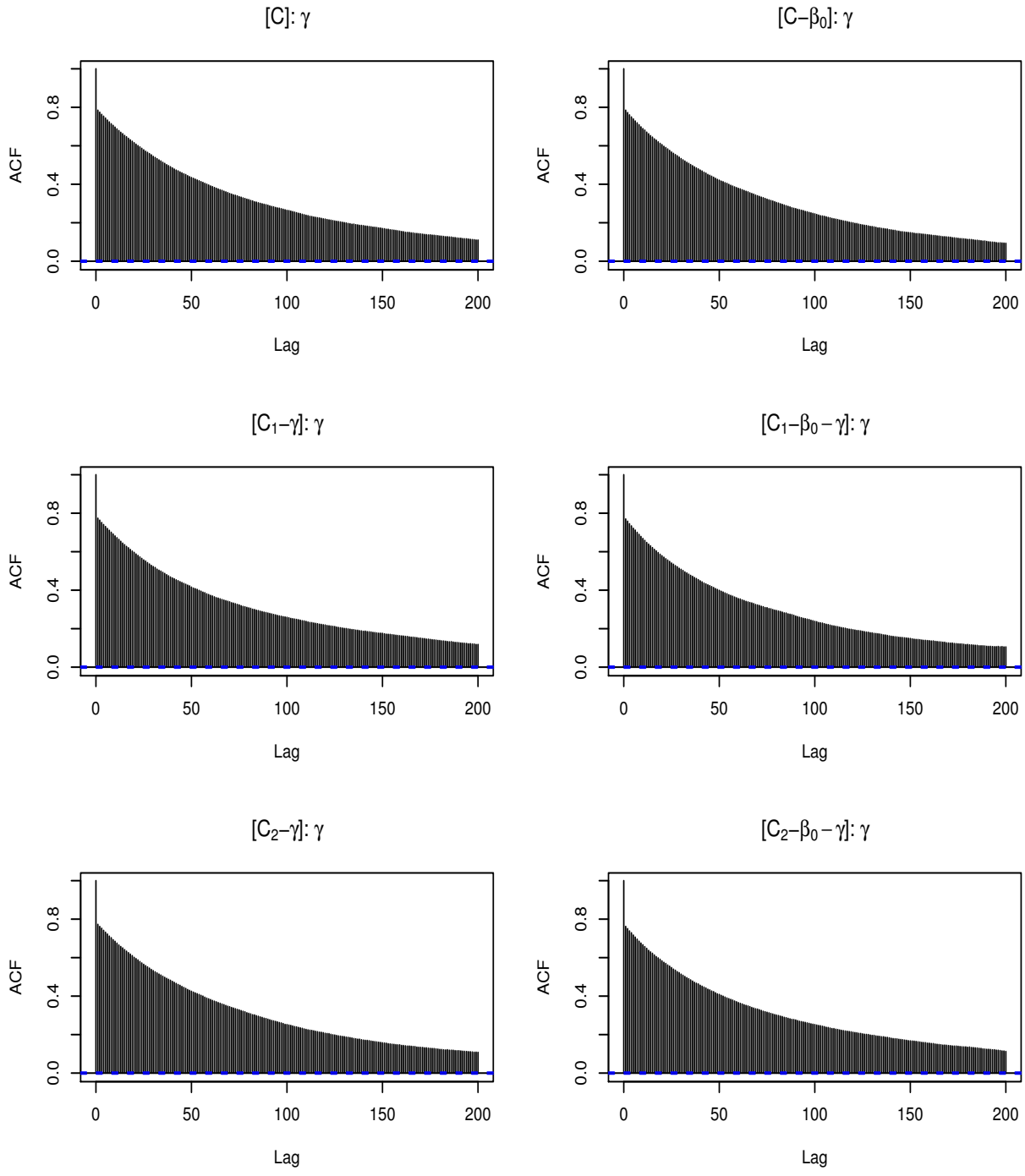

Figure 2.5: ACFs of parameter γ, using the centered algorithms presented in Table 2.1 for dataset D1.

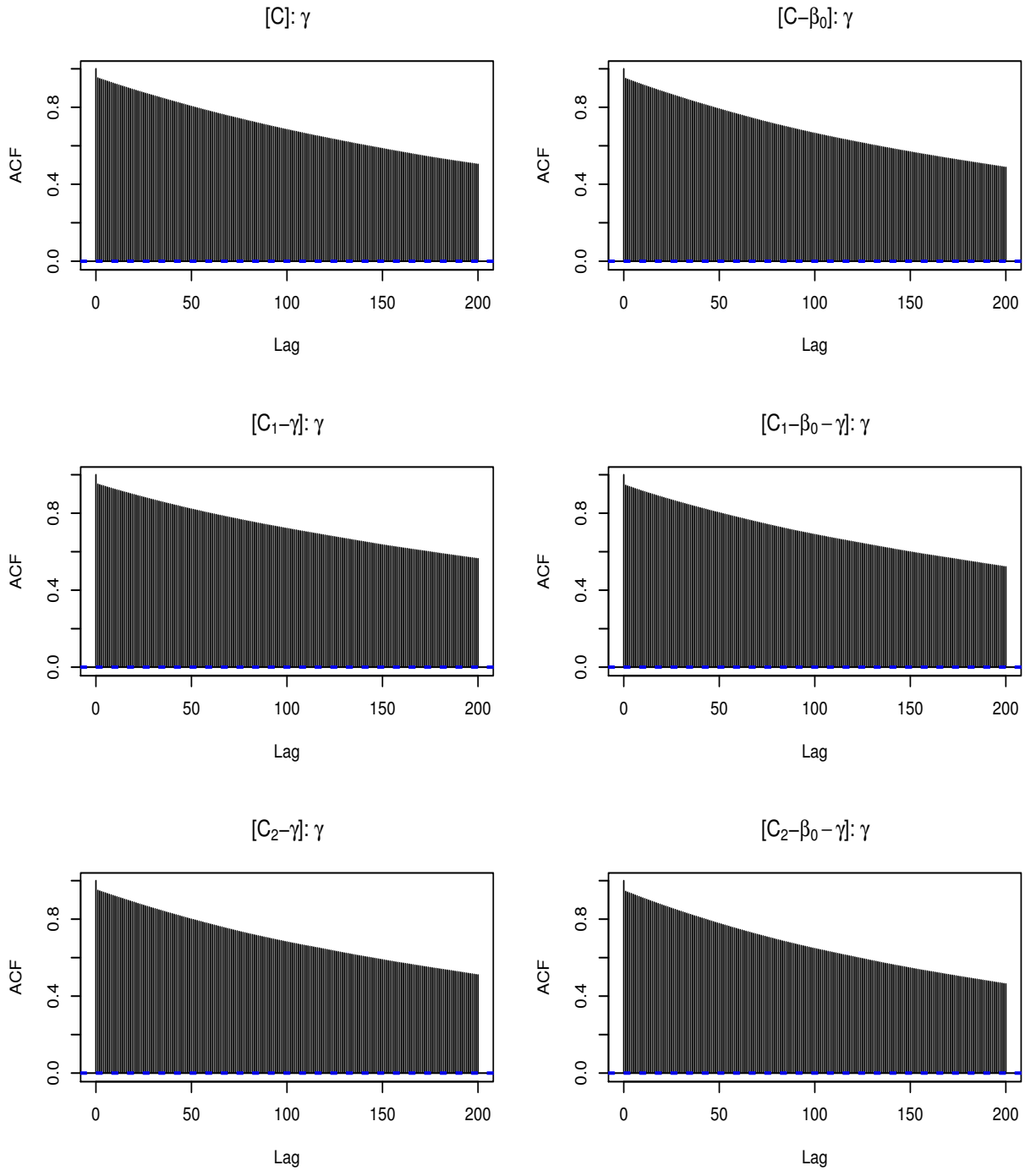

Figure 2.6: ACFs of parameter γ, using the centered algorithms presented in Table 2.1 for dataset D2.

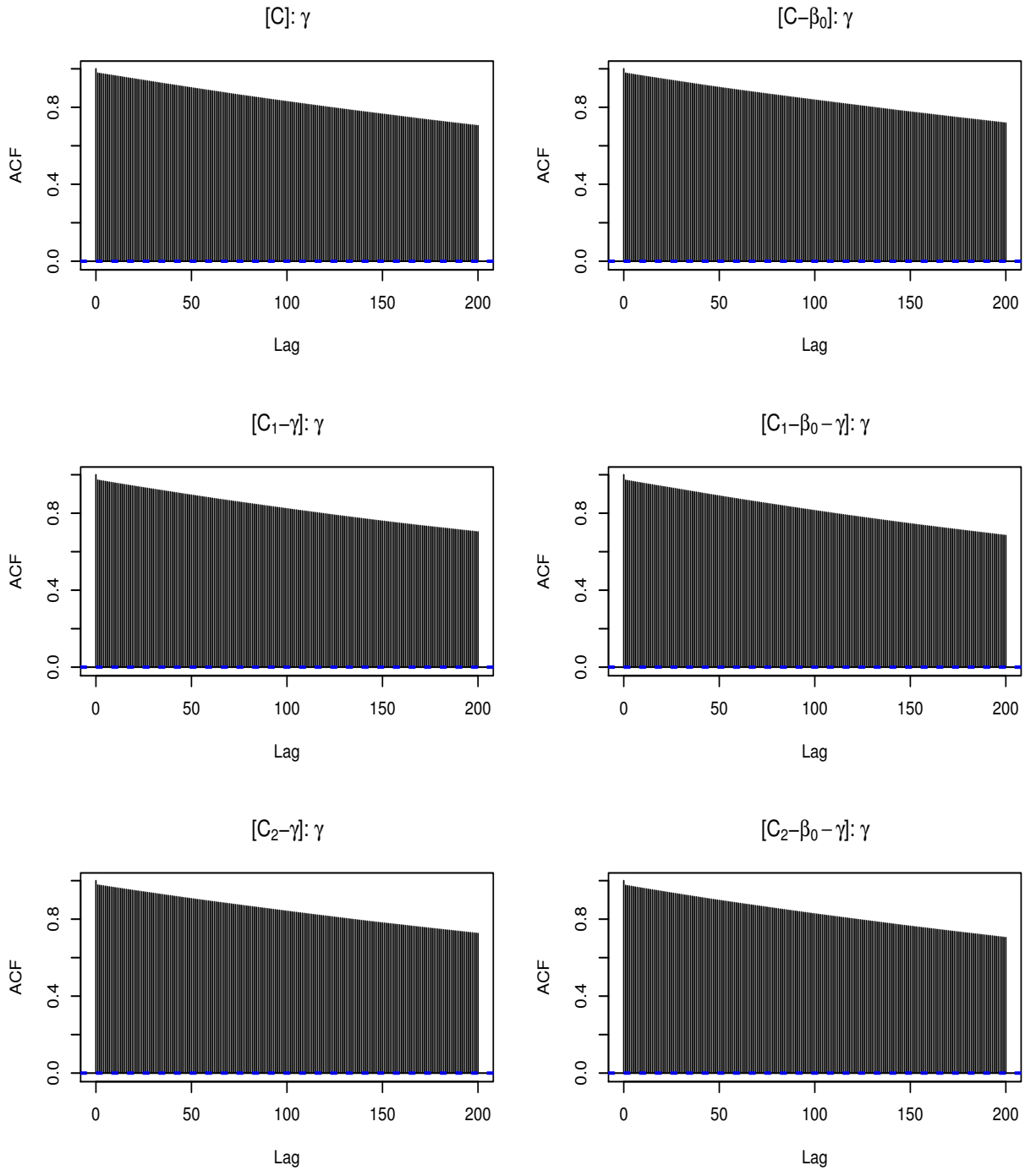

Figure 2.7: ACFs of parameter γ, using the centered algorithms presented in Table 2.1 for dataset D3.

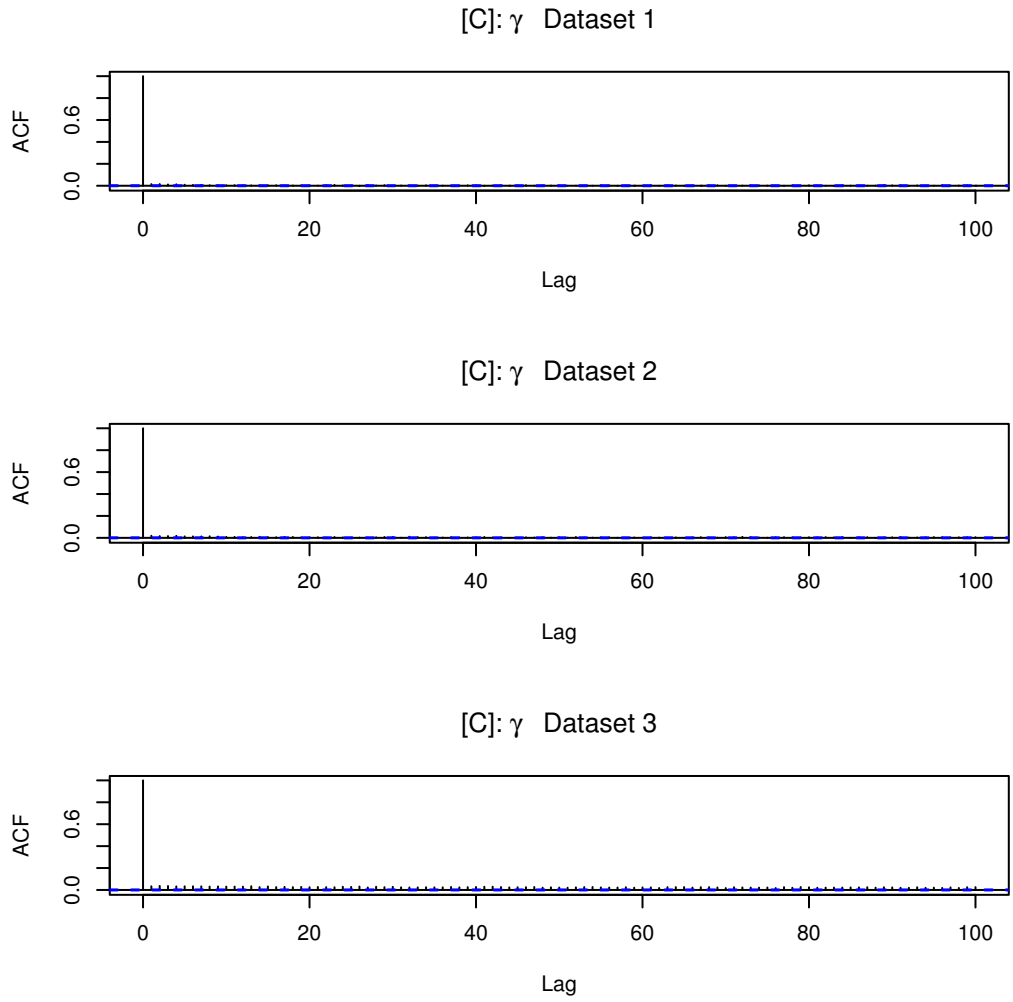

Figure 2.8: ACFs of the parameter $\psi = \beta_0/\gamma$, using samples of the parameters β_0 and γ, obtained from the $[C]$ algorithm (see Table 2.1). Each ACF refers to the different datasets (see Table 2.5).

heavily correlated regardless the dataset used to obtain the posterior samples. It also indicates that when the variance of the infectious period decreases (such as in Dataset 3) the correlation increases. Such as a strong dependence between γ and $\overline{I}$ can be explained by adopting a similar argument as in Neal and Roberts (2005):

$$\begin{aligned} R_i - I_i &\sim Ga(\alpha, \gamma), \text{ for } i = 1, \ldots, n_I \\ \sum_{i=1}^{n_I} (R_i - I_i) &\sim Ga(\alpha n_I, \gamma) \end{aligned}$$

Thus for large n_I or α, the parameter γ and the sum of the infectious periods $\sum_{i=1}^{n_I} (R_i - I_i)$ are *a-priori* heavily dependent. If these two were the parameters of interest, then this *a-priori* correlation would have caused mixing problems in the case of a two-state Gibbs sampler since it is well known that the convergence of the algorithm is linked to the correlation between parameters, see Amit (1991) and Roberts and Sahu (1997). However, things are more complicated in practice since the MCMC schemes used so far involved a deterministic scan update of the each of infection times I_i, $i = 1, \ldots, n_I$.

Things deteriorate if both α and n_I are large. Intuitively, the more correlated the two parameters are *a-priori*, then the more we expect them to get *a-posteriori* and this leads to a poor mixing of the MCMC. It is clear that the strong correlation between each of the model parameters (β_0,γ) and the missing data ($\boldsymbol{I}$) needs to be broken in order to improve the mixing of currently used MCMC algorithms.

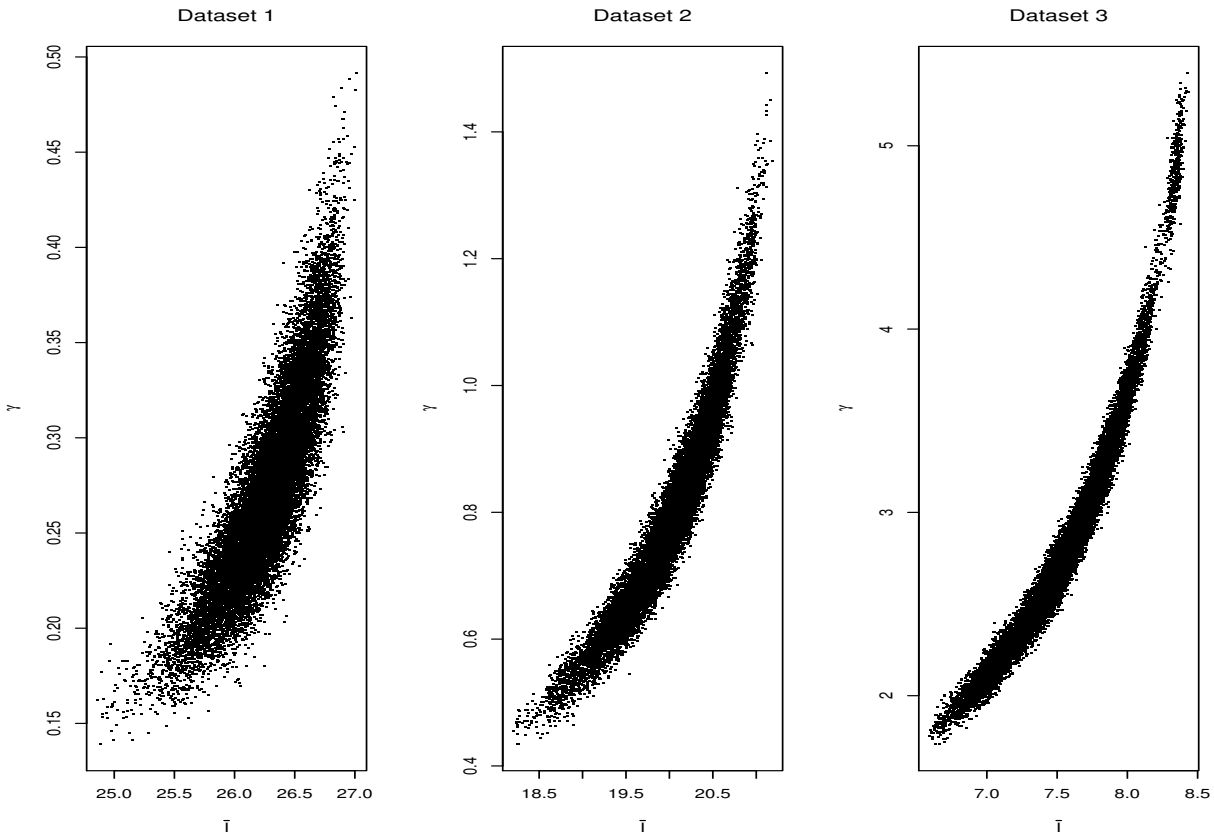

Figure 2.9: Scatter plot between γ and $\overline{I}$ obtained from their posterior samples using the $[C]$ algorithm for each of the three different datasets.

2.6.4 Algorithms Based on Centered Reparameterisations

In this section we are interested in drawing samples from the posterior distributions of the parameters of interest, such as β_0 and γ via the centered reparameterisations which have been discussed in Section 2.3. First, we apply the $[C_\psi]$ algorithm which does not involve any additional computational cost compared to any of the centered algorithms $[C]$ (see Table 2.1). Then, we apply the algorithms where either ψ or γ is integrated out from the full posterior distribution, i.e. the algorithms $[C_\psi - \psi]$ and $[C_\psi - \gamma]$ respectively.

Regarding the infection times, we update them in the same way as we did in Section 2.6, i.e. a 10% *deterministic scan* in order to facilitate a fair comparison between the centered−reparameterised and the standard algorithms. Furthermore, when γ

(or ψ) is updated via Metropolis-Hastings as in the $[C_\psi - \psi]$ (or $[C_\psi - \gamma]$) algorithm, a multiplicative random walk Metropolis step was chosen. Appropriated tuning of the variance of the random walks was chosen such that an approximately $20\% - 25\%$ was achieved for both of the parameters (see for example, Roberts et al., 1997). When the γ is integrated out, we update the infection times in a similar manner as we did when we used the $[C_2 - \gamma]$ algorithm (see Sections 2.6.2.1 2.6). Note that by using any of the algorithms in Table 2.2 values for β_0 are obtained from the resultant sample of (ψ, γ), since $\beta_0 = \psi\gamma$.

Table 2.7 gives the *integrated autocorrelation time* (see Section 1.10) for the algorithms presented in Table 2.2. The clear message from these results, is that in most of the cases, centered reparameterisation does not significantly improve the existing centered algorithms. Moreover, some of the algorithms under the centered reparameterisations (see Table 2.2) perform slightly worse than the standard (see Table 2.1); for instance, $[C_\psi - \psi]$ algorithm. An explanation for its poor performance is that by integrating ψ out, γ and $\boldsymbol{I}$ become more (conditionally) dependent which results in poorly mixing Markov chains (see also Figure 2.4). Nevertheless, under the assumption of an uninformative infectious period, such the one assumed for the simulated dataset 1, $[C_\psi - \psi - \gamma]$ gives comparable results to $[C_1 - \beta_0 - \gamma]$ algorithm. Intuitively, the reason for not observing a significant improvement by applying MCMC algorithms under centered reparameterisations could be explained as follows. By transforming from $(\beta_0, \gamma, \boldsymbol{I})$ to $(\psi, \gamma, \boldsymbol{I})$ we manage to break the conditional dependence between γ and β_0 but not the dependence between γ and $\boldsymbol{I}$.

Table 2.7: Estimates of the integrated autocorrelation function of γ using the centered reparameterised algorithms for datasets D1, D2, D3

Algorithm	**D1**	**D2**	**D3**
$[C_\psi]$	139.55	263.08	334.72
$[C_\psi - \psi]$	148.43	288.72	334.43
$[C_\psi - \gamma]$	119.58	266.55	342.61

2.6.5 Non-Centered Algorithms

We revisit the working example (see also Sections 2.6, 2.6.5) and we are interested in applying the various PNC which already have been proposed by Neal and Roberts (2005) as well as the EPNC algorithms proposed in Section 2.5. In this section we will compare the performance of the various centered and non−centered algorithms using the simulated datasets shown in Table 2.5 and running the MCMC algorithms for the same number of iterations as we did for the centered algorithms (see also Section 2.6). Papaspiliopoulos et al. (2003, Sec. 4) showed that a PNC algorithm in some specific cases can perform better than a fully NC algorithm and therefore in our simulation study we consider different proportions of non−centering, i.e. 10%, 30%, 50%, 70% and 90% to study the effect of the amount of non−centering adopted in the different algorithms. The *integrated autocorrelation time* is also computed and is used as a measure of assessing and comparing the various approaches. We should note that although a wide range of options were proposed to update the parameter γ, for simplicity, we decide to update γ using the "Pseudo-Gibbs" approach.

Table 2.8 is very informative about the merits of the NCP in an epidemic context. The results show that the more informative the infectious period is, i.e. α increases, the optimal algorithm becomes increasingly non−centered. For $\alpha = 0.5, 2.0, 5.0$ the optimal non−centered algorithms have $\delta = 0.7, 0.9, 0.9$. A similar conclusion was

also drawn from the simulation study which was performed in Neal and Roberts (2005, Sec 6.2). Although, they considered a homogeneously mixing population, they showed that as α increases, δ increases too. Note that regardless of the distribution of the infectious period an appropriate non−centered algorithm out performs the standard centered algorithms.

In addition, Table 2.8 gives us the ability to compare the variations of the standard PNC algorithms. There is some evidence that if the "best" centered algorithm is chosen rather than always the $[C-\beta_0]$ to implement the second Step of the EPNC algorithm, can provide at least a comparable or in some specific cases a better mixing algorithm. This is the case, for instance, when the dataset 1 is used, and the $[EPNC_1-\beta_0-\gamma]$ out performs $[C-\beta_0]$. On the other hand, this approach does not seem to offer much improvement for infectious period assumed in datasets 2 and 3.

Furthermore, Table 2.8 allows us to determine the effect of choosing an alternative way to update γ than by RWM as is done in the $[PNC-\beta_0]$. Recall that the comparison is done only for the "Pseudo-Gibbs" sampler. It turns out that when less non−centering is chosen then such an independence sampler behaves better than the RWM. This is not the case, when a high proportion of non−centering is needed where the IS offers badly-mixing Markov chains similar to the ones obtained via a centered algorithm.

Concluding, we should mention that Table 2.8 does not take into account the computational time needed to implement a centered or a non-centered algorithm with the latter being two times slower than the former. This is due to the fact, that within the non−centered framework we need to update γ via a Metropolis Hastings step and this requires computation of the quantities:

$$\prod_{i=1,i\neq k}^{n_I}\left(\sum_{j\in Y_i^U} d_{ji}\right) \quad \text{and} \quad \sum_{i=1}^{n_I}\sum_{j=1}^{\mathcal{N}} d_{ij}(R_i\wedge I_j - I_j\wedge I_i)$$

which when they are computed from scratch are quite costly, especially when

n_I or/and $\mathcal{N}$ are large Nevertheless, even when the cpu is time is taken into account, the PNC algorithms still offer a better mixing of the Markov chain than the centered ones.

We should also mention that we decided to choose a 10% deterministic scan to draw samples of $(\boldsymbol{I}, \gamma)$ via the centered algorithms. Unlike Neal and Roberts (2005) who argued that repeating the Step 3 in the PNC algorithm (see Section 2.4.3) did not improve the efficiency considerably, in contrast, when samples of $(\boldsymbol{I}, \gamma)$ were drawn via a 100% deterministic scan the mixing was significantly better. Nevertheless, there has been a considerable increase in the computational cost. That is, the cpu time needed to run the 100% deterministic scan algorithm increases when the final size of the epidemic becomes larger. Therefore, the choice of the percentage of deterministic scan should be made based on the actual cpu time needed to run the corresponding algorithm.

Table 2.8: Estimates of the integrated autocorrelation time of the parameter γ for the different PNC and EPNC algorithms

Algorithm	**Update** γ	10%NC	30 %NC	50 %NC	70% NC	90% NC
Dataset 1						
$[C]$	–	130.00				
$[PNC]$	RWM	125.38	96.11	95.08	115.38	183.72
$[PNC - \beta_0]$	RWM	126.68	95.85	84.35	104.09	180.11
$[EPNC]$	IS	103.71	99.05	114.42	170.05	234.04
$[EPNC - \beta_0]$	IS	109.21	95.40	108.83	191.55	238.79
$[EPNC_1 - \beta_0 - \gamma]$	RWM	114.09	77.32	76.59	65.80	70.59
$[EPNC_2 - \beta_0 - \gamma]$	IS	116.40	88.61	83.30	85.01	95.80
Dataset 2						
$[C]$	–	275.01				
$[PNC]$	RWM	271.58	229.41	174.39	123.82	58.81
$[PNC - \beta_0]$	RWM	272.79	241.94	192.79	108.12	56.38
$[EPNC]$	IS	266.97	225.75	224.66	225.07	239.00
$[EPNC - \beta_0]$	IS	268.96	234.44	217.21	211.92	241.11
$[EPNC_1 - \beta_0 - \gamma]$	RWM	279.36	234.60	170.72	102.11	62.74
$[EPNC_2 - \beta_0 - \gamma]$	IS	267.47	244.95	211.61	198.43	195.40
Dataset 3						
$[C]$	–	338.99				
$[PNC]$	RWM	325.90	292.68	244.16	158.22	62.51
$[PNC - \beta_0]$	RWM	324.96	297.11	235.98	165.05	61.19
$[EPNC]$	IS	321.14	304.04	278.84	274.74	279.67
$[EPNC - \beta_0]$	IS	323.29	299.01	285.95	264.03	273.92
$[EPNC_1 - \beta_0 - \gamma]$	RWM	340.25	302.32	236.05	163.21	65.75
$[EPNC_2 - \beta_0 - \gamma]$	IS	337.35	309.76	281.03	277.18	244.83

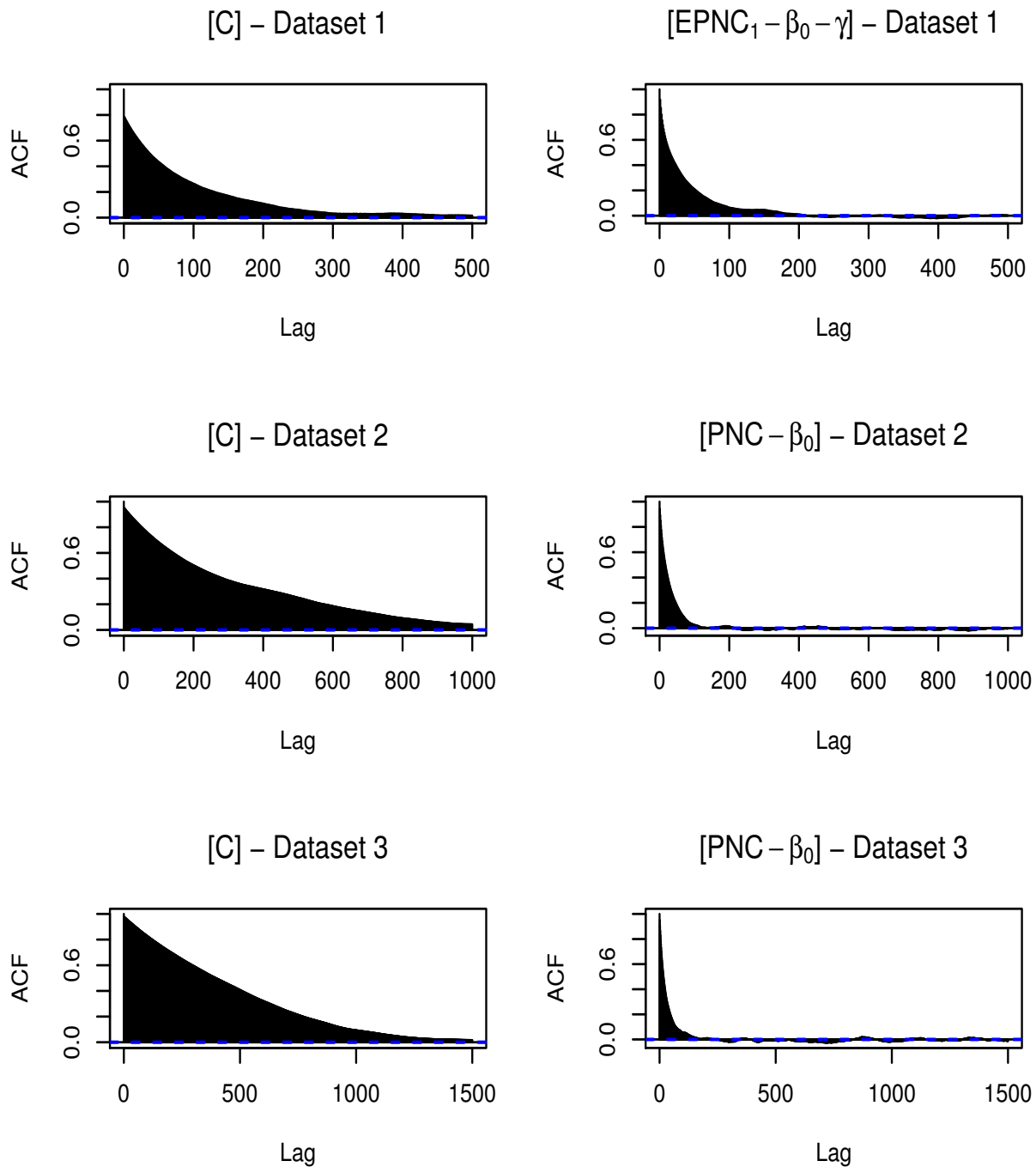

Figure 2.10: Comparison of ACFs of γ between the centered and the optimal PNC algorithm for the different datasets. Details on the nomenclature of the algorithms are given in Tables 2.3 and 2.4

2.6.6 Conclusions

The extensive simulation study presented in this section offers a useful guide on the choice of MCMC algorithms for partially observed stochastic epidemics.

The main finding of this simulation study is that when the size of the dataset increases and the variance of the infectious period decreases then the centered algorithms perform badly and a non−centered algorithm should be considered instead. Note that this is very important, since in many real applications such as Foot and Mouth or Avian Influenza the size of the population is much larger than the one in the simulated datasets.

Another finding of the study is that the algorithms based on centered parameterisations, do not offer much improvement. On the other hand, non−centered algorithms offer significantly better mixing, especially when the dependence between the infection times and the model parameters is high. As we have shown, this is often the case when the number of the individuals who ultimately contracted the disease (n_I) is relatively large. In addition we have observed that modifications and extensions of the currently available in the literature NC algorithms for stochastic epidemics (Neal and Roberts, 2005), can lead to well mixing algorithms.

In practice, if we are interested in designing an MCMC algorithm to draw inference for a partially observed a stochastic epidemic model then we should take into account the following. First, it is preferable to use a (repeated) single site update (*deterministic scan*) rather than a block update for the infection times. Moreover, it turns out that about 10% of them should be updated at each step of the algorithm (see also Neal and Roberts (2006)) since performing a 100% deterministic scan can be very computationally costly without gaining much more relative efficiency. Also, the proposal which is taken from the likelihood equation of the model should be used rather than a random walk Metropolis in order to avoid problems of tuning.

The crucial choice refers to which NC algorithm out of those shown in Sections

2.4 and 2.5 should someone choose. The results obtained in this section, suggest that if the variance of the infectious period is small (i.e. unobserved data more informative about the parameters) then δ (percentage of non−centering) should be large. On other the hand, if the infectious period has large variability then δ should be small. Having decided on the value of δ we can then choose the algorithm which offers the best mixing. For instance, if δ needs to be large, then the $[PNC]$ should be preferred, while on the other hand if δ is small, then the $[EPNC]$ should be in favor. Concluding, we should note that any choice should be made with care and the reader is also referred to Table 2.8 where the performance of each of the algorithms is shown.

2.7 Discussion

In this chapter we have focused on drawing inference for a stochastic epidemic model which extends the GSE in two ways. Apart from assuming a more general and epidemiologically motivated infectious period, the model presented in Section 2.2 allows for heterogeneity in the individuals by assuming an individual−specific infection rate. In other words, unlike the GSE which considers a common infection rate β, we suppose that an infected individual i makes a contact with the susceptible individual j with rate $\beta_{ij} = \beta_0 \cdot h_{ij}$. The deterministic function can incorporate the various characteristics of the individual and also the structure of the population; recall Section 2.2.1 for the formulation of β_{ij}.

We assumed that only the removal times are observed and the infection ($\boldsymbol{I}$) times of the individuals are unknown. We first presented the existing (centered) MCMC algorithms which have been proposed in the literature for partially observed stochastic epidemics. Although various modifications have been proposed, regarding the target distribution for which the MCMC algorithm is designed and the proposal distribution for the infection times, we show that the mixing of the centered algorithms deteriorates when the of the individuals becomes more informative and as the number of infected individuals increases. The simulation study which was performed considering different infectious periods indicated a high correlation between (γ) and the infection times ($\boldsymbol{I}$) which increases as the variance of the infectious period decreases.

The key property of the basic reproduction number R_0 that can be estimated without the need of observing the infection times, unlike the infection and the removal rate, allowed us to introduce a centered reparameterisation which involves a quantity proportional to R_0. Nevertheless, although such reparameterisations manage to break the marginal dependence between β_0 and γ do not succeed in eliminating the dependence between γ and $\boldsymbol{I}$. Therefore, this explains intuitively why MCMC algorithms under such centered reparameterisation do not significantly improve

the standard centered algorithms.

Since the strong *a posteriori* correlation between γ and $\boldsymbol{I}$ which causes problem of mixing has to be broken we applied the non−centered (NC) methodology (including partially non−centered algorithms) for stochastic epidemics which has been firstly introduced by Neal and Roberts (2005). Although this methodology has been derived in terms of a homogeneously mixing population we showed that it can be easily implemented for the HMSE as well.

A crucial difference between the GSE and the HMSE is that by construction the observed data for the latter are more informative about the parameters than the former. For example, suppose that an HMSE model associates the infection rate with the distance among individuals. It is then likely, that an individual will become infected from those who are in short distance. This is not the case in the GSE, where any infective individual can infect any other susceptible with the same probability, regardless its characteristics such as location etc.

The above explanation could lead to the argument that a NCP does not have much to offer in comparison the CP. Nevertheless, it can be shown by simulation that the need of a NCP is also essential in the case of heterogeneously mixing population. Further improvements on these algorithms have been obtained by extending the current methodology and deriving *efficiently* non−centered parameterisations (ENCP). Such reparameterisations depending on the choice of the distribution of the infectious period, lead to considerably more rapidly convergent Markov chains than some of the NCP and in consequence better than the conventional centered algorithms even when taking into account the computational time needed.

In general, partially non−centered algorithms become more useful as the final size of the epidemic (n_I) increases. The examples studied in this chapter are concerned with relatively large initially susceptible population ($\mathcal{N} = 501$) and a large final size of the epidemic (n_I). Thus, the ENC methodology presented here is of great practical interest since most of epidemics that are of interest have final

size relatively large.

Finally, we should mention that although models such the GSE and the HMSE are relatively simple, they are very challenging for standard MCMC methods. On the other hand, even though the methodology presented in this chapter has mainly focused on the HMSE, it can be very easily extended to models with more complicated structure, such as incorporating latent periods (eg. SEIR) or models involving other states (eg. notification). The main idea is to break the prior link between missing data and the model parameters via non−centered parameterisation which will make them *a priori* independent. In conclusion, non−centering has much to offer for inference problems in epidemics due to the nature of outbreak and the need of extensive data−augmentation schemes to derive inference for realistic models.

Chapter 3

Bayesian Analysis of the 2001 UK Foot-and-Mouth epidemic.

3.1 Introduction

Foot-and-Mouth disease (FMD) is considered one of the most important of all these infections because it can spread rapidly between livestock species. In general, FMD is rarely lethal to adult livestock, but causes blisters on the mouth and feet which often leads to a significant drop in milk production in dairy cattle. It also causes very slow weight gain in other livestock (Alexandersen et al., 2003). The economic effects of infection within a country are dramatic; prevention of export of meat and milk to other countries eliminates a vital source of revenue.

The main aim of any control policy is to achieve disease-free status as quickly as possible by having the minimum impact on the livestock community. However, minimizing the time and the disturbance are often in conflict (Keeling, 2005) and determining which is the correct balance between the two is a critical decision that must be taken.

This chapter is mainly concerned with performing a fully Bayesian analysis to analyse the FMD outbreak which took place in the UK in 2001. First, we will

briefly review the models which have been used during the outbreak. Then, we will describe in a detail how an HMSE model (see Section 2.2), could capture the disease's dynamics. We will also show how the methodology presented in Chapter 2 can be applied to efficiently draw Bayesian inference for the model's parameters. Finally, we compare our findings with the conclusions obtained from other approaches on modelling the FMD outbreak and discuss limitations of our study.

3.2 Previous Work on Modeling of the 2001 FMD

Prior to 2001 there were relatively few attempts to model the spatial spread of FMD or any other livestock disease. Morris and coworkers developed a variety of spatial simulation models (Sanson, 1993, Sanson et al., 1993, 1999). Naturally, there was increasing interest in analyzing such models during the 2001 FMD outbreak. Three different models, among others, were used in order to predict the disease's dynamics, assess the existing control measures and provide information to support the decision-making process. In this section we will review the general ideas of these models and briefly mention their methodology and main findings. Following Keeling (2005), we will refer to as the Imperial model (Ferguson et al., 2001a,b), the Cambridge-Edinburgh model (Keeling et al., 2001, 2003), and InterSpread (Morris et al., 2001). We will also review the most recent work by Diggle (2006) and Deardon et al. (2006).

3.2.1 InterSpread

InterSpread (Sanson et al., 1999) is a computer program which is used to simulate (stochastically) an epidemic in discrete time. It was founded upon the research by R.S. Morris and coworkers in the early 1990s and particularly, in the Ph.D. thesis of Sanson (1993). InterSpread was used by DEFRA during the FMD outbreak in

2001, in order to predict the spread of the epidemic. Simulating an FMD epidemic via InterSpread involves representing biological processes, including their inherent variability, by sampling from statistical distributions. InterSpread initially uses as input the spatial location of all farms and animal markets in conjunction with other relevant data such as coordinates of any control zones. In addition, detailed information on the the number of cattle, sheep, pigs, goats and deer in each farm at the start of the epidemic is part of the initial input of the program. The model simulation is initiated with either the index case in the outbreak or with the sequence of specific farms already confirmed with the disease of a chosen date, which represents the start of the simulation period. The model is then used to predict the temporal and spatial spread of the infection of the disease by taking into account the different factors which can influence the spread.

The transmission mechanism adopted by InterSpread is briefly explained as follows. For each farm confirmed with FMD an estimated date of infection is determined by subtracting a species-specific incubation period from the date on which clinical signs where first detected. Each farm is assumed to be infectious on or just prior to the date of appearance of clinical signs, depending on the species present.

3.2.1.1 The Model

The transmission rate of FMD from an infectious to a susceptible farm is modeled stochastically, with the probability of infection depending on the distance between the two farms, route of transmission and the number of the different species of animals in each farm. An infectious farm makes contacts with susceptible farms via one of the following four mechanisms:

- movement of animals as a result of sales to other farms or markets,
- local spread to nearby farms due to movements of personnel,
- long-distance windborne spread if meteorological conditions are conductive to this pathway,

- spread from dairy tanker movements.

A large number of risk factors are used by InterSpread. For instance, the proportion of dairy farms with lactating dairy cattle, the maximum length of tanker routes and the probability of farm being selected for particular tanker route, are taken into account to model the probability of infection (see Tables 2 and 3 of Morris et al., 2001). A farm remains infectious until control measures have been completed and varies according to the stage of the disease process and the adoption of control policies. A set of parameter values is then chosen (see Sanson, 1994, and the references therein) to predict the spread of the disease.

3.2.1.2 Methods and Results

In February 2001, when the epidemic was first detected in the UK, the InterSpread model was used by DEFRA to predict the spread of the infection (Morris et al., 2001). One of its primary uses was to compare short-time model predictions with the observed cases in order to target specific areas for discovering cases which need further investigation.

Moreover, InterSpread was used to evaluate the various control strategies adopted by DEFRA during the outbreak. In the first series of strategies, the effects of varying the number of farms slaughtered around each farm diagnosed as infected was assessed. In addition, in the second series of strategies, the effect of increasing the time to slaughter of farms of unknown status (pre-emptive slaughtering) was also assessed. Finally, the effect of vaccination alone as a control measure and the effectiveness of a combination of vaccination and slaughter were both assessed. Each of the specific strategies was simulated for 200 days commencing from April 10, 2001, and five iterations of each variant were produced. For each simulation, the total number of farms which became infected, the mean date of eradication and the proportion of iterations where eradication was achieved within 200 days were recorded.

Morris et al. (2001) reported that it is crucial to slaughter all susceptible animals on affected farms as rapidly as possible (after diagnosis) and to slaughter animals on high-risk farms before signs of the disease can appear. Furthermore, the authors claim that if the goal is to eradicate the disease most rapidly and most effectively, then the provision of additional recourses to allow intensification of the stamping-out policy was clearly the best solution identified by their approach.

3.2.2 The Cambridge - Edinburgh Model

The Cambridge-Edinburgh (CE) approach is to describe the dynamics of the FMD disease by developing a stochastic, explicitly spatial, "individual-based" model where individuals are represented as discrete points in time and space. The CE model is a stochastic *SEIR*−type model (in discrete time) which is initialised with the location of all the farms in the UK and their livestock as recorded at the last census. The CE model requires the same input such as InterSpread, but model's transmission mechanism is much simpler. We shall describe in detail the formulation of the CE model and then refer to the methodology used by Keeling et al. (2001) to draw inference for the model's parameters.

3.2.2.1 The Model

Each farm is classified as either susceptible, incubating (exposed), infectious or slaughtered. Heterogeneity of the farm is incorporated by allowing the susceptibility and infectiousness of farms to vary with the type and the number of livestock. Therefore, the probability of infection is associated with the number and species of animals per farm as well as with the distance between infectious and susceptible farms. Denote by $P(j,t)$ the probability that a previously uninfected (susceptible) farm j is infected within the time interval $(t, t+1]$:

$$P(j,t) = 1 - \exp\left\{-\sum_{i \in I_t} \beta_{ij}\right\} \tag{3.1}$$

where I_t denotes the set of infectious farms at time t. Moreover, β_{ij} is defined as follows:

$$\beta_{ij} = K(\rho(i,j)) \times (\epsilon n_i^c + n_i^s) \times \left(\xi n_j^c + n_j^s\right) \tag{3.2}$$

where n_i^c and n_i^s denotes the number of cattle and sheep for farm i respectively while ϵ (ξ) represents the relative infectiousness (susceptibility) of cattle to sheep. The terms $T_i = (\epsilon n_i^c + n_i^s)$ and $S_i = (\xi n_i^c + n_i^s)$ are considered the farm's infectivity and susceptibility respectively. $K(\rho(i,j))$ is the transmission kernel which shows how infectivity decreases with the distance $\rho(i,j)$ of the two farms. While livestock number and species contribute towards the rate of transmission, the latent and the infectious periods were treated as fixed without any variability or differences between farms. Each farm is considered to be exposed (incubation period) for 5 days and the period from infection to reporting is taken to be 9 days.

3.2.2.2 The Methodology

Keeling et al. (2001) used what they term as a hybrid mixture of maximum likelihood estimation and repeated stochastic simulations of model. They consider the transition from a susceptible state to an exposed state to be based on a time dependent Poisson process. Therefore, given a record of infection events at discrete time points t during the epidemic, the likelihood is simply a product over those time points so that the likelihood can easily be obtained:

$$L(\boldsymbol{\theta}) = \prod_t \left\{ \left(\prod_{i \in \text{ Susceptibles at time } t+1} (1 - P_i) \right) \left(\prod_{i \in \text{ New cases at time } t+1} P_i \right) \right\}.$$

The authors first estimated the model's parameters via maximum likelihood methods. They claimed that these estimates did not offer an adequate fit and therefore, they fitted a different model. They wanted to adjust for instance for the fact that between June 2000 (census) and February 2001, when the epidemic began, there was substantial movement of livestock, in particular the movement of sheep out

of the upland areas of Cumbria and Wales into lowland regions. Their new model takes into account that farm's infectivity and susceptibility could vary for different counties in the UK (such as Cumbria and Devon). In order to make inference for the parameters of this new model they performed least-squares fit at the county level to match both the temporal pattern of case reports as well as the regional patterns of spatial cases. The maximum likelihood estimates of their first model were used as initial guess for the adopted least-squares method.

3.2.2.3 Results

Apart from estimating parameters associated with farm's infectivity and susceptibility, culling and vaccination strategies have been examined with this model (Keeling et al., 2001, 2003) in order to assess the approach adopted by DEFRA of the culling of contiguous premises (CP). However, CP culling is very difficult to be precisely modelled since only the location of the farm is recorded and information about their neighbours (boundaries) is not available (Keeling, 2005). The Cambridge-Edinburgh model suggested that if a "well-targeted" large-scale vaccination was applied early in the epidemic, it would have been beneficial. The authors also argued that an earlier implementation of the control strategies and earlier detection of the first cases could have led to a significant reduction of the total number of infected farms.

3.2.3 The Imperial Model

The Imperial model is a deterministic SIR-type model (see Section 2.1.2.1) and treats the farm as the individual unit in a similar way to InterSpread and the CE model. Its initial version (Ferguson et al., 2001a) was formulated during the early stages of the outbreak and therefore was very simplistic. As a crude approximation, the differences between the farms (size, species etc) were ignored and the model focused only on local and long-range transmission.

Although the traditional, homogeneous−mixing, deterministic SIR model (see also Section 2.1.2.1) ignores any spatial structure, the Imperial model attempts to adjust for the spatial effect by adopting the methodology presented in Keeling (1999) for modelling the behaviour of individuals in a fixed network. The assumption made by Ferguson et al. (2001a) is that the total infectious pressure that a susceptible farm j is subjected to can be separated in two sources; pressure from the locally connected farms and pressure from the farms which are in long distance. This formulation is along the lines to the work by Ball et al. (1997), where models with two levels of mixing (local and global) are described. Before explaining in detail the Imperial model, we briefly outline the framework used to adjust for a local spatial spread.

3.2.3.1 The Network's Structure

We assume that the contact structure forms a network of links between farms, with all links being of equal strength. Such a network is often referred to as a graph. A network involving $\mathcal{N}$ farms can be described by a matrix $G \in \{0,1\}^{\mathcal{N}^2}$, where

$$G_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases}$$

As all links are bidirectional and self contact is not allowed, the following two constraints upon the matrix as placed: $G = G^T$ and $G_{ii} = 0$. The number of connected *pairs* and *triples* in the graph can be calculated as follows:

$$\text{Number of pairs } := ||G|| = n\mathcal{N}$$

$$\text{Number of triples} = \left|\left|G^2\right|\right| - \text{trace}\left(G^2\right)$$

where $||G|| = \sum_{ij} G_{ij}$ is the sum of all the elements in the matrix and n is therefore the average of neighbours per node. The number of triples is calculated as the number of nodes which are joined by two connections, given that the nodes are

distinct. A *triangle* is three linked nodes with the same start and end point. We define ϕ the ratio of triangles to triples and this is a simple measure of how interconnected the local neighbourhoods are:

$$\phi = \frac{\text{number of triangles}}{\text{number of triples}} = \frac{\text{trace}\,(G^3)}{||G^2|| - \text{trace}\,(G^2)}$$

When ϕ is large, the members of a connected pair is connected to many common nodes, whereas when ϕ is small there a few common nodes and long-range connections dominate.

In order to consider the dynamics of farms, the following set of functions which inform us about the state of each node are defined. Let

$$A_i = \begin{cases} 1 & \text{the farm at node } i \text{ is of type A} \\ 0 & \text{otherwise.} \end{cases}$$

This allows us to define rigorously the number of single, pairs and triples of each type:

$$\begin{aligned} \text{singles of type A} \quad &:= [A] \quad &= \sum_i A_i \\ \text{pairs of type A-B} \quad &:= [AB] \quad &= \sum_{ij} A_i B_j G_{ij} \\ \text{triples of type A-B-C} \quad &:= [ABC] \quad &= \sum_{ijk} A_i B_j C_k G_{ij} G_{jk} \end{aligned}$$

This method of counting means that pairs are counted once in each direction so that $[AB] = [BA]$ and that $[AA]$ is even.

For illustration, using these tools, Keeling (1999) considers the spread of a disease through a network of nodes. In particular he describes the dynamics of the disease's spread within a deterministic SIR model framework (see 2.1.2.1). Within such a framework, differential equations describe the behavior of A-B pairs instead of the behavior of individuals. Therefore, we define τ, the transmission rate across

a connection, to be β/n whereas we let $(1/\gamma)$ to be the length of the infectious period. Although there exist nine distinct types of pairs, due to symmetries and the fact that the sum over all pairs remains constant, only the five differential equations are necessary:

$$\begin{aligned}
\frac{d[SS]}{dt} &= -2\tau[SSI] \\
\frac{d[SI]}{dt} &= \tau([SSI] - [ISI] - [SI]) - \gamma[SI] \\
\frac{d[SR]}{dt} &= -\tau[RSI] + \gamma[SI] \\
\frac{d[II]}{dt} &= 2\tau([ISI] + [SI]) - 2\gamma[II] \\
\frac{d[IR]}{dt} &= \tau[RSI] + \gamma([II] - [IR])
\end{aligned} \tag{3.3}$$

Furthermore, Keeling (1999) provides estimators of R_0 and the final size of the epidemic within this context of a fixed network.

3.2.3.2 A Pair-Based Transmission Model

We shall now describe the initial model which was proposed by Ferguson et al. (2001a). They authors first were interested in quantifying the effect of a long-range infection compared to a "local" infection. Contact tracing for all FMD-affected farms (provided by DEFRA) has produced data on the spatial scale of disease transmission. The probability density function, $f(\rho(i,j))$, of the distance $\rho(i,j)$, from the source FMD-affected farm to the farm it infects was parameterised as follows:

$$f(\rho(i,j)) = p \cdot \frac{g(\rho(i,j))}{\mathcal{N}} + (1-p) \cdot K(\rho(i,j)) \tag{3.4}$$

with probability p that the infection arose uniformly over the area surrounding the index case (represent mass action mixing) and with probability $(1-p)$ that the infection arose from local spread in the proximity of the FMD-affected farm characterised by a local kernel, $K(\rho(i,j))$. The radial density of farms with sheep,

cattle and/or pigs distance $\rho(i,j)$ from the average FMD-affected is denoted by $g(\rho(i,j))$ and was determined by the data from the census (June 2000). The following parametric model of the kernel was used:

$$K(\rho(i,j)) = \frac{\exp\{-a\rho(i,j)^b\}g(\rho(i,j))}{\int_0^\infty \exp\{-a\rho(i,j)^b\}g(\rho(i,j))}. \tag{3.5}$$

Ferguson et al. (2001a) reported parameter estimates for a, b and p by fitting $f(\rho(i,j))$ to the distribution of distances $\rho(i,j)$ between identified infectious contacts.

The authors then defined a model which combined a traditional mass-action term (see Section 2.1.2) to describe initial long-range contacts, with a spatial correlation (formulated as described in Section 3.2.3.1) to adjust for local transmission and the structure of contact network between neighbouring farms. They stratified the population of farms in a susceptible class, S, sequential infection classes $I_i, i = 1, \ldots M$ and the slaughtered class, R. Classes can be seen as types according to the previously mentioned construction of the contact network. The authors, for conciseness and clarity, presented only a simple model with two infected classes: E (uninfectious) and I (infectious). Therefore, such a model can be seen as a deterministic SEIR-type model. The dynamics of the disease can be represented by the following set of differential equations:

$$\begin{aligned}
\frac{d[S]}{dt} &= -(\tau+\mu+\omega)[SI] - p\frac{\beta}{\mathcal{N}}[S][I] \\
\frac{d[E]}{dt} &= p\frac{\beta}{\mathcal{N}}[S][I] + \tau[SI] - \nu[E] - \mu[EI] \\
\frac{d[I]}{dt} &= \nu[E] - \sigma[I] - \mu[II] \\
\frac{d[SS]}{dt} &= -2(\tau+\mu+\omega)[SII] - 2p\frac{\beta}{\mathcal{N}}[SS][I] \\
\frac{d[SE]}{dt} &= \tau([SSI] - [ISE]) - \mu([SEI] + [ISE]) - \omega[ISE] + p\frac{\beta}{\mathcal{N}}([SS] - [SE])[I] \\
\frac{d[SI]}{dt} &= \nu[SE] - (\tau+\mu+\omega)([ISI] + [SI]) - p\frac{\beta}{\mathcal{N}}[SI][I] \\
\frac{d[EE]}{dt} &= \tau[ISE] - 2\mu[EEI] - 2\nu[EE] + 2p\frac{\beta}{\mathcal{N}}[SE][I]
\end{aligned}$$

$$\begin{aligned}\frac{d[EI]}{dt} &= \nu[EE] - \mu([EI] + [IEI]) - (\nu + \sigma)[EI] + p\frac{\beta}{\mathcal{N}}[SI][I] \\ \frac{d[II]}{dt} &= 2\nu[EI] - 2\sigma[II] - 2\mu([II] + [III])\end{aligned}$$

The number of triples (for instance [EEI]) are approximated with the method provided in Section 3 of Keeling (1999). $\tau = (1-p)\frac{\beta}{n}$ is the transmission rate across a contact, where p is the proportion of contacts that they are long-range. ν is the rate transition rate from the uninfectious to the infectious state while σ is the transition rate from the the infectious to the removal state. μ is the rate at which farms in the neighbourhood of an infected farm are culled and ω is the rate at which farms are vaccinated in ring vaccination. It is also assumed that vaccination has no effect on previously infected farms. The above set of differential equations reveal that new infections occur with the following rate:

$$\text{rate of new infections } = (1-p) \cdot [SI] \cdot \frac{\beta}{n} + p \cdot [S][I] \cdot \frac{\beta}{\mathcal{N}} \tag{3.6}$$

3.2.3.3 A Spatially Explicit Per−Farm Hazard Model

Ferguson et al. (2001b) extended their earlier work (Ferguson et al., 2001a) to take into account the heterogeneity between the farms. Apart from classifying farms in terms of their status, they additionally structured the them by their livestock, classifying each farm as either cattle, sheep, pig or small (fewer than 100 animals). The analysis framework used what they term as a spatially explicit per−farm hazard model in discrete time. It was formulated to allow simultaneous estimation of the spatial transmission, infectiousness, susceptibility and time−varying transmission rates. Infectiousness is dependent on farm type, specified by the species mix and number of animals per farm, susceptibility is dependent on both farm type and farmland fragmentation.

Each of the farms is indexed by (k, l, i) with k and l respectively denoting farm-type-dependent infectiousness and susceptibility classes and i indexing farms within

the (k, l) class. The relative infectiousness of farms in infectious class k is denoted by T_k. Similarly, the relative susceptibility of farms in a susceptible class l is denoted by S_l.

The model in its general form allows the infectiousness of farm i within the class (k, l), $n_{kli}(d)$, to vary with the number of days, d, since infection on the farm. Nevertheless, in the work presented by the authors, they assumed that infectiousness does not vary from the day after infection until the date on which the farm was culled. The probability at which farm i of class (k, l) infects farm j of class $(k^{'}l^{'})$ is given as follows:

$$\beta_{kli\ k^{'}l^{'}j} = \frac{T_k \cdot (I_j - I_i) \cdot n_{k^{'}l^{'}i} \cdot T_{\rho\left(k^{'}l^{'}j,kli\right)}}{\sum_{k^{''}l^{''}i^{''}} T_{k^{''}} \cdot (I_j - I_i^{''}) \cdot n_{k^{''}l^{''}i^{''}} \cdot T_{\rho\left(k^{'}l^{'}j,k^{''}l^{''}i^{''}\right)}}$$

where $n_{kli}(d)$ denotes the relative infectiousness of farm (k, l, i), d days after infection, T_k denotes the relative infectiousness of farms in class k. Let I_i denote the infection time of farm i and $T_{\rho\left(kli,k^{'}l^{'}j\right)}$ the relative infectiousness of an infected farm (k, l, i) to a susceptible farm $(k^{'}, l^{'}, j)$ distance $\rho\left(kli, k^{'}l^{'}j\right)$ away. In addition, the authors have also focused in deriving farm-specific estimates of the "effective reproductive" number R_{kli} which is obtained by summing the proportions of infections attributed to each farm weighted according to their susceptibility. The estimate R_{kli} is then corrected for neighbourhood depletion and gives the farm-specific R_0.

3.2.3.4 Methodology

Ferguson et al. (2001a) reported estimates for the parameters which are associated with the transmission kernel (a, b) and the probability of long-range infection (p) by fitting $f(\rho(i, j))$ to the distribution of distances $\rho(i, j)$ between identified infectious contacts. The pair-based model was fitted to the 3 fully recorded incidence time series (report, confirmation, and slaughter), assuming that the data are Poisson distributed. The authors then derived estimates for the first date of infection, R_0

before and after the introduction of movement restrictions.

For the spatially explicit per−farm hazard model, the parameters such as farm's infectivity and farm's susceptibility, are estimated iteratively by adjusting parameters such that the observed and predicted number of infections in each class (infectious, susceptible) were equal. Furthermore, the effect of farm fragmentation is estimated by assuming that the susceptibility was linearly proportional to fragmentation and fitting the slop parameter such that the average number of fragments in the farms expected to be infected by IPs (over the course of the epidemic) equalled the average number of fragments on that day.

On the basis of these parameter estimates the likelihood of the observed epidemic was calculated based on the farm type, fragmentation and location of each farm in the country. For each day, expected infection probabilities are obtained from the estimated hazards scale to sum to the observed number of infections across the country at that day. The authors used univariate likelihood profiles with respect to each of the parameters of interest were used to calculate confidence bounds on the parameters.

Finally, the authors indicate that a full likelihood approach is much more computationally intensive than the adopted iterative procedures since the former relies on multidimensional optimization methods. Nevertheless, they claimed that their adopted parameter estimation techniques very close to the maximum likelihood estimates.

3.2.3.5 Results

The Imperial group focused on an extensive description of the times (also referred to as "delays") from infection-to-report and report-to-slaughter and especially how these distributions varied over the epidemic. Moreover, they were interested in estimating farm-specific R_0 before and after the placement of the control policies. They suggested that culling of infected premises and dangerous contacts from the

start of the epidemic could have reduced the number of farms lost by 45%. The same model was also used to investigate the use of the ring culling and vaccination. They concluded that both can be used to control the epidemic with vaccination requiring relatively larger rings (Ferguson et al., 2001a,b).

3.2.4 A Partial Likelihood Approach

Diggle (2006) is concerned with drawing formal statistical inference for the FMD 2001 outbreak. The author's approach is to a specify a model for a spatio-temporal point process through its conditional intensity at location x and at time t, given the history of the process up to time t. It is assumed that the data consist of all relevant events in a pre-specified spatial region A and time interval $[0, T_{obs}]$. Then, a parametric model for the underlying point process is specified and the goal is to make inference for the model parameters.

3.2.4.1 The Model

Diggle (2006) models the conditional transmission rate from an infected farm i to a susceptible farm in a form similar to the CE model. Let n_i^c and n_i^s denote the numbers of cows and sheep held on farm i. Denote by $\mathbb{I}_{ij}(t)$ an at-risk indicator for transmission of infection from farm i to farm j at time t, if farm i is infected and not slaughtered by time t, and farm j is not infected and not slaughtered by time t. A central feature of the model is the transmission kernel which has the following form:

$$K(\rho(i,j)) = \exp\left\{-(\rho(i,j)/\phi)^{\kappa}\right\} + \rho$$

where $\exp\left\{-(\rho(i,j)/\phi)^{\kappa}\right\}$ in which the powered exponential term corresponds to spread of the infection over short distances, whilst the parameter ρ allows for long-distance infections, occurring far from all currently infectious farms. Let $\beta_{ij}(t)$ denote the conditional rate of transmission from farm i to farm j, given the history $\mathcal{H}_t$. The model has the following form:

$$\beta_{ij}(t) = \beta_0(t) \cdot T_i \cdot S_j \cdot K(\rho(i,j)) \cdot \mathbb{I}_{ij}(t) \tag{3.7}$$

where $\beta_0(t)$ is an arbitrary baseline hazard and

$$T_i = \epsilon \cdot n_i^c + n_i^s \quad \text{and} \quad S_j = \xi \cdot n_j^c + n_j^s.$$

The parameters ϵ and ξ represent the relative infectiousness and susceptibility, respectively, of cows to sheep. The model is very similar to the model proposed by Keeling et al. (2001) except from the kernel, $K(\rho(i,j))$. The forms of farm's infectivity and susceptibility are identical.

3.2.4.2 The Methodology

We shall briefly describe the methodology adopted by the author to draw inference for the parameters of interest. Denote by $\mathcal{H}_t$ the complete history of the process up to time t and let by $\beta(x,t|\mathcal{H}_t)$ the conditional intensity for an event at location x and time t, given $\mathcal{H}_t$. Therefore, for data which consist of the location of all the events in the specified area and the time interval, $(x_i,t_i) \in A \times [0,T_o] : i = 1,\ldots,n$, with $t_1 < t_2 < \ldots < t_n$, the log-likelihood function can be derived as follows (see for example, Daley and Vere-Jones, 1988):

$$L(\theta) = \sum_{i=1}^{n} \log \beta(x_i,t_i|\mathcal{H}_t) - \int_0^{T_o} \int_A \beta(x,t|\mathcal{H}_t)\ \mathrm{d}x\ \mathrm{d}t$$

where θ is the parameter associated with the intensity function. It is often the case where the form of the conditional intensities may be intractable or/and the integral term might be impractical. Diggle (2006) comments that although Monte Carlo methods are widely available for such problems (see for example, Møller and Waagepetersen, 2004), these methods often need careful tuning to each application, and the cost of developing them turns out to be an obstacle to their routine use.

The author proposes as an alternative, a computationally simpler approach to

make inference for models which are defined through their conditional intensity. A partial likelihood is proposed which can be obtained by conditioning on the locations x_i and times t_i and taking into account the resulting log−likelihood for the observed time-ordering of the events, $1, \ldots, n$. We need to adjust for right-censored event-times, we denote by $\mathcal{R}_i$ the risk-set at time t_i. Then we let:

$$p_i = \frac{\beta(x_i, t_i|\mathcal{H}_{t_i})}{\sum_{j\in\mathcal{R}_i} \beta(x_j, t_i|\mathcal{H}_{t_i})} \tag{3.8}$$

Then, the partial log-likelihood is

$$L_p(\theta) = \sum_{i=1}^{n} \log p_i. \tag{3.9}$$

The partial likelihood defined by the above equations is a direct adaption to the space-time of the seminal proposal in Cox (1972) for proportional hazards modelling of survival data. As discussed in Cox (1975), estimates obtained by maximising the partial likelihood preserve the general asymptotic properties of maximum likelihood estimators (MLE) although they might be less efficient that then full MLEs. Moreover, Diggle (2006) indicates that although some parameters of the original model may be unidentifiable from the partial likelihood, this is not a problem if non-identified parameters are nuisance parameters.

When the conditional intensity function can be expressed as

$$\beta(x, t|\mathcal{H}_t) = \beta_0(t)g(x, t|\mathcal{H}_t)$$

for some function $\beta_0(t)$, it follows that the partial likelihood provides no information about $\beta_0(t)$. If $g(\cdot)$ is indexed by parameters θ, then the partial log-likelihood is

$$L_p(\theta) = \sum_i \log\left(g(x_i, t_i|\mathcal{H}_{t_i})\right) - \sum_i \log\left\{\sum_{j>i} g(x_j, t_i|\mathcal{H}_{t_i})\right\}$$

Within the FMD model, for any farm i the relevant conditional intensity is $\beta(x_i, t_i|\mathcal{H}_{t_i}) =$

$\sum_j \beta_{ji}(t_i)$ and the partial likelihood follows by substitution of these conditional intensities into Equations (3.8) and 3.9). The partial likelihood is maximised by using the Nelder-Mead simplex algorithm (Nelder and Mead, 1965).

3.2.4.3 Results

Diggle (2006) provides estimates for the parameters of interest ϵ, ξ, ϕ and ρ. Qualitatively similar conclusions were obtained to those derived by Keeling et al. (2001) in the sense that cattle are more infective and more susceptible than sheep. The author also focused on possible extensions of the model by considering the fact that the infectivity and susceptibility for each individual farm may be sub-linear to number of animals.

3.2.5 An Individual-Level-Model's Approach

Very recently, Deardon et al. (2006) have focused on drawing Bayesian inference for individual-level models. The probability of a susceptible individual being infected from the infectious pressure is modelled by considering potential factors of infection (eg. distance). Models of this structure have already been proposed in the literature (see for example Gibson, 1997, Keeling et al., 2001). Such an approach provides models which are very intuitive and flexible. In addition, they fit very naturally to a Bayesian framework where the imputation of the missing data is straightforward via MCMC methodology.

Nevertheless, such models are typically highly computationally costly to analyse, especially when dealing with large data sets. That is due to the difficulty of calculating the full likelihood. Deardon et al. (2006) are mainly concerned with providing methods to minimise this computational cost. At this stage, we will first describe their model and then briefly their methodology.

3.2.5.1 The Model

The authors formulate an $SEIR$ model in discrete time, which assumes that at any given time t, an individual farm i can be in one of four states: $i \in S$ implies farm i is susceptible to FMD; $i \in E$ implies farm i has been exposed to the disease (i.e. has been infected), but is not yet infectious; $i \in I$ implies farm i is infectious; $i \in R$ implies that farm i has been removed from the population, in this context through the culling of animals. The time is measured in days and the following assumptions are made: once exposed to FMD, the farm would remain in that exposed state for 5 days, and then become infectious for 4 days. After this point, it is assumed that symptoms would be visible and so the farm would be reported and subsequently removed from the population through animal slaughter.

The probability of a susceptible farm j becoming infected during the time interval $[t, t+1)$ is given as follows:

$$P(j,t) = 1 - \exp\left\{-\sum_{i \in I_t} \beta_{ij}\right\} \tag{3.10}$$

where I_t denotes the set of infectious farm at time t and β_{ij} is defined as follows:

$$\beta_{ij} = S_j \times (K(\rho(i,j)) \times T_i + \epsilon \cdot |E(t+1) \setminus E(t)|) \tag{3.11}$$

where T_i and S_i denotes the infectivity and susceptibility for farm i respectively. They are defined as follows:

$$\begin{aligned} T_i &= \epsilon_c (n_i^c)^{\zeta_1} + \epsilon_s (n_i^s)^{\zeta_2} \\ S_i &= \xi_c (n_i^c)^{\zeta_3} + \xi_s (n_i^s)^{\zeta_4} \end{aligned}$$

The vector (ξ_s, ξ_c) denotes the susceptibility vector where ξ_s and ξ_c describe the rate of increase in susceptibility of a susceptible farm per additional sheep and cow respectively. Similarly, the transmissibility vector, (ϵ_s, ϵ_c), where ϵ_s and ϵ_c

are parameters describing the rate of increase in infectious pressure per additional sheep and cow, respectively, that an infectious farm exerts on the susceptible population. Following similar notation to the previous approaches, n_i^c and n_i^s denote the number of cattle and sheep in farm i respectively. The set of parameters $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \zeta_3, \zeta_4)$ allows us to assume that the effect of the number of animals of different species to the infection probability could be non-linear.

$K(\rho(i,j))$ denotes a distance−based kernel where $\rho(i,j)$ denotes the distance between farms i and j. The authors prefer to use a geometric kernel rather than an exponential, which could have also been used. They argue that the former has more mass in the tails and allows for the possibility of longer infections. They also mention that after experimentation with the FMD data, the geometric kernel failed to adequately describe short range transmissions as the geometric shape was largely determined by its tail which describes the longer-range transmissions. On the other hand, they believe that over shorter distances, different disease transmission dynamics come in to play. Therefore, they suggest a threshold distance δ_0, to be determined, within which we assume a constant disease transmission rate, k_0. Thus, they introduced the following kernel

$$K(i,j) = \begin{cases} k_0, & 0 \leq \rho(i,j) \leq \delta_0 \\ \rho(i,j)^b, & \delta_0 < \rho(i,j) \leq \delta_{\max} \\ 0, & \text{otherwise} \end{cases}$$

with parameters k_0, δ_0, b. The parameter $\delta_{\max}$ is fixed *a priori* and set to 30km. It is mentioned that changing $\delta_{\max}$ has an effect on the results; however, increasing $\delta_{\max}$ above 30km produces relatively little change in the other results.

The model also allows for spontaneous infection which is unexplained by the susceptibility, transmissibility and kernel components of the model. Such infections are considered as *sparks* infections and in such an example it allows for infection beyond the $\delta_{\max}$ limit (e.g. long distance movement of vehicles or people for example). The authors suppose that the risk of a spark infection on farm i at time t

is affected by the farm's susceptibility and by the number of newly exposed farms during the interval $[t, t+1)$, denoted by $|E(t+1) \setminus E(t)|$.

It is easy to see that the model as shown in (3.11) is an extended version of the models shown in (3.2) and (3.7). The CE model does not explicitly model the distance kernel and only assumes a linear effect of the number of animals to the infection rate. Diggle's model considers a parametric form of the distance kernel and also allows for a non-linear effect of the covariates, which is common for sheep and cattle. On the other hand, the model presented in Deardon et al. (2006) has much more flexibility regarding the distance kernel and the effect of the covariates.

3.2.5.2 The Methodology

We briefly review the methodology adopted by the authors. Inference for the parameters is drawn within a Bayesian framework. The likelihood is given below:

$$f(S, E, I, R|\text{parameters}) = \prod_t \left\{ \left(\prod_{i \in E(t+1) \setminus E(t)} P(i,t) \right) \left(\prod_{i \in S(t+1)} (1 - P(i,t)) \right) \right\} \tag{3.12}$$

It has already been mentioned that such models are highly costly computational due to the products which appear in (3.12). For instance, for any t, the $S(t)$ consists of around $150,000$ elements and the calculation of the (3.12) at every MCMC iteration becomes demanding.

Deardon et al. (2006) provide methodology to avoid recalculating (3.12) at each iteration. This is achieved by linearising the model through a Taylor series expansion. Then, the time-consuming summations in the likelihood can be partitioned into two groups: those which are computationally expensive but need to be calculated only once; and those which are much quicker to compute but change with the parameter values and therefore need to be recalculated throughout the simulation. Finally, this approach also assumes that the infection status of those farms culled without knowing whether they have been infected or not is unknown and hence,

the time of infection of these farms must be imputed. Note that for the infected premises, their infection times are assumed to be known.

3.2.5.3 Results

Deardon et al. (2006) provided estimates regarding farm's infectivity and susceptibility as well as the transmission kernels. Their findings are similar to the results obtained by the other approaches. In addition, the authors argued that the tracing data kernel (provided by DEFRA) overestimates the risk of short-distance infection and underestimate that of long-distance infections. Furthermore, they also indicated that the linear assumption made in Keeling et al. (2001) was questionable in terms of transmissibility, whereas a linear approximation looks more reasonable in the case of susceptibility. Finally, regarding the assessment of the adopted control policies by DEFRA, Deardon et al. (2006) suggest that perhaps CP culling was not carried out in the most efficient manner.

3.2.6 Preliminary Conclusions

In this section we described the models and the corresponding methodology adopted by five different approaches (Morris et al., 2001, Keeling et al., 2001, Ferguson et al., 2001a,b, Diggle, 2006, Deardon et al., 2006) to infer about the dynamics of the 2001 UK FMD outbreak. Diggle (2006) and Deardon et al. (2006) perform statistical inference based solely on the likelihood. On the other hand, Ferguson et al. (2001a,b) and Keeling et al. (2001) use methods which make use of the likelihood of the observed data given the parameters in conjunction with stochastic simulation. That is, given a set of parameter values the model is simulated forward to predict the most likely spread of the epidemic.

Concluding, InterSpread is a very flexible and powerful modelling tool. It is not used to draw inference for the parameters but to simulate a variety of models, from very simple to very complex ones given some specific assumptions. The large

number of transmission roots allow us to construct models which reflect the real epidemic process. On the other hand, the more parameters we insert, the more informative and detailed the available data should be. In addition, Keeling (2005) reports that InterSpread is much more computationally costly to simulate than the other two models which were used during the outbreak (Imperial and Cambridge - Edinburgh). Extra details used in InterSpread should be weighted against difficulties in parameterisation with expert opinion being required to estimate many quantities of interest.

The stochastic simulations of the CE are much computationally cheaper than InterSpread due to simpler transmission mechanism of the former. However, robustness of the obtained results should be investigated since changes in the parameter estimates could have potentially led to very different conclusions regarding control policies. In addition, parameter's uncertainty is performed via sensitivity analysis, nevertheless it questionable whether the parameters of the model satisfy orthogonality.

Although the initial model by (Ferguson et al., 2001a) was able to capture the temporal dynamics of the 2001 FMD, the model suffers from being deterministic. Therefore, questions which have strong stochastic element such as the duration of the epidemic cannot be accurately answered. Furthermore, since the model always predicts an average epidemic, extreme situations such as when a significant number of infections occurs, or other unlikely scenarios are not really encountered. Apart from not being stochastic, the other main feature of the Imperial model is that does not explicitly model the spatial spread of the disease. The latter is only taken into account, by externally defining the number of the local connections of each farms. The way these farm-specific local connections are defined can potentially have an effect on the parameters and lead to different estimates.

The model proposed by Diggle (2006) is very similar to the CE model, nevertheless the difference in the two approaches relies on the methodology used to infer the parameters. Although the implementation of maximum partial likelihood method

is relatively straightforward, the obtained estimators might be less efficient than their likelihood counterparts. Such methods are well-suited to routine use, provided that the parameters of interest remain identifiable.

Deardon et al. (2006) are the first to adopt a fully Bayesian approach to model the 2001 FMD outbreak. In order to capture the dynamics of the disease, they propose a very detailed and flexible model in discrete time. They also incorporate within the Bayesian framework the unknown status of the culled farms which have been identified by DEFRA as dangerous contacts. Nevertheless, they treat the farm's latent and infectious periods fixed. Concluding, the authors focus on deriving methods to reduce the computational cost of such an approach and make it feasible for large data sets, such as the 2001 UK epidemic.

Despite the big differences of the methodology of the approaches by Morris et al. (2001), Keeling et al. (2001), Ferguson et al. (2001a,b), similar predictions were made regarding the type of controls which were needed to prevent the epidemic from spreading. Nevertheless, prediction of epidemic risks on the basis of plug-in parameter estimates is likely to be highly inaccurate no matter how well-informed these estimates are. Moreover, no level of sensitivity analysis can compensate for a rigorous statistical analysis. A sensitivity analysis will mainly be based on considering orthogonality between the parameters and in stochastic epidemic models such an assumption is often implausible.

3.3 A Fully Stochastic Epidemic Model

Our goal is to adopt a fully likelihood-based approach because this will allow all relevant information from the data to be extracted. In addition, it is important to account for stochasticity in the evolution of the epidemic in time and secondly, the risk analysis needs to take into account risk due to *parameter* uncertainty.

In this section, we propose to use a heterogeneously mixing stochastic epidemic

model (see Section 2.2) to capture the dynamics of the FMD disease in the UK. Adopting a fully Bayesian approach, a natural framework is offered to incorporate the unobserved infectious periods whereas all the other approaches discussed so far assumed fixed (and known) infectious periods. In addition an alternative transmission kernel is used in order to capture long-range infections without the need of extra parameters.

Our main goal is to draw inference for the parameters associated with farm's infectivity and susceptibility using efficient MCMC algorithms. In general we are interested in drawing conclusions on the mechanism that could reveal how the disease was spread. Furthermore, we also compare the obtained results with those derived from the other approaches and comment on any differences or similarities between them.

3.3.1 The Data

Two years after the outbreak in 2003, the Department of Environment Food and Rural Affairs (DEFRA) uploaded a spread sheet in their website (`www.defra.gov.uk`). The name of the file was "DataForModellersOct03.xls" and contained information about the infected premises. The key entries of this file are shown in Table 3.1.

The same file also provides the "infection dates "of the infected premises (IP). This is either estimated as 5 days prior to date of earliest lesions, regardless of species. In this case the "Infection Status" is set to "E". On other the hand, "infection date" is considered the date of a known contact with infected farm. This is the case where "Infection Status" is set to "C". For cases confirmed on serology (infection status set to 'S') the infection date is estimated as 10 days. Within our framework, (see 2.2) we are able to draw formal inference about the infection times of the IPs and therefore we discard this information from DEFRA.

Regarding the *Date Slaughtered*, it is explained by DEFRA that there can be multiple slaughter dates on DCS due to voluntary culls and previous DC status

etc. Therefore the slaughter date is closest to the date it became infected. The variables *Dairy* and *Beef*, which refer to the total number of dairy animals and cattle respectively, are firstly imported from the DCS database and then they were corrected from telephone reports. DEFRA has also provided us with another file which includes information about the uninfected premises, which contains the x and y coordinates, and the number of different species for each farm.

Table 3.1: Information on the infected premises

Variable Name	**Description**
IP	Sequence number for every Infected Premises.
X	x-coordinate of Map reference.
Y	y-coordinate of Map reference.
Infection Status	'E' if the 'Infection Date' has been estimated, 'C' if there is a known date of contact, or 'S' if FMD was confirmed on the basis of serology.
Date of Report	Date of report to MAFF SVS HQ at Page Street. Imported from DCS database.
Date Slaughtered	Date slaughter was completed. Updated from DCS regularly, but will not always be current.
Date confirmed	The date the premises was confirmed to be infected. Imported from DCS database.
Dairy	Number of dairy animals on premises.
Beef	Refers to total cattle on property.
Sheep	Total sheep on property. Imported from DCS database.
Pigs	Total pigs on property. Imported from DCS database.
Goat	Total goats on property. Imported from DCS database.
Deer	Total deer on property. Imported from DCS database.
Infecting species	First species to be infected on the IP.

Figure 3.1 shows the spatial distribution of the infected and uninfected farms in the UK. It is clear that the counties which were mostly affected by the FMD are Devon (south-east) and Cumbria (north-west). Diggle (2006) argues that because the two counties are geographically well separated, they should be treated informally as replicates of the same natural experiment, thus allowing to compare parameter estimates and pool as appropriate. Following this argument, we decide to analyze

only the data in Cumbria since it had more infected premises than Devon (see Figure 3.2).

Due to the limited availability of detail in the spread sheet, as Keeling et al. (2001), Diggle (2006) and Deardon et al. (2006) we choose as model covariates the number of cattle and number of sheep for each farm. Note that a small number of farms which appears to have zero number of cattle (n_c) and sheep (n_s), have been excluded from our study. The variable *Date Slaughtered* is considered as the *removal times* within the context of an HMSE. The corresponding *infection times* are assumed to be unknown. Table 3.2 presents the summary statistics for the two covariates. Figure 3.3 shows the distribution of the size of susceptible farms in Cumbria and reveals skewness and heavy tails. In other words there are a few farms with significantly larger size than the average.

Once the data are cleaned, the resulting dataset consists of $\mathcal{N} = 5378$ farms in total. At the end of the outbreak, $n_I = 1021$ got infected. Some farms were culled without knowing their infection status (eg. dangerous contacts). If information about these farms was available then the status before their culling (infected, susceptible) could have been imputed. Alternatively, a simpler approach would have been to remove them from the susceptible population once the animals have been slaughtered. Due to the absence of such data, the fact that farm animals may be slaughtered after they become infected but before the disease is diagnosed is ignored with our approach.

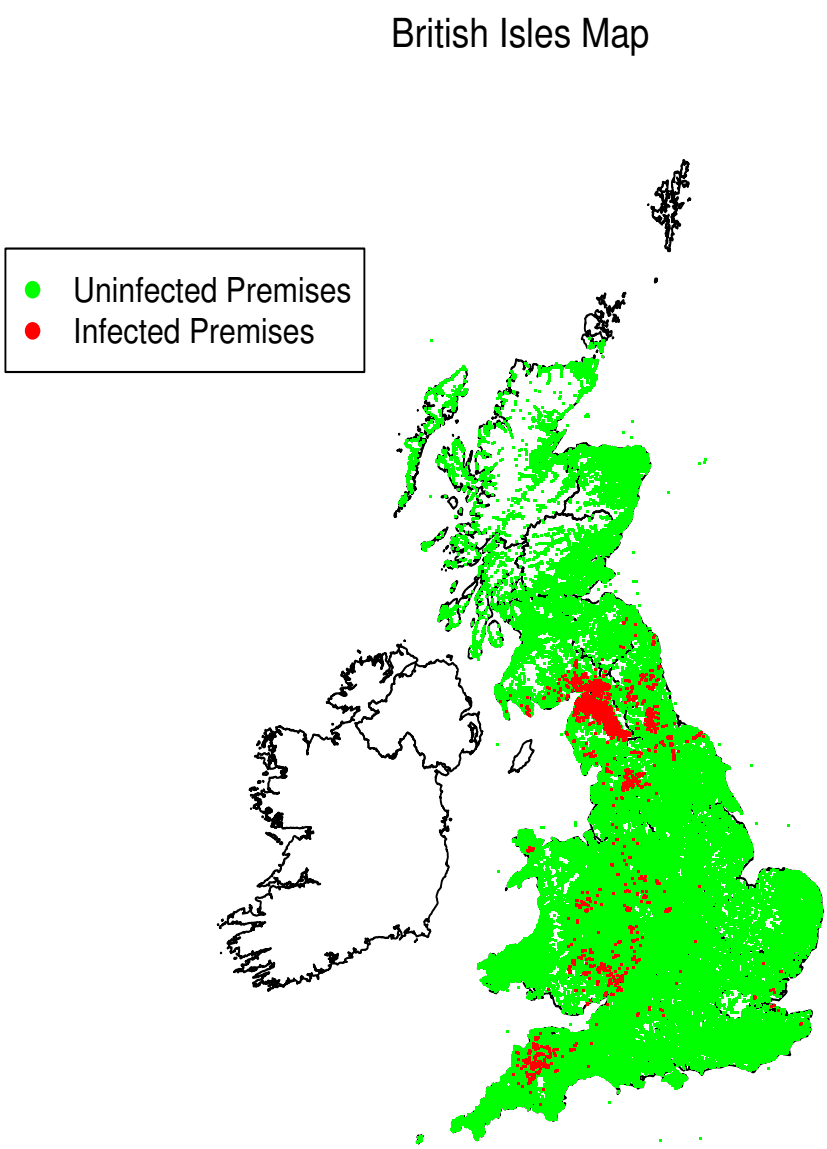

Figure 3.1: The spatial distribution of susceptible farms in the UK at the start of the outbreak (green) and of the infected farms at the end of the outbreak (red).

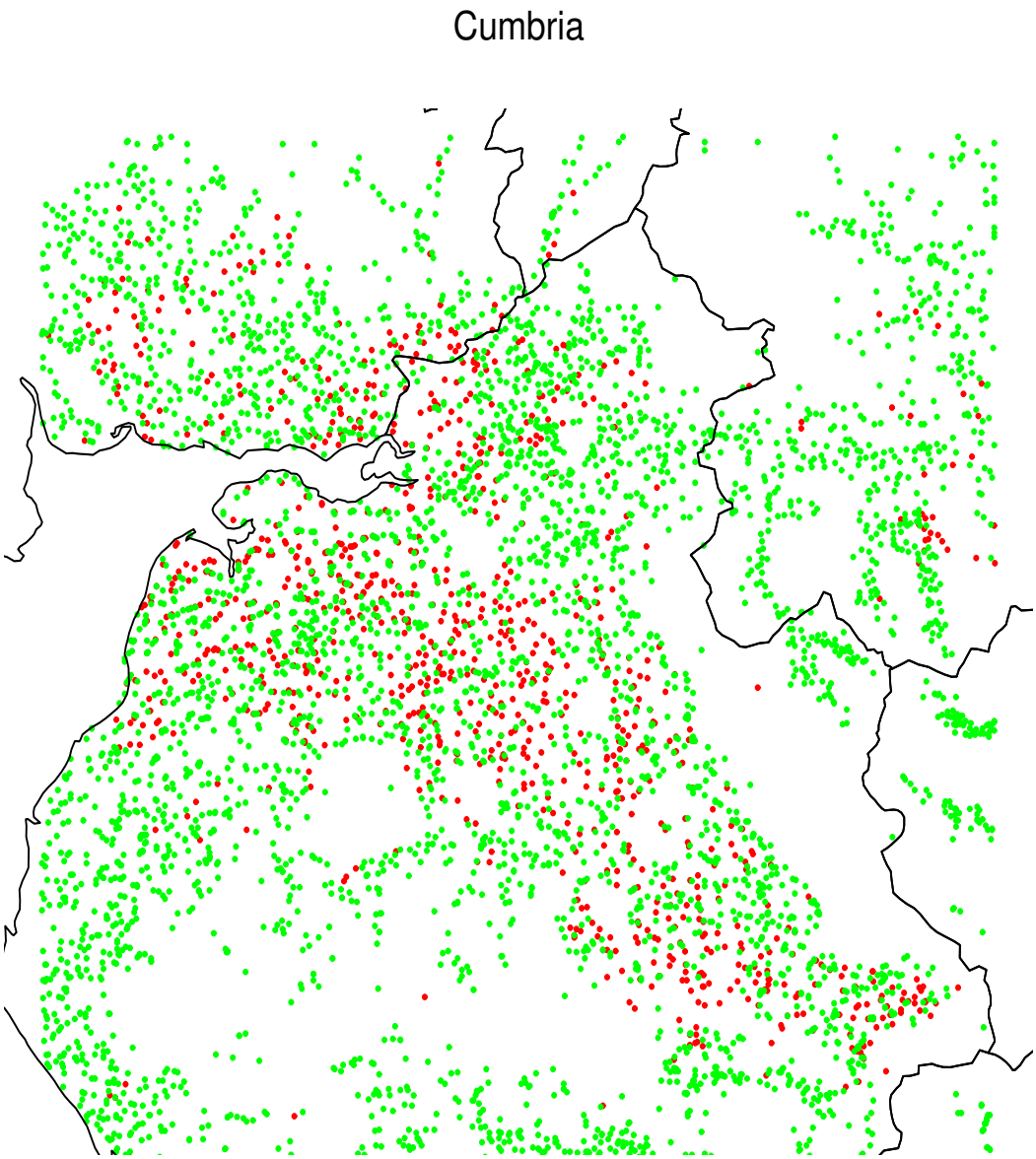

Figure 3.2: The spatial distribution of susceptible farms in Cumbria at the start of the outbreak (green) and of the infected farms at the end of the outbreak (red).

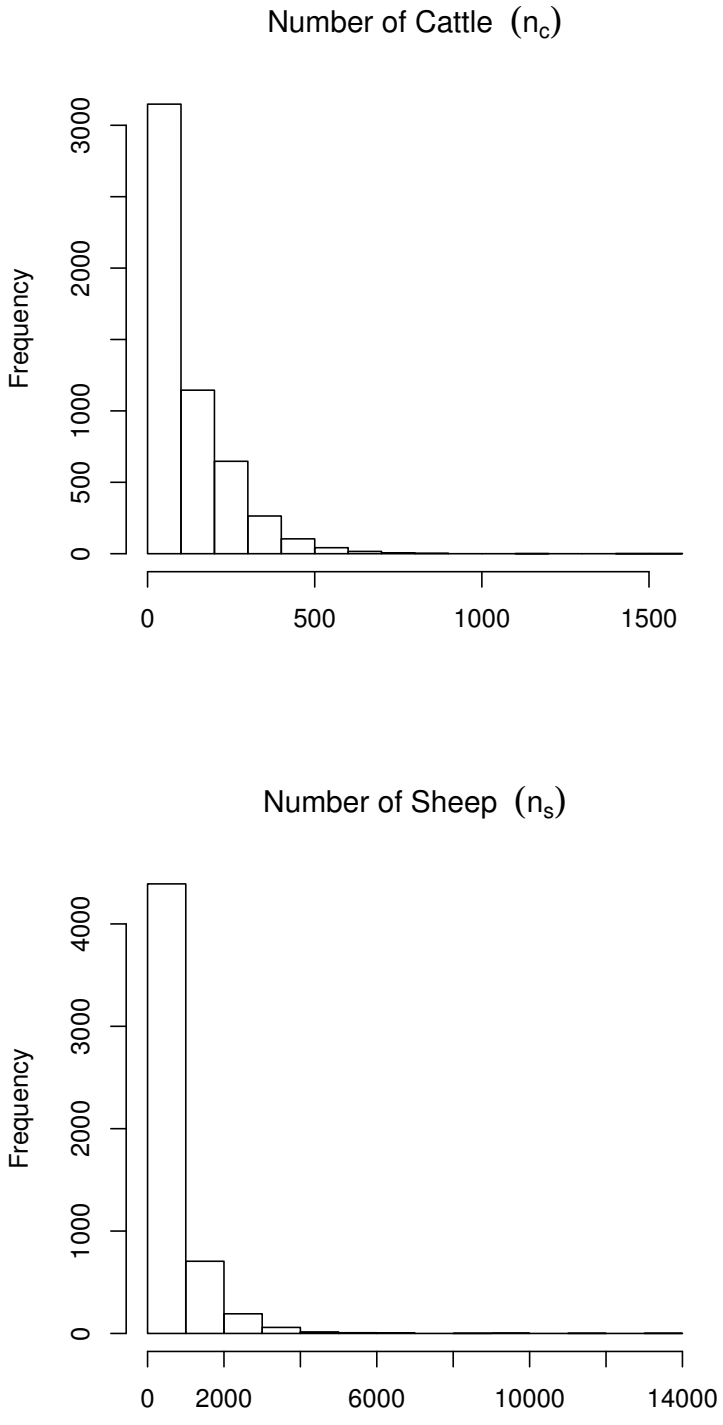

Figure 3.3: Histograms of the number of cattle and sheep for the all the susceptible farms in Cumbria

Table 3.2: Summary statistics for the number of cattle and sheep of each farm in Cumbria

	Min.	1st Quart.	Median	Mean	Std. Error	3rd Quart.	Max.
Cattle	0	9	73	111.7	127.43	171	1593
Sheep	0	10	227	534.5	812.583	739.0	13240.0

3.3.2 The Model

Our approach relies on constructing an HMSE (see Chapter 2, Section 2.2) with appropriate assumptions about the infection and the removal rate which we describe in this section in detail.

Infectious Period

At any time t, the farm can be in one of the three states: susceptible, infected or removed. Within the FMD context, a previously infected farm is considered to be removed when all its animals have been slaughtered. Once a farm is infected, is remains infectious for some time which we assume is Gamma distributed with mean α/γ and variance α/γ^2:

$$R_i - I_i = Ga(\alpha, \gamma).$$

We assume that the shape parameter of the Gamma distribution, α, is known and equal to 4. The particularly chosen value of α, leads to a bell-shaped distribution where the mean (or the mode) and the variance depends only on γ. Figure 3.4 reveals the various shapes of such distribution for different values of γ. We should note that by assuming an Exponential infectious period i.e. $\alpha = 1$, such flexibility is not possible.

The commonly adopted approach is to assume that the infectious period of a farm is fixed and known. For instance, and Keeling et al. (2001) and Deardon et al.

(2006) assume that a farm is in the exposed state for five days and then remains infectious for four days. Diggle (2006) considers that the reporting date is the infection date plus a constant time τ, corresponding to the latent period of the disease plus any reporting delay. Although such assumptions about the infectious period and hence the (unknown) infection times, can reduce the computational cost considerably, it is of interest to quantify the effect such assumptions have on parameter's inference.

Suppose that within the context of an SIR model the individual's infectious period is distributed as the random variable D. Lefèvre and Picard (1993) showed that if we replace D with its mean μ, then we tend to predict an epidemic with a smaller number of ultimate susceptibles surviving the disease, i.e. an epidemic with a larger final size. For illustration, suppose two different infectious periods, $D_1 \sim Ga(\alpha, \gamma)$ and $D_2 \sim Ga(\kappa\alpha, \kappa\gamma)$. Then, we can show that that once you increase the variance of the infectious period the final size gets smaller. Nevertheless, exact results in terms of the infection and removal rates (instead of the final size) do not seem to exist. Therefore, such assumptions about the length of the infectious periods should be made with extra caution. In contrast, within our Bayesian framework, infection times are treated as unknown parameters and hence inference is made based on the observed data (removal times).

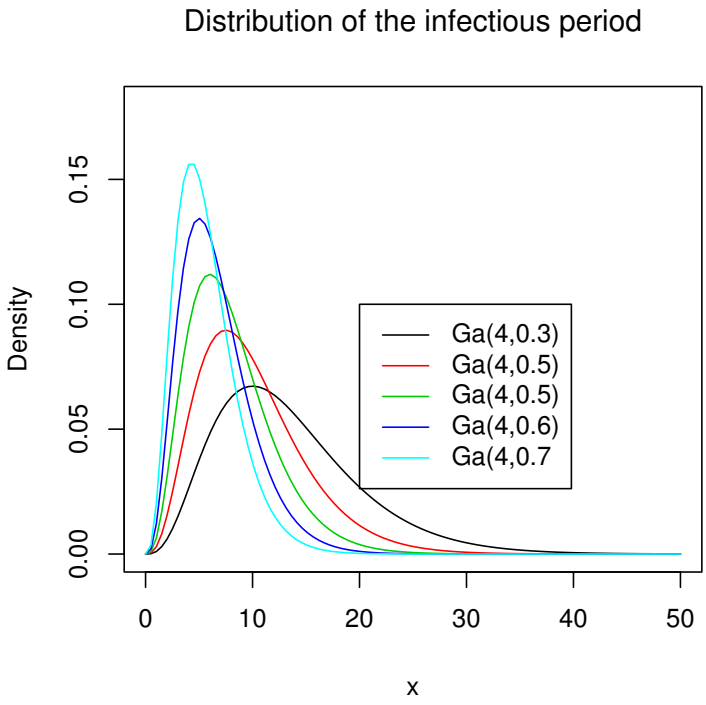

Figure 3.4: Different distributions for the infectious period given specific values of the shape and the scale parameter

Infection Rate

We model the infection rate by taking into account the available covariates which are the location and the size of the farm. Following the approaches by other groups (see for example, Keeling et al., 2001, Diggle, 2006, Deardon et al., 2006) we take into account the number of different species in each farm and in particular the number of sheep and cattle and in each farm. Since FMD was primarily confined to cows and sheep, this approach seems sensible. The infection rate is modelled as follows:

$$\begin{aligned} \beta_{ij} &= \beta_0 \cdot K(i,j) \cdot \left(\epsilon \cdot (n_i^c)^\zeta + (n_i^s)^\zeta\right) \cdot \left(\xi \cdot (n_j^c)^\zeta + (n_j^s)^\zeta\right) \qquad (3.13) \\ K(\rho_{ij}, \delta) &= \frac{\delta}{\rho_{ij}^2 + \delta^2} \end{aligned}$$

where β_0 represents the overall (average) infectious rate, ϵ and ξ the relative in-

fectiousness and susceptibility, respectively, of cows to sheep. n_i^c and n_i^s refer to the number of cattle and sheep for farm i respectively. The parameter ζ identifies whether the infectivity or susceptibility for each of the farms is linear or sub-linear in the numbers of animals. Note that Diggle (2006) has also adjusted for a non-linear effect of the covariates. Moreover, Deardon et al. (2006) have proposed a more general model which allows for different degree of non-linearity among the number of different species and the status of the premise.

The Kernel

We model explicitly the environmental spread by assuming a Cauchy kernel $K(\rho_{i,j}, \delta)$, which is associated with parameter δ. Denoted by $\rho(i, j)$, is the Euclidean distance between farms i and j. We shall explain in this section why such a kernel is chosen.

We should note that a transmission kernel, known as *DEFRA tracing data distance kernel* was used by DEFRA during the outbreak. It was estimated by veterinarians by taking into account information on infected premises (eg. location, dangerous contacts) and making a best guess at the source of the infection. It could be argued that such a subjectively-based distance kernel is likely to overestimate the effect of short distance infections and on the other underestimate the effect of the long distance infections (see also, Deardon et al., 2006).

The approach adopted by Imperial does not explicitly model the transmission kernel between farms, while the Cambridge-Edinburgh model assumes a kernel with a fixed and known parameter which is more sharply peaked than exponential. Diggle (2006) uses an exponential-type kernel (see Equation 3.7) which allows the relative importance of long-range transmission in the spread of the disease to be measured. Similarly to Deardon et al. (2006), we would like to consider long-range infections (sparks). Therefore we use a Cauchy kernel which has heavier tails than the exponential or the geometric and allows for long-range infections without the need of any extra parameter.

The Distance Between Farms

An important issue regarding the association of the infection rate with the distance between the farms is whether or not the Euclidean metric is the most appropriate measure. Other distances could be used, such as the minimum walking distance. Nevertheless, the answer to the above question is problem-specific and cannot be generalised very easily. Depending on the landscape of the area where the outbreak is taking place and the disease's characteristic, it can be argued that in many cases the Euclidean metric is not applicable.

For example, consider the case that between farms, i and j, there exists a lake or a mountain and that the disease cannot be spread by wind. Obviously, the Euclidean distance does not seem to be an appropriate measure. On the other hand, we should be very careful when using measures like minimum walking distance especially when i and j are adjacent farms.

Regarding the FMD outbreak in the UK, Savill et al. (2006) showed that Euclidean distance between infectious and susceptible premises is a better predictor of transmission risk than shortest and quickest routes via road, except where major geographical features intervene. Therefore, they concluded that a simple spatial transmission kernel based on Euclidean distance suffices in most regions, probably reflecting the multiplicity of transmission routes during the epidemic.

Concluding, due to the lack of geographical information on the landscape of Cumbria and the difficulty of obtaining metrics such as minimum walking distances and also taking into account the results Savill et al. (2006), we used the Euclidean distance for our spatial kernel.

3.3.3 Results

In this section we present the results obtained from our Bayesian analysis. Since any fully Bayesian analysis consists of prior's specification about the parameters,

we first refer to our chosen prior distributions. Then by adopting the methodology in Chapter 2, a 25% partially non−centered algorithm was applied to obtain samples from the posterior distributions of the parameters of interest.

Priors

Consider the following vector of the unknown parameters $\boldsymbol{\theta} = (\beta_0, \gamma, \delta, \epsilon, \xi, \zeta)$. We specify the the following prior:

- $\pi(\beta_0) \equiv Ga(0.001, 0.001)$
- $\pi(\gamma) \equiv Ga(0.001, 0.001)$
- $\pi(\delta) \equiv Ga(1, 0.1)$
- $\pi(\epsilon) \equiv Ga(1, 0.001)$
- $\pi(\xi) \equiv Ga(1, 0.001)$
- $\pi(\zeta) \equiv Ga(1, 0.001)$

It is easy to see that all the assigned priors are fairly uninformative about the parameters. Because the infection times also have to be treated as parameters, we assume a uniform prior over the label of the initially infected farm and also on its corresponding (initial) infection time. We can visualise the state of our knowledge about the parameters of interest by plotting the density of each of the corresponding posterior distributions.

Key Parameter of the Infectious Period

The key parameter which characterizes the infectious period is the scale parameter of the Gamma distribution, $R_i - I_i \sim Ga(4, \gamma)$. Figure 3.5 shows the kernel density estimates of the posterior distribution of γ and the average infectious period, $4.0/\gamma$.

It turns out that the estimated average infectious period is approximately 7.5 days with the 95% highest posterior density region (see Section 1.3.3) of $(7, 8.5)$. Such an estimated value is not very different from the one that Keeling et al. (2001) and Deardon et al. (2006) have assumed. Recall that within the context of both approaches, the authors consider an SEIR-type model, which incorporates apart from the infectious period(I→R), an incubation period (E→I) as well. The incubation period was taken to be 5 days and the period from infection to reporting 9 days. There an estimated average infectious period of 7.5 is quite similar. In addition, we should note that the prior assigned to parameter γ has no influence on the posterior distribution. In other words, the information about γ is extracted mainly from the data.

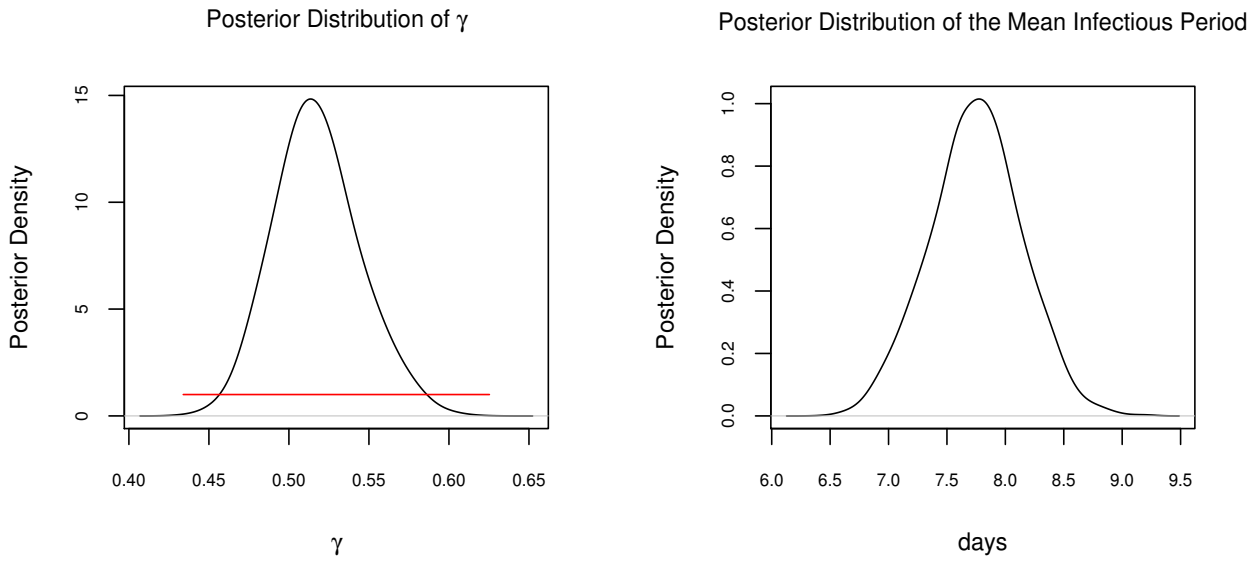

Figure 3.5: Posterior distribution of parameter γ (left) and the corresponding mean infectious period (right)

Spatial Kernel

The key parameter which drives the shape of the spatial kernel in our model is δ. The most likely value of the posterior distribution of δ is around 0.0065 . Figure 3.6 shows a 95% highest posterior density region $(0.0055, 0.0087)$ for the shape of the spatial kernel, $K(\rho(i, j))$, based on the posterior samples of $\pi(\delta|\boldsymbol{R})$. It is

remarkable that there is not much uncertainty about δ. Furthermore, it can be easily seen that the effect of $K(\rho(i,j))$ to the infection rate is very little when the distance between two farms is greater than 4 Km.

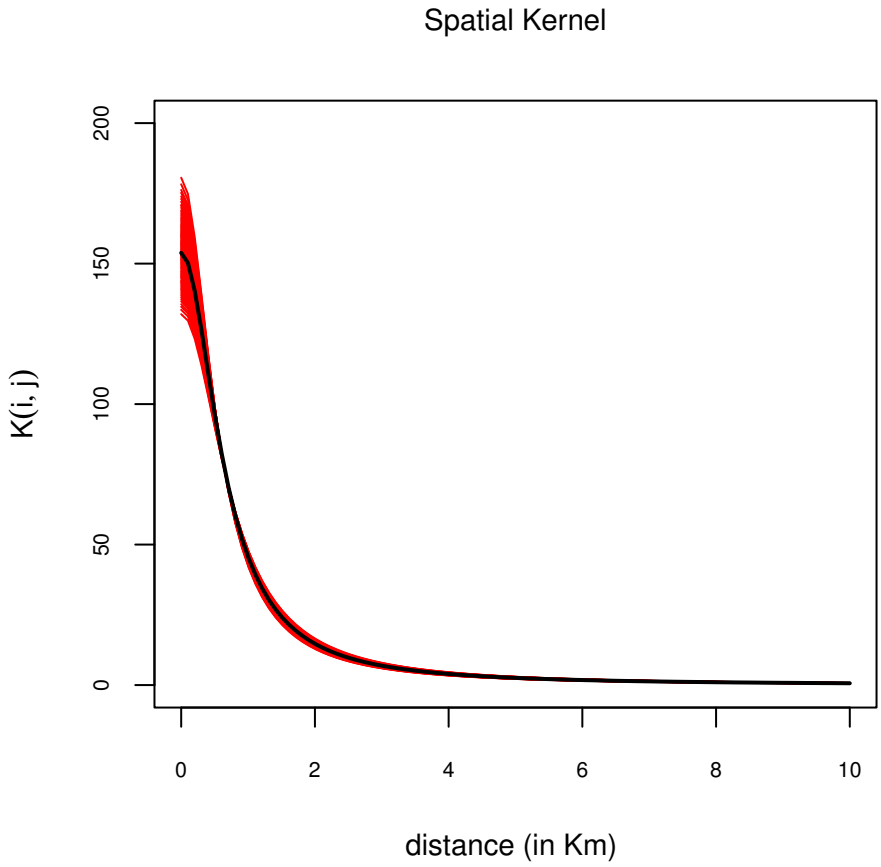

Figure 3.6: A 95% highest posterior density region of $K(i,j)$. The black line refers to the "average" shape of the Kernel, based on the posterior mean of $\pi(\delta|\boldsymbol{R})$

It is of interest to compare the the effect of the kernels used by other approaches to the infection rate (see Table 3.3).

Before performing any proper comparison, we define the *relative kernel's effect* (r_δ) as follows:

$$r_\delta = \frac{K(\rho(i,j),\delta)}{K(0,\delta)} \tag{3.14}$$

In other words, r_δ is obtained as the ratio of the transmission kernel evaluated for two farms which are far from each other at distance ρ ($K(\rho(i,j))$), divided by the

Table 3.3: The form of the different kernels used for the 2001 UK FMD outbreak

A Fully Stochastic Model	**Kernel**
Our approach:	$K(\rho(i,j)) = \frac{\delta}{\delta^2+\rho(i,j)^2}$
Pseudo-Likelihood (Diggle, 2006):	$K^{'}(\rho(i,j)) = \exp\{-(\rho(i,j)/\delta)^{0.5}\} + \rho$
ILM (Deardon et al., 2006)	$K^{''}(\rho(i,j)) = \begin{cases} k_0, & 0 \leq \rho(i,j) \leq \delta_0 \\ \rho(i,j)^{\delta}, & \delta_0 < \rho(i,j) < \delta_{\max} \\ 0, & \text{otherwise} \end{cases}$

value of the kernel evaluated for two farms whose distance is 0 ($K(0)$). Figure 3.7 shows the relative kernel's effect for the different kernels. It is easy to see that for distances less than 5 Km there are considerable differences, whereas for distances greater than 5 Km, the differences are small.

Infectivity and Susceptibility

The infection rate (β_{ij}) consists of three main parts; the transmission kernel, farm's infectivity and farm's susceptibility (see Equation 3.13). The parameters which are associated with farm's infectivity and susceptibility, are ϵ and ξ respectively. We assume that farm's infectivities or susceptibilities could be non−linear in the number of cows or sheep.

The relationships between animal numbers and susceptibility, and animal numbers and transmissibility are estimated via the marginal posterior means of the corresponding parameters ϵ, ξ, and ζ. Figure 3.8 shows the posterior distribution of the model parameters whereas Table 3.4 provides us with the parameter estimates as well as with their corresponding 95% highest posterior density region (HPDR). We have chosen the median of each of the posterior distribution to be our loca-

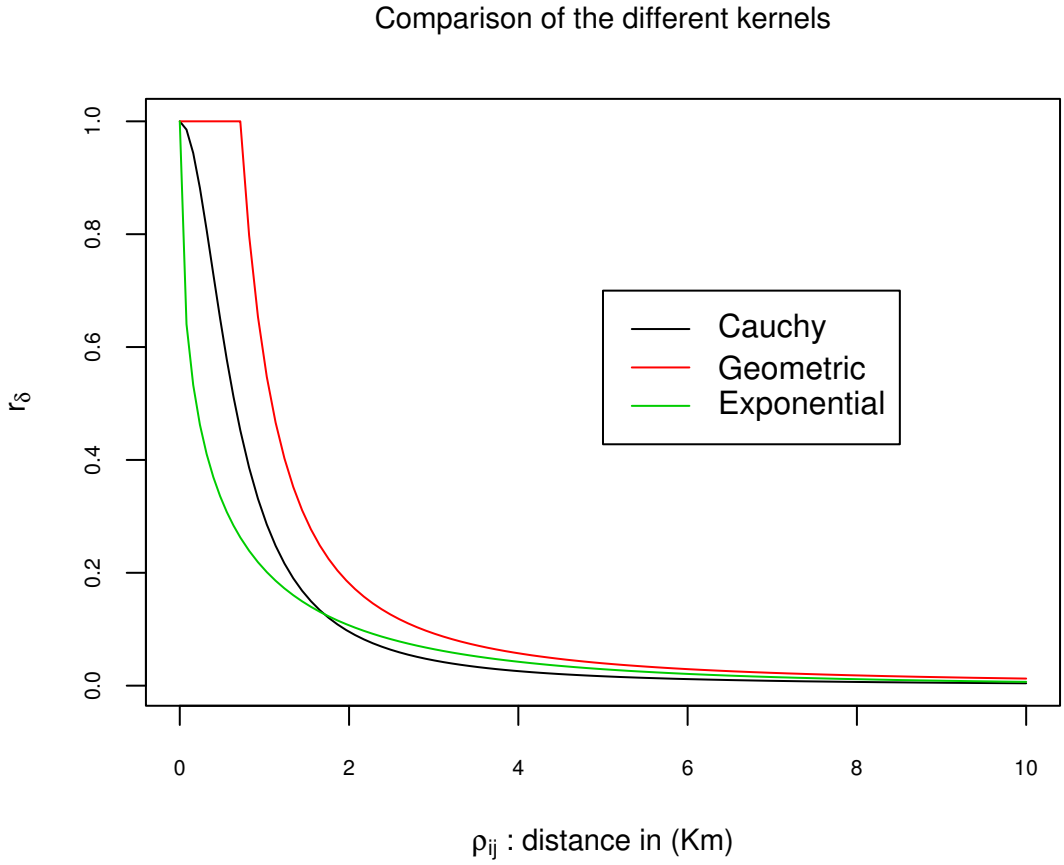

Figure 3.7: The relative kernel's effect for the Cauchy-, geometric- and exponential-type kernels

tion measure. Note that the distribution of ϵ cannot be considered as symmetric and therefore the approximately obtained 95% HPD is not very reliable. This is not the case for the other parameters, since each of them can be assumed to be symmetrical.

Table 3.4 reveals that each individual cow was more likely to transmit the disease ($\epsilon = 1.45$), and also likely to be more susceptible to the disease than each individual sheep ($\xi = 2.45$). Such results lead to qualitatively similar conclusions of those obtained by Keeling et al. (2001) although the authors reported rather different estimates ($\epsilon^{'} = 1.61$ and $\xi^{'} = 15.2$). We should note though the two crucial differences between the two approaches. First, the estimates by Keeling et al. (2001) are based to the total number of farms in the UK while we have focused only in Cumbria; secondly, they have considered the case where $\zeta = 1$ while we allow for non-linearity.

Moreover, Diggle (2006) and Deardon et al. (2006) have also reported different point estimates for the corresponding parameters. Diggle (2006) who also considers a common non-linear effect of the number of animals, reported $\epsilon^{''} = 1.42$ and $\xi^{''} = 36.17$. We should note that the latter parameter is estimated very imprecisely as the reported 95% confidence interval is $(0.19, 692.92)$ and therefore any comparison should be made with care. Deardon et al. (2006) reported $\epsilon^{'''} = 0.57$ and $\xi^{'''} = 7.14$, having assumed a different effect of non$-$linearity for the different species (cows and sheep). Summarizing, although the reported parameters' estimates obtained via the different approaches are very different, all methods lead to similar conclusions about how transmissibility and susceptibility varies with the number of different species in a farm.

Regarding the assumption about a potential non$-$linear effect of the covariates, it turns out that the linear assumption made in the paper by Keeling et al. (2001) is questionable ($\zeta^{'} = 1$). We derived a point estimate of ζ, about 0.32 which is significantly less than one. This is in agreement with the conclusions drawn by Diggle (2006) ($\zeta^{''} = 0.13$) and Deardon et al. (2006). Note that the latter

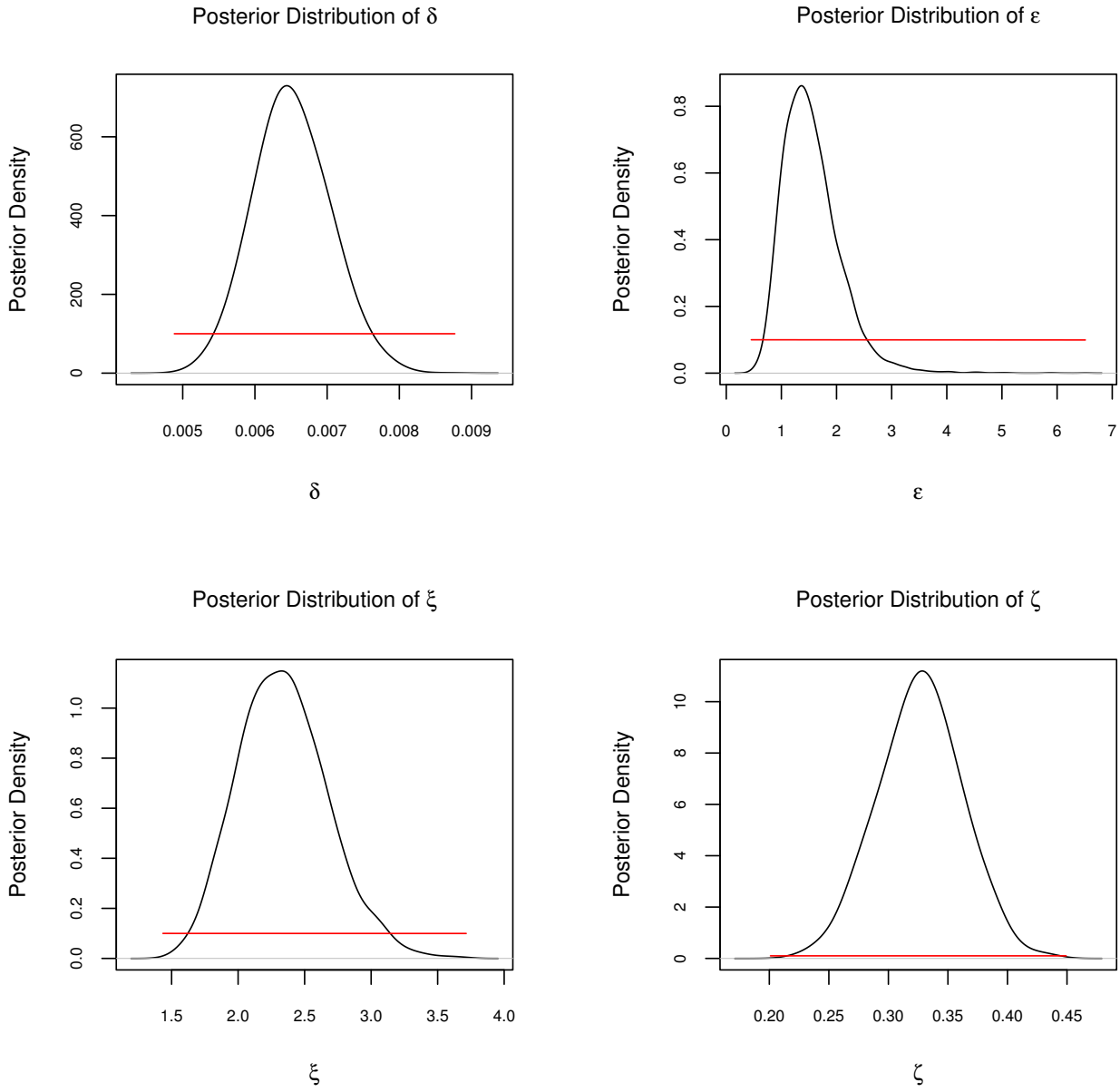

Figure 3.8: Posterior distributions of the model parameters. Red line shows the prior distributions.

Table 3.4: Parameter estimates and approximate 95% highest posterior density region for model's parameters

Parameter	Estimate	95 % Highest posterior density region
Relative infectivity of cattle to sheep (ϵ)	1.45	$0.79, 6.51$
Relative susceptibility of cattle to sheep (ξ)	2.32	$1.74, 3.71$
Non-linear effect of number of animals (ζ)	0.32	$0.25, 0.44$

approach assumes a different non-linear effect for farm's infectivity and susceptibility depending also on the different species. The authors concluded that a linear approximation looks more reasonable in the case of susceptibility since the estimated effect of ζ for a susceptible farm is fairly close to one for both cattle and sheep. On the other hand, in terms of transmissibility a strong non-linear effect is found for both sheep and cattle since. For illustration Figure 3.9 shows how the farm's infectivity and susceptibility increases according to the number of sheep and cattle.

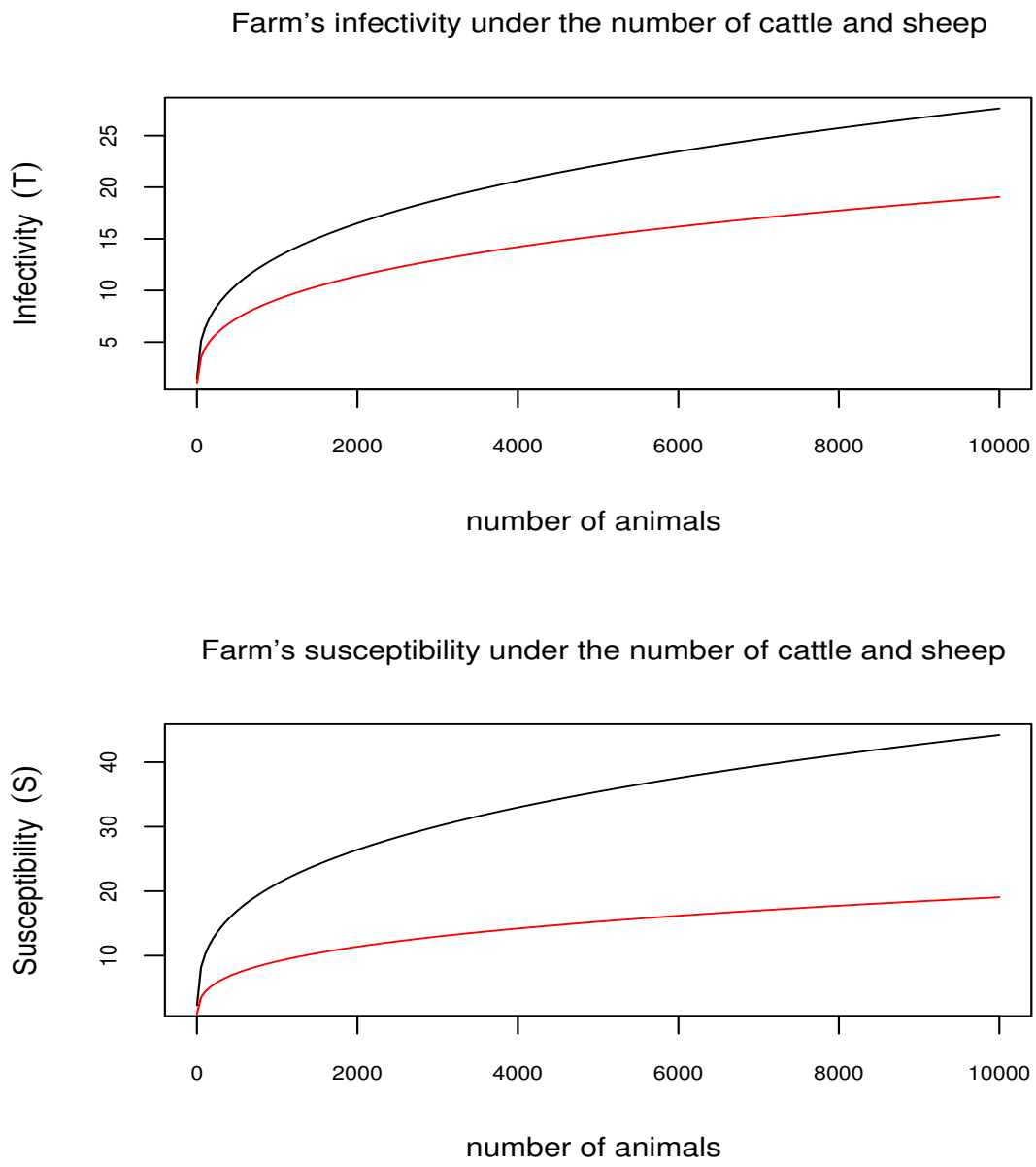

Figure 3.9: Average posterior farm's infectivity under the number of cattle (green) and sheep (red), $T = \epsilon n_c^{\zeta}$ and $T = n_s^{\zeta}$ respectively (top) and susceptibility, $S = \xi n_c^{\zeta}$ and $S = n_s^{\zeta}$ respectively (bottom).

3.3.4 Limitations

Perhaps the most important limitation of our study is the level of detailed information on the dataset which we were given. First, although we have very detailed data for the infected premises, we have no information for farms which were culled

without knowing their status; i.e. whether they were infected or not at the time they were slaughtered. Such information can potentially have an effect for the parameters, such as the spatial kernel or/and farm's infectivity and susceptibility. Moreover, regarding the uninfected farms, the available knowledge is only based on census data. It is uncertain, whether the number of animals in each farm during the outbreak matches the number recorded in the census.

Another important issue which has already been mentioned is how appropriate the use of the Euclidean distance is in comparison to other measures such as minimum walking distance or quickest route. Especially in Cumbria there exists a rich landscape with hills and lakes and therefore if such measures were used it may have lead to different kernel parameter estimates. Nevertheless, such measures cannot be calculate very easily since they are very computationally demanding. It is of interest whether such measure can be used in conjunction with the Euclidean distance to obtain the "optimal measure" for the distance between two farms.

Comparisons between our findings and those obtained from the other studies should be made very carefully. This is due to the fact that we have analysed the outbreak which took place in Cumbria while other studies (Keeling et al., 2001, Deardon et al., 2006) analysed the epidemic which took in the whole UK.

3.3.5 Conclusions

In this chapter we performed a fully Bayesian analysis of the 2001 FMD UK outbreak. Before doing so, we described the approaches adopted by other researchers during the outbreak. These range from a set of deterministic differential equations to a complex simulation model, reflecting the elements that the researchers felt were most important and those that could be neglected. In addition, we briefly reviewed some recently developed work which provides formal statistical inference based solely on the likelihood. The similarities and the differences in the model and the approaches on how inference is made have also been discussed.

Then we introduced a a fully stochastic, heterogeneously mixing epidemic model to capture the dynamics of the outbreak. The infection rate was associated with two main risk factors; the size and the location of the farm. Unlike the other studies, we have assumed that the farm's infectious period is unobserved since the date of infection is not known. A transmission kernel based on the Cauchy distribution was used to allow for long-range infections.

Compared to the other studies, qualitatively similar conclusions were drawn about farm's infectivity and susceptibility. Cattle were found to be more infective and more susceptible to the disease than sheep. In addition, we found a strong non-linear relationship between the infection rate and the number of cattle and sheep in each farm. Moreover, the key feature of the transmission mechanism in our model was the spatial kernel. Our results suggested that over long distances (> 5 Km), the effect of the kernel is very little.

Although many new measures have been implemented to prevent FMD from arriving in the UK, we are still at risk from this and other infections (Yu et al., 1997). It is not certain that a future outbreak will have the same epidemiological characteristics or affect the same livestock to the same degree as the previous one. Therefore, the methods of inference should be very flexible and easy to be adapted to a different model. Within the framework of an HMSE (see Section 2.2) this is straightforward.

The British government considers vaccination to be a viable defence in the face of an outbreak because vaccination is used to combat a wide variety of human and animal diseases. Vaccination especially within the context of human diseases would seem very intuitive. However, reactive vaccination against FMD would need to be implemented in the face of an epidemic and would thus require prohibitively large amounts of trained labour to vaccinate all susceptible livestock rapidly (Keeling et al., 2003). Modelling strategies can play an important role, informing the optimal use of limited resources.

It is probable that as our understanding of spatiotemporal disease dynamics increases and our quantitative knowledge of FMD epidemiology grows, so, too, will models become more complex. This will lead to models which include more factors that may influence the disease dynamics than the ones which have been currently used. Advancements will need to be made if such models are to be of use. In order for such models to be applicable, for far more data (in terms of covariates) are required on a range of facets that may modify the susceptibility to infection or the risk of transmission is essential. In addition, factors such as topography and farm management are both important elements in the spread of infection that could be included in future models if the data were available.

A common feature was absent from the models discussed in this chapter. All models treated the farm as a single unit, such that all the animals became infected *en masse*. In practice, the infection will initially spread through the farm before spreading between farms. However, predicting such dynamics is complicated by our lack of knowledge of the infection at this level of detail. Until such information is available treating the farm as a single infectious unit is as reliable as attempting to simulate within-farm epidemics (Keeling, 2005).

Concluding, one of the most important advancements for the future would be to combine the expertise of modellers, veterinarians and those responsible for implementing policy. Therefore, veterinary judgement, experience and local knowledge can provide the most accurate assessment of infection risk for particular farms. Such expert's opinion can be very easily incorporated by adopting a Bayesian framework. The methodology presented in Chapter 2 will then offer a variety of robust MCMC algorithms to draw inference for the parameters of interest.

Chapter 4

Future Work

In this chapter we first discuss various extensions to the methodology presented in Chapter 2. Then we briefly refer to future work on modelling the 2001 UK FMD outbreak. Finally, we also present some ongoing work on modelling a potential Avian Influenza outbreak in the poultry industry of the UK.

4.1 Methodology

4.1.1 Infectious Periods

It has already been mentioned that although the GSE and the HMSE are very simple models, they can become surprisingly very challenging for modern MCMC methods due to the high dependence between model parameters and infection times. Nevertheless, we have shown how mixing problems can be overcome by introducing partially (or completely) non−centered parameterisations.

So far, we have considered an HSME where each individual i, remains infectious for some time, say D_i, which is Gamma distributed with with mean α/γ and variance α/γ^2:

$$D_i \sim Ga(\alpha, \gamma).$$

Throughout the simulation study which was performed in Section 2.6, we have assumed that α was known. It would be very interesting to assess the performance of centered and the non−centered algorithm when both α and γ are unknown and inference needs to be drawn for both of them. Preliminary work shows that when α is large or the number of infected individuals increases, then the mixing of the Markov chains deteriorates. In addition, there exists high correlation between α and γ. Although the mixing for both α and γ is bad, the mixing of average infectious period, α/γ, is much better. Intuitively, this can be explained by the fact, that the observed data contain more information about the mean infectious period, rather than the parameters α and γ separately. In this case, a centered parameterisation in a similar manner to those presented in Section 2.3 can be potentially useful in order to break the correlation between the model parameters. However, such a reparameterisation will not break the dependence link between γ and the missing data (infection times) and therefore a (partially) non−centered seems to be appropriate to improve the mixing of the standard (centered) algorithms.

Finally, if other distributions for the infectious period of the individuals are assumed, such as a Weibull with parameters α and γ, i.e.

$$D_i \sim Weib(\alpha, \gamma),$$

then it will be of a particular interest to see the performance of the various centered and/or non−centered algorithms and whether mixing problems occurs with such an infectious period.

4.1.2 Epidemics in Progress

The approach via which Bayesian inference was drawn for the HMSE relies on the assumption that the epidemic was completed by the time T_{obs} we observed it. In practice, it is often the case to be mostly interested in drawing inference for the parameters associated with the chosen model while an epidemic is in progress.

In Section 2.1.8.2 we have shown how MCMC methods can provide inference for the parameters of the GSE, taking into account the uncertainty about the infectious status of each the individuals, i.e. whether they have been infected or not at time T_{obs} we observe the epidemic. Although the importance of GSE model in epidemic theory is widely known, it is not always a realistic model to model many real life outbreaks. Mainly this is due to the fact the heterogeneity between the individuals in the population is not taken into account. Therefore, we will briefly describe in this section our approach for performing real-time inference about the parameters of the HMSE model.

Our approach is along the lines of the approach by O'Neill and Roberts (1999). Consider the structure of the HMSE model as described in Section 2.2 with the following forms of the infection and removal rates.

$$\begin{aligned} \beta_{ij} &= \beta_0 \cdot h_{ij} \\ R_i - I_i &\sim Ga(\alpha, \gamma) \end{aligned}$$

where I_i and R_i denote the infection and the removal time of the infected individual i respectively. Denote by n_I, n_s and n_R the number of infected, susceptibles and removed individuals by time T respectively. An individual which is infected but not detected by time T is considered as *occult* individual.

We propose to implement the following MCMC algorithm:

MCMC Algorithm

(Repeat the following steps)

1. Start the chain with initial values for the parameters $\boldsymbol{\theta}^0$ and the set of infection times $\boldsymbol{I}^0$;
2. Update β_0 and γ via Gibbs steps;
3. Choose one of the following steps with equal probability:
   (a) Move an infection time;
   (b) Add an infection representing an *occult* individual;
   (c) Delete an infection time which has been previously added;
4. Goto 1.

Step 3 is described here in more detail. We move the time of an existing infection, or add or delete an infection time as described above. The three events are performed with equal probability. We denote by $\boldsymbol{I} - \{t\} + \{s\}$ a move, $\boldsymbol{I} + \{s\}$ an addition, and $\boldsymbol{I} - \{t\}$ a deletion. Thus whenever we add an infection, n_I increases by 1, and n_S decreases by 1. The reverse is true for a deletion, while for moving an infection time, n_I and n_S remain constant. This is implemented as follows:

3a. **Move an infection time**: We choose an infection time to move from a discrete uniform distribution, $U(1, n_I)$, and propose a replacement infection time drawn from distribution a chosen distribution with probability density function $q(\cdot)$. We accept the proposed value with probability

$$1 \wedge \frac{f(\boldsymbol{I} - \{t\} + \{s\}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})}{f(\boldsymbol{I}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})} \times \frac{q(\boldsymbol{I}|\boldsymbol{I} - \{t\} + \{s\})}{q(\boldsymbol{I} - \{t\} + \{s\}|\boldsymbol{I})}$$

3b. **Add an infection**: We choose an occult infection from the susceptibles

using a discrete uniform distribution $U(1, n_S)$. An infection time is then chosen from the uniform distribution $U(\min(\boldsymbol{I}), T_{obs})$. We accept such an addition with probability:

$$1 \wedge \frac{f(\boldsymbol{I} + \{s\}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})}{f(\boldsymbol{I}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})} \times \frac{n_S(T_{obs} - I_k)}{m+1}$$

where m is the number of previously added infections prior to the addition.

3b **Delete an infection**. We choose an infection time to delete from a discrete uniform distribution over the premises that have been previously added. We accept such a move with probability:

$$1 \wedge \frac{f(\boldsymbol{I} - \{t\}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})}{f(\boldsymbol{I}, I_m|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})} \times \frac{m}{(T - I_k)(n_S - 1)}$$

where m is the number of previously added infections prior to the deletion.

Such an algorithm is very similar to the one used by O'Neill and Roberts (1999) in the context of the GSE. However, the authors have considered a very small dataset which consists of 10 individuals. In real life outbreaks the initial susceptible population is typically very large, For instance, in the 2001 UK FMD outbreak, the farms at risk in the beginning of the outbreak, were about 120,000. Moreover, suppose that we are interested in modelling a potential Avian Influenza outbreak in the poultry industry of the UK. In this case, the farms at risk are about 40,000.

Simulation studies have shown so far, that when the number of the susceptible farms increases, the MCMC algorithm described above becomes very computationally costly and very inefficient. Intuitively, this can be explained by the fact that due to heterogeneity, choosing individuals uniformly and assign them an infection time does not always increase the likelihood and many of such proposed steps get rejected. Therefore, this leads to very slow mixing Markov chains. Moreover, the same holds when choosing uniformly individuals to delete their infection times.

An alternative and more efficient approach is to take into account, the infectious pressure (as defined in Section 2.1.6.3) that an individual gets at any time. In other words, it is more sensible to choose an individual among the susceptible population to add an infection time for, with probability which is proportional to the infectious pressure which is subjected. Note that the infectious pressure is determined through the infection rate, β_{ij} which takes into account a series of potential risk factors. Similarly, when choosing individuals in order to delete their infection time, it makes more sense to be chosen with probability conversely proportional to their infectious pressure.

Nevertheless, although such an approach can become very effective compared to simpler one, when the size of population increases, the computational cost of the former is much larger than the computational cost of the latter. This is because, if an addition is chosen, then the infectious pressure for the individuals in the susceptible population (which is typically large) must be calculated. An interesting question is whether it is necessary to compute the infectious pressure at every step of the MCMC algorithm or not. For instance, one could calculate the individual's infectious pressure only every 100 say, or more, iterations. Alternatively, if it is believed that the epidemic has been "established" in the sense that plausible estimates for the parameters can be obtained, then these estimates could be used for computing the corresponding individual's infectious pressure.

In conclusion, another issue which should carefully been considered in the future, is the effect of the number of additions, moves or deletions to the mixing of the Markov chains. The illustrative example in Section 2.6 showed that the more infection times we choose to update, the better the mixing it gets; although the algorithm runs shower in time. In the case of an epidemic in progress, addition and deletion of infection times take place as well. Therefore, it should be examined what is the trade off between the number of times we shall repeat Step 3 and the performance of the MCMC algorithm.

4.2 Applications

4.2.1 A Comprehensive Bayesian Analysis of the 2001 FMD Outbreak

Although we adopted a fully Bayesian analysis to analyse the outbreak which took place in Cumbria, it would be of interest to assess the assumption by Diggle (2006). He claims that because the two counties (Cumbria and Devon) are geographically separated, it shall be treated informally as two replicates of a natural experiment, thus allowing to compare parameter estimates and pool as appropriate. In addition, an analysis considering the total number of farms in the UK would allow us to compare our methodology with the one adopted Deardon et al. (2006) in terms of using a discrete or a continuous model setup.

Having a more detailed and accurate dataset than the currently available, a more sophisticated analysis can be performed where we will be able to incorporate and infer about the infectious status of the farms which had animals slaughtered without being identified as IPs.

Adopting the methodology in Section 4.1.2, it would be very interesting to investigate after which day of the outbreak and onwards there was no much more variation in the parameter estimates. This for instance, could lead to an assessment of DEFRA's policies regarding for controlling strategies.

4.2.2 Modelling a Potential Avian Influenza Outbreak in the UK

It is remarkable the threat that governed UK when the poultry industries of many other European countries were affected by the Highly Pathogenic Avian Influenza (HPAI) disease. In this section we will summarize ongoing work on modelling a potential a potential AI outbreak in the UK.

4.2.2.1 The Model

We use a realistic complex stochastic model for the evolution of an outbreak, parameterised by a number of unknown parameters. Whilst expert opinion can be relied upon to provide reasonable estimates of many of these parameters, the absence of an H5N1 epidemic in the UK poultry industry to date inevitably implies that some uncertainty about parameter values remains.

Our approach to the problem will be Bayesian since

- we would like to incorporate expert's opinion in the form of the parameter prior distributions;
- we would like to derive a real-time risk assessment tool as the epidemic evolves;

We propose to model a potential outbreak of Avian Influenza in the UK using a stochastic epidemic model of the form of an HMSE.

Consider a total population of size $\mathcal{N}$ farms. At any given time point each farm i can be in one of four states:

- **S**usceptible premises (SP) do not have the disease and are able to be infected by it;
- **I**nfected premises (IP) have the disease and are able to infect susceptibles. Their infectivity increases as a function of time.
- **N**otified premises (NP) have been detected as having the disease and are subject to government-imposed movement restrictions. However, they are still capable of infecting susceptibles by other means, such that their infectivity is curbed at a lower level.
- **R**emoved premises (RP), in the case of AI, have had their flocks culled and therefore play no further part in the epidemic.

Thus the only transitions we allow are: from susceptible to infected, from infected to notified, and from notified to removed.

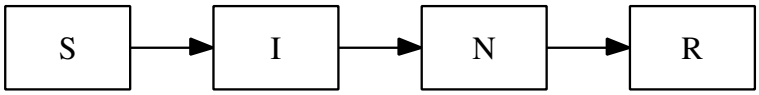

Figure 4.1: The four compartments of the SINR model.

We assume that the epidemic is observed up to a certain time, say T_{obs}. Denote by $n_I \leq N$ and $n_R \leq N$, the number of individuals who got infected and removed by time T_{obs} respectively. In general, $n_I \leq n_R \leq N$. The SINR model has the following properties:

we consider a heterogeneously mixing population, with a *time-dependent* transmission rate from farm i to farm j:

$$\beta_{ij}(t) = T_{ij} \cdot h(t) \tag{4.1}$$

where $h(t)$ represents how a farm's infectivity changes over time up to a maximum of T_{ij}. It is defined as follows:

$$h(t) = \frac{1}{\mu + \nu \exp\{-t\}}, \qquad \mu, \nu > 0 \tag{4.2}$$

It easy to see that different values of μ and ν give a large variety of different shapes of the farm's *infectivity profile*.

The quantity T_{ij} involves six parameters which drive the infection rate. $\boldsymbol{I}, \boldsymbol{N}, \boldsymbol{S}$ define the sets of the infected, notified and removed farms respectively:

$$T_{ij} = \begin{cases} \beta_{ij}, & i \in \boldsymbol{I}, j \in \boldsymbol{S} \\ \\ \beta^{\star}_{ij}, & i \in \boldsymbol{N}, j \in \boldsymbol{S} \end{cases}$$

Parameters β_{ij} and $\beta^{\star}_{ij}$ are modelled appropriately so as we distinguish between possible disease transmission by managemental contacts, and environmental factors such as rodents, walkers, and short-range dust-borne infection. Therefore, factors such as the probability of any two farms being connected by the requisite means of contact per day and the distance between them are taken into account. Since an IP has movement restrictions imposed upon it, we assume that all contact components are severed, and only environmental transmission is possible. Furthermore, we assume that since the biosecurity of the IP is likely to have been increased, the rate of spatial spread for a NP is lower than for an IP.

In contrast to the model used so far, we assume that the time from infection to notification $(D|I = N - I)$ is distributed according to the following distribution:

$$P(D > d) = \exp\{-a(\exp\{b \cdot d\} - 1)\} \tag{4.3}$$

where $a, b > 0$.

4.2.2.2 Challenges

Bayesian inference for the parameters of interest can be drawn via the algorithm mentioned in this chapter. However, the size of the susceptible farms is relatively large and therefore, care is needed in order to construct robust and feasible algorithms. Moreover, unlike the HMSE model (as defined in Section 2.2), the SINR model assumes a fixed and known period between notification and removal.

It would be interesting to see how well the centered algorithm performs in the case where the parameters a and b are unknown. Intuitively, these parameters are *a priori* dependent with the infection times and therefore the implementation of a non−centered reparameterisation sounds essential. Questions similar to those mentioned in Section 4.1.2 should also be answered so as to be able to provide precise answers in a reasonable time and inform governmental organisations about optimal policies.

4.3 Computational Issues and Parallel Computing

It can be easily understood that when we are aiming to analyse real life applications such as FMD or AI, the datasets we are concerned with, are of very high dimension. An increase in the dimension of parameter's space would result to more computationally costly algorithms. Although such problems can be overcome by the methods introduced in Deardon et al. (2006), there are still parts in the algorithm which can slow down the time that an algorithm needs to run. Methods of parallel computing seem to be ideal for static Monte Carlo algorithms (eg importance sampling), however their applicability within a dynamic Monte Carlo framework (i.e. MCMC) is not very straightforward. That is because we need to keep track of the states which the Markov chain has visited. Nevertheless, parallel computing methods have much to offer in evaluating parts of the likelihood which can be linearised, for example those which appear in the likelihood of an HMSE (see Section 2.2). Therefore, we are interested in making use of such algorithms in order to be able to efficiently apply our methodology in real-time.

Part II

A New Class of Semi–Parametric Time Series Models

Chapter 5

Latent Branching Trees

5.1 Introduction

There are many aspects of a number of observed univariate time series which cannot be adequately modelled by standard time time series modelling. For instance, non-Gaussian marginal distributions and dependence beyond autocorrelations. It is also often the case that the assumption of Normally distributed observations is inappropriate because the variable being modelled has a positive and highly skewed distribution, eg. wind speeds and daily flows of a river. Therefore, in this chapter, we are mainly concerned with the construction of a class of semi-parametric time series models of infinite order where we will be able to specify the marginal distribution of the observations in advance and then build their dependence structure around them. Such a class of models can be very useful in cases where data are collected over long period and it might be relatively easy to indicate their marginal distribution but much harder to infer about their correlation structure.

Regarding the structure of this chapter, we first briefly review previous work on modelling time series with non-Gaussian margins and various correlation structures. In addition, we explain our motivation by giving examples of time series we will be interested in analysing. In Section 5.4 we introduce a stochastic process,

which we term it a *latent branching tree* (LBT) and constitutes the base of the class of time series we will develop. Furthermore, in Section 5.5, we will describe in detail the general properties of the LBT and also show some illustrative data sets generated by such a construction (Section 5.6). Exact simulation for a LBT and methods for drawing Bayesian inference for the associated parameters of interest, are presented in Sections 5.7 and 5.8 respectively. Sophisticated algorithms which could improve the efficiency of the standard algorithms are illustrated via applications on simulated datasets in Sections 5.9 and 5.10. An application on some genome scheme data is presented in Section 5.11 and finally, we discuss potential extensions on the current methodology in Sections 5.12 and 5.13.

5.2 Literature Review

The development of methods for constructing stationary time series models with marginals of choice has been of particular interest in the literature. As it has already been mentioned, typically such models are useful when marginal inspection from the data is feasible. Early work of this type of construction, outside the Gaussian framework can be found in Lawrance and Lewis (1977), Jacobs and Lewis (1977) and Gaver and Lewis (1980) where positive real-valued AR-type models with Gamma marginal distributions were proposed. There has also been a number of other examples concerned with Exponential (Gaver and Lewis, 1980, McKenzie, 1982), Gamma (Gaver and Lewis, 1980) and mixed Exponential (Lawrance, 1980) distributions. There has been an extensive literature on constructing Markov models with short-term dependence behavior; a key work on the development of such stationary time series with pre-specified marginals is that of Lawrance and Lewis (1985), whilst more recently, Joe (1996) and Jørgensen and Song (1998) have highlighted a unified approach for constructing stationary AR-type models.

Models with short-range memory like ARMA model and Markov processes are well known (see for example, Brockwell and Davis, 1991) and often used in practice.

Apart from Markov models, there has been a considerable interest in constructing infinite order models and developing methods in order to analyse data for long-range dependence. Datasets with such a behavior often appear in geophysics, hydrology and astronomy. Beran (1992) provides a review of statistical methods for data with long-range dependence. The two best known classes of stationary processes with slowly decaying correlations are increments of self-similar processes (in the Gaussian case so-called fractional Gaussian noise) and fractional ARIMA processes.

Self-similar processes and the corresponding increment processes were first introduced to statistics by Mandelbrot and collaborators (see for example, Mandelbrot and Van Ness, 1968). Brownian motion is self-similar and was known for long time, while Lamperti (1962) points to the fact that normalized sums of random variables converge to self-similar processes (see for details Beran (1994)).

Fractional ARIMA models were introduced by Granger and Joyeux (1980) and Hosking (1981). They are a natural generalisation of standard ARIMA(p,d,q) models defined in Box and Jenkins (1970). Later, a generalisation of fractionally ARIMA models have been proposed by Gray et al. (1989).

Summarizing, it can be easily seen that there has been a considerable amount of focus and interest on developing stationary time series with pre-defined marginal distribution of the observations and short-term covariance structure around them. In addition there has also been an extensive literature on statistical methodology for modelling long-range dependent data. Most of the work cited so far, relies on rather different techniques to derive the appropriate margins and dependence. In this chapter, we will introduce a unified framework based on a latent stochastic process through which a class of infinite order time series models with the desired properties is obtained.

5.3 Motivation

In this section, we will first refer to some motivating examples in order to illustrate the idea behind the time series we are interested in modelling. Then, we will review the work by Neal (2003) which motivates the adoption of the introduced stochastic process in Section 5.4.

5.3.1 Examples

Suppose we are interested in modelling a time series such as the one shown in Figure 5.1. The histogram of the data reveals that they are Normally distributed and the corresponding ACF plot indicates slowly decaying autocorrelation (see Figure 5.2). A transparent characteristic of this data set is that a significant number of clusters exist. Since there seems to exist some dependence structure within the cluster, analysis of such data via standard change-point problem does not seem appropriate. That is, because such an analysis requires the assumption of *iid* observations within the clusters in order to make the inference feasible. From a practical point of view, such data structures often appear in genome scheme data; in particular, at sequences of a DNA where there exist long genome segments homogeneous in C+G, termed as *isochores*.

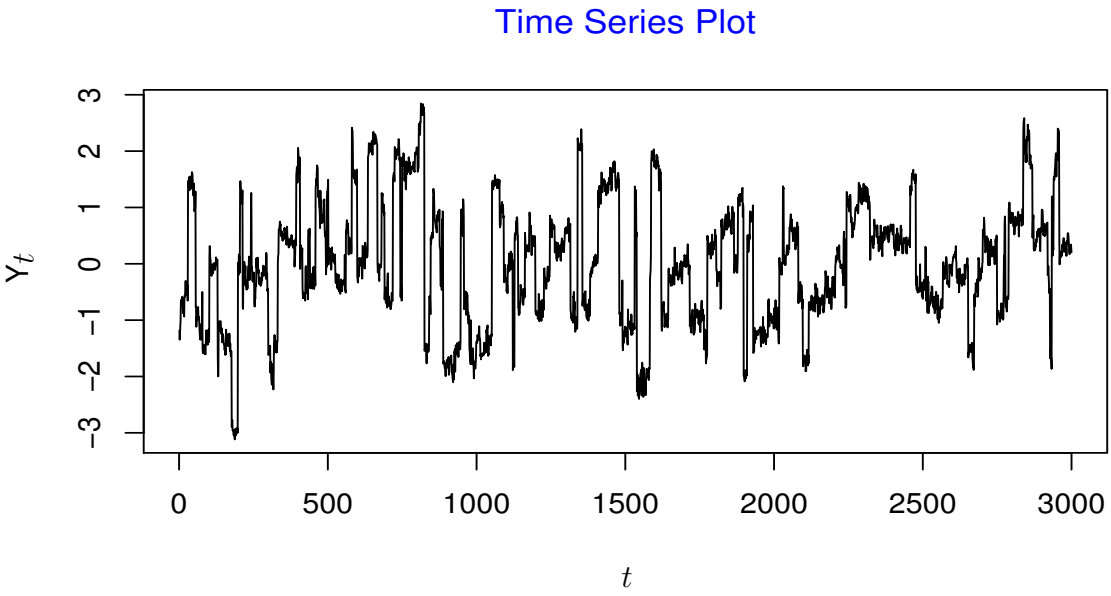

Figure 5.1: A series 3,000 observations collected over time.

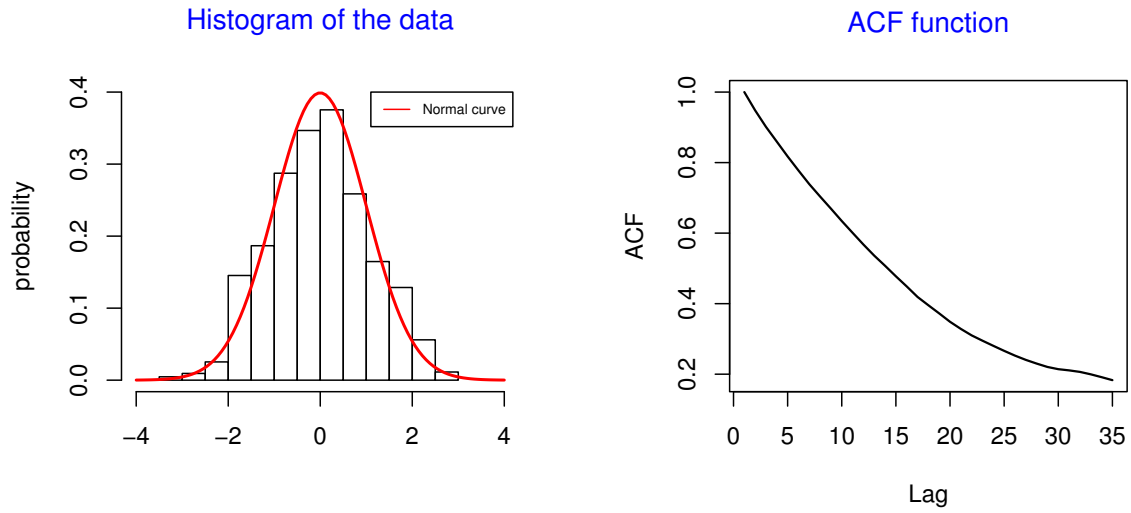

Figure 5.2: Histogram (left) and ACF plot for the data shown in Figure 5.1(right).

5.3.2 Dirichlet Diffusion Trees

Neal (2003) introduced a class of prior distributions over distributions which can be seen as generalization of Dirichlet mixture models (Ferguson, 1973, 1983, An-

toniak, 1974). Unlike simple mixtures, these priors can capture the hierarchical structure which is present in many distributions. The construction of these priors is based on an underlying latent process, defined as a *"Dirichlet diffusion tree"(DDT)*. Such a process, also provides a hierarchical clustering of the data with probabilistic indications of uncertainty. In this section we will describe the procedure of obtaining a DDT in order to randomly generate a data set of n points each a vector of p real numbers in which the data points are drawn independently from a common distribution drawn from the prior.

In general, each point in the data set drawn from a DDT is generated by following a path of diffusion process. In principle any diffusion process can be used although the author only considered a Gaussian diffusion process. The first data point is generated in the following way: the Brownian motion begins from an origin which is fixed at zero and operates for a length of time which without any loss of generality, can also be fixed at one. The end point of this first path is the first point in the data set.

The second point in the data set is also generated by following a path from the same origin ($t = 0$) and for the same time ($t = 1$). This second path initially follows the first but after some time, T_d say, it diverges to another which is independent of the remainder of the first path. The end point of the second path is second data point.

In general the i_{th} point in the data set is generated by following a path from the origin that initially coincides with the path to the previous $i - 1$ data points. If the new path has not diverged at a time when paths to past data points diverged, then it chooses between the previous path with probabilities proportional to the number of paths that went each way. Once a path diverges, the new one moves independently of the previous paths.

The distribution of the divergence time, T_d can be expressed in terms of a divergence function $\alpha(t)$. Neal (2003) explores the properties of a DDT by looking at

data sets which have been created by using a variety of divergence functions $\alpha(t)$. Different choice of the divergence function will lead to different behavior of the prior generated by the DDT.

However we should note that a density function must be absolutely continuous and a prior generated by a DDT must satisfy this condition. Therefore the selection of a divergence function $a(t)$ must be very careful. Neal (2003) has investigated empirically when the DDT produce an absolutely continuous density by looking at distances to nearest neighbors in a large sample from the distribution.

In order to prove that the prior is "exchangeable" he proved a stronger property that the probability density for producing a data set along with its underlying tree structure (locations and times) is the same for any ordering of the data points. Once this is proved, then the exchangeability follows by summing over all possible tree structures and integrating over times and location of divergences.

Concluding, the author also provides a number of generated data sets and also briefly describes how MCMC methods can be applied to sample from the posterior distribution as long as with the parameters of the underlying structure of the tree such as divergence times, locations and hyperparameters such as diffusion variances. In the following section we will show how a modification of the DDT can lead to the generation of a time series model with specific properties.

5.4 Construction of a LBT

In this section, we describe in detail the general latent branching tree construction. The result of such a construction will be the generation of a data set containing n say, points, i.e. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ which follow a specified distribution. Throughout the construction of the LBT we will state the similarities and the differences between it and the DDT.

The first data point Y_1, is drawn by a diffusion denoted by $X_1(t)$ which starts from

a fixed origin and operates for a fixed length of time. In principle, *any diffusion* can be used but for mathematical convenience and without loss of generality we use a standard Brownian motion. Furthermore, we assume that $X_1(t)$ operates for a time interval of length equal to 1 starting from 0, i.e. $X_1(t),\ 0 \leq t \leq 1$. The value of the Brownian motion at time $t = 1$ is considered as the first data point denoted by $Y_1 = X_1(1)$.

Consider the following Brownian motion, $X_2(t),\ 0 \leq t \leq 1$ where:

$$\begin{array}{rcll} X_2(t) & = & X_1(t) & \text{for} \quad 0 \leq t \leq \tau_1 \\ X_2(t) - X_2(\tau_1) & \perp & X_1(t) - X_1(\tau_1) & \text{for} \quad \tau_1 < t \leq 1 \end{array}$$

where $a \perp b$ denotes that "a is independent of b". In words, the process $X_2(t)$ traverses the same path as $X_1(t)$ up to a (divergence) time point τ_1 and then it diverges to another path which is independent of the previous one. The second data point is the value of the second Brownian motion at $t = 1$, i.e. $Y_2 = X_2(1)$. These first two steps are exactly the same as the ones you need to produce the first two data points from the DDT. However the generation of the following data differs.

In general the ith data point is the value of the (i)th Brownian motion, i.e. $Y_i = X_i(1), i = 0, \ldots, n$ where:

$$\begin{array}{rcll} X_i(t) & = & X_{i-1}(t) & \text{for} \quad 0 \leq t \leq \tau_{i-1} \\ X_i(t) - X_i(\tau_{i-1}) & \perp & X_j(t) - X_j(\tau_{i-1}) & \text{for} \quad j = 1, \ldots, i-1 \text{ and } \tau_{i-1} < t \leq 1 \end{array}$$

In general, in order to generate n data points, n diffusions and $n - 1$ divergence points are required. For illustration, a visualization of a latent branching tree is shown in Figure 5.3. We decide in advance the *"jump distribution"* to be a Uniform $[0, 1]$, i.e. $\tau_i \sim U[0, 1]$. Firstly, the first BM path from 0 to 1 is simulated and $Y_1 = 1.55$ is obtained (upper-left). The upper-right plot, shows that the second BM follows the same path as the first one up to time $\tau_1 = 0.122,$ where $X(\tau_1) = 0.05$

and then it follows another (independent) path which leads to the second data point $Y_2 = -2.63$. In a similar way the third data point $Y_3 = -2.88$ is generated where the third BM diverged at time $\tau_2 = 0.595$, where $X(\tau_2) = -1.819$ (bottom-left). Finally the fourth BM follows the common path of the previous three up to time $\tau_3 = 0.312$, where $X(\tau_3) = -1.05$ and then it diverges by traversing another path which leads to the fourth data point $Y_4 = 0.015$ (bottom-right).

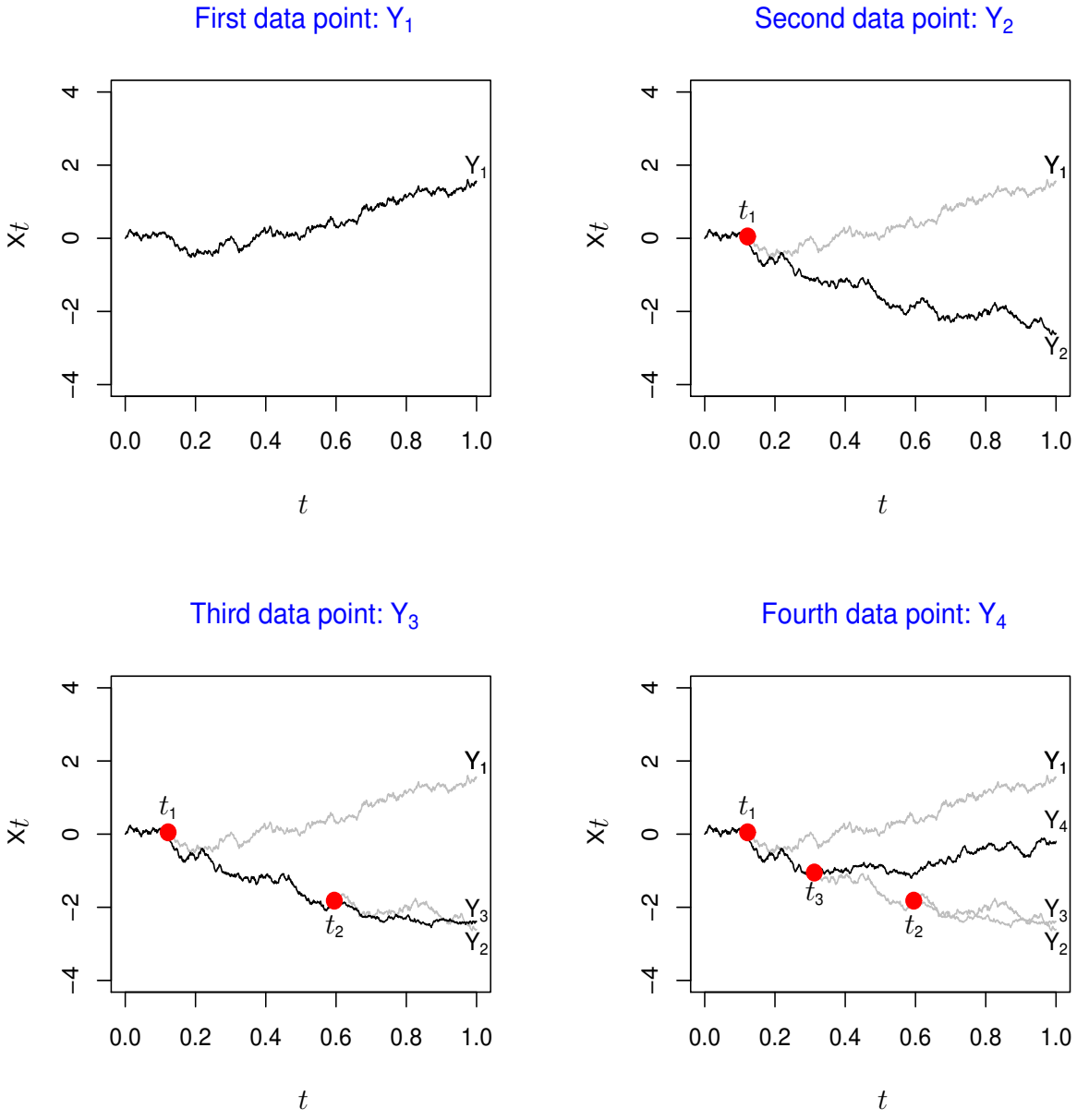

Figure 5.3: Generation of four real data points (Y_1, Y_2, Y_3, Y_4) where the *"jump distribution"* is Uniform$[0, 1]$. (see the text for more details)

5.5 General Properties of a LBT

5.5.1 Marginal Distribution of the Data

By construction, whatever the *"jump distribution"* is, the marginal distribution of each of the data points, $Y_1, \ldots, Y_n$, is $N(0,1)$. This is due to the choice of the diffusion, i.e. Brownian motion in this case. Although at first glance this sounds restrictive, in fact it is not, because we can derive different marginal distributions for the data either by adding an another level to hierarchy of the model or by a choosing a different diffusion than a Brownian motion.

Firstly, we describe how by appropriate transformations we can get a variety of non-Normal distributions having as origin a $N(0,1)$ random variable. For example, if $Y_i \sim N(0,1)$ then $(Y_i)^2 \sim X_1^2$. Table 5.1 shows the appropriate transformations to obtain a variety of very well known distributions. Note that:

$$\Phi(m) = \frac{1}{\sqrt{2\pi}} \int_0^m e^{-\frac{k^2}{2}} dk,$$

Table 5.1: A variety of transformations of a standardized Normal Variable $X \sim N(0,1)$

Distribution	**Parameters**	**Transformation**
Normal	μ, σ^2	$\mu + \sigma X$
logNormal	μ, σ^2	$\exp\{\mu + \sigma X\}$
Gamma	a, b	$ab\left(X\sqrt{\frac{1}{9a}} + 1 - \frac{1}{9a}\right)^3$
Exponential	λ	$-\frac{1}{\lambda}\log\left(\frac{1}{2} + \Phi(X)\right)$
Uniform	a, b	$a + (b-a)\left(\frac{1}{2} + \Phi(X)\right)$

Instead of adding another level to the hierarchy ($\boldsymbol{\tau} \rightarrow \boldsymbol{X}(\tau) \rightarrow \boldsymbol{Y}$) we can choose a different Lévy process. Nevertheless, in order to make the construction of the

LBT feasible, it is essential that we can simulate from the such a process and also, that we are able to write explicitly the conditional density of intermediate points of the diffusion.

Choosing a different process will lead to a different marginal distribution. If we choose a Gamma process (i.e. independent Gamma increments) then the realizations will not longer be Normally distributed but they will follow a Gamma distribution. If we are interested in modeling heavy-tailed distributions then we might need to use a Cauchy process.

In general, apart from real-valued time series we are also able to produce integer-valued observations from discrete distributions. This can be achieved by choosing a process which will lead to discrete observations. For instance, a Poisson process will produce Poisson variables. Moreover, one can also use a *random walk* in discrete space such that a discrete distribution will be generated.

The above extensions reveal the flexibility of a LBT and give an idea how one can generalize it in order to obtain different marginal distributions of the realizations.

5.5.2 Covariance Structure

The most important property of a construction of a LBT is the correlation structure of observations. Suppose that only two data points have been generated from such a stochastic process, Y_1, Y_2 say, where the 2nd Brownian motion diverged at time τ_1. The covariance between Y_1 and Y_2 is:

$$\text{cov}(Y_1, Y_2) = \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2] = \mathbb{E}[Y_1 Y_2] = \mathbb{E}_{\tau_1}[\mathbb{E}[Y_1 Y_2|\tau_1]].$$

This expectation can be easily calculated by taking into account that because of the independence between the normal increments the expectations $\mathbb{E}[X(\tau_1) \cdot (Y_2 - X(\tau_1))]$, $\mathbb{E}[(Y_1 - X(\tau_1)) \cdot X(\tau_1)]$ and $\mathbb{E}[(Y_1 - X(\tau_1)) \cdot (Y_2 - X(\tau_1))]$ are equal to

zero. Therefore,

$$\begin{aligned}\mathbb{E}[Y_1Y_2|\tau_1] &= \mathbb{E}[(X(\tau_1)+Y_1-X(\tau_1))\cdot(X(\tau_1)+Y_2-X(\tau_1))] \\ &= \mathbb{E}[X^2(\tau_1)] \\ &= \text{var}(X(\tau_1)) \\ &= \tau_1\end{aligned}$$

Hence,

$$\mathbb{E}[Y_1Y_2] = \mathbb{E}_{\tau_1}[\tau_1] = \tau_1$$

i.e. the covariance between the first two data points it is equivalent to the time that the 2nd Brownian motion diverged. i.e. the greater the time τ_1 the higher the correlation becomes. We can then state the following theorem:

Theorem 1 *Let $Y_1, Y_2, \ldots, Y_n$ be the data points generated by a latent branching tree and denote by $\tau_1, \ldots, \tau_{n-1}$ the divergence time points, where $\tau_i \sim F$, $i = 1, \ldots, n-1$. The covariance of Y_1 and Y_n is equal to the expected value of the minimum of the τ_i's, i.e.*

$$\text{cov}(Y_1, Y_n) = \mathbb{E}[\min(\tau_1 \ldots, \tau_{n-1})]$$

Proof Denote by τ^* the minimum of the divergence points:

$$\tau^* = \min(\tau_1, \ldots, \tau_{n-1}).$$

Each of the data points can be written as follows:

$$\begin{aligned}Y_1 &= X(\tau^*) + W_1 \\ Y_n &= X(\tau^*) + W_2\end{aligned}$$

where $W_1 \sim N(0, 1-\tau^*)$ and $W_2 \sim N(0, 1-\tau^*)$. Both these independent random variables can also be seen as sums of m, $m > 0$ independent normal variables each of them having mean zero and variance σ_i^2 which satisfy the condition $\sum_{i=1}^{m} \sigma_i^2 = 1 - \tau^*$. The covariance between the first and the last data point is

$$\begin{aligned} \text{cov}(Y_1, Y_n) &= \mathbb{E}_Y[Y_1 Y_n] \\ &= \mathbb{E}_{\tau^*}[\mathbb{E}_Y[Y_1 Y_n | \tau^*]] \end{aligned}$$

By following a similar approach as we did for the first and the second data point we can evaluate the following expected value:

$$\begin{aligned} \mathbb{E}_Y[Y_1 Y_n | \tau^*] &= \mathbb{E}_Y[(X(\tau^*) + W_1) \cdot (X(\tau^*) + W_2)] \\ &= \mathbb{E}_Y[X^2(\tau^*)] \\ &= \text{var}(X(\tau^*)) \\ &= \tau^* \end{aligned}$$

Therefore it is easy to derive that:

$$\mathbb{E}_Y[Y_1 Y_n] = \mathbb{E}[\tau^*] = \mathbb{E}[\min(\tau_1, \tau_2, \ldots, \tau_{n-1})] \tag{5.1}$$

□

It can be easily seen from (5.1) that the covariance between Y_1 and Y_n depends on F and especially, on the expected value of the minimum of the divergence time points. Different distributions will give a different covariance structure of the time series $Y_1, Y_2, \ldots, Y_n$.

We have considered a variety of well known distributions in order to study the behavior of the expected value of their minimum which characterizes the nature of the realizations obtained via a LBT. Table 5.2 summarizes the rates of decay for each of the distributions which have been studied in the Appendix. The different

decays are due to the different amount of probability mass around zero for each of the distribution. It can be seen that a very rich class of covariance structures can be obtained which could range from very short to long memory time series depending on the different *"jump distribution"* . A visualization of Table 5.2 is given by the Figure 5.4.

Table 5.2: Rate of decays $O(\cdot)$

"Jump Distribution"	**Density** $f_X(x)$	**Rate of decay** $O(\cdot)$
Uniform, $U(0,1)$	1	$\propto \frac{1}{n}$
Beta$(a,1)$	$a \cdot x^{a-1}$	$\propto \frac{1}{n^{1/a}}$
Bernoulli(p)	$p^x \cdot (1-p)^{1-x}$	$\propto (1-p)^n$
Truncated Exponential(λ)	$\frac{\lambda}{1-e^{-\lambda}} \cdot e^{-\lambda x}$	$\propto -\frac{1}{\lambda}\frac{n+e^{-\lambda}}{n+1}$
Special	$x^{-2}e^{1-\frac{1}{x}}$	$\propto \frac{1}{\log n}$

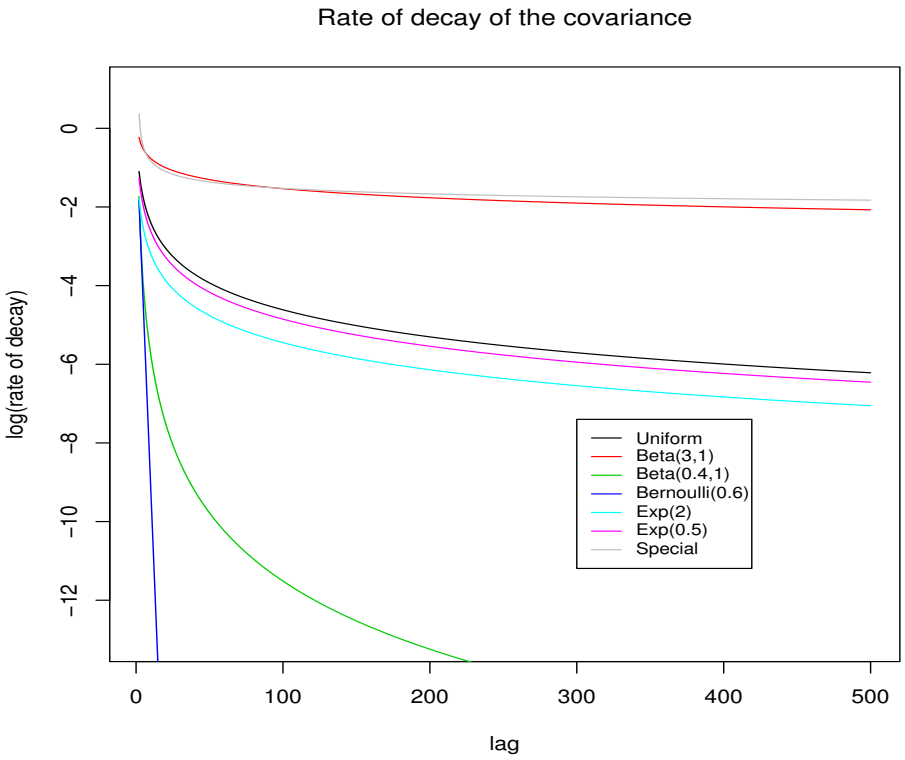

Figure 5.4: Rate of decay of the covariance, $O(\cdot)$

5.5.3 Differences between LBT and DDT

While constructing a LBT, each of the diffusions $X_i(t)$, $i = 1, \ldots, n$ follows the same path only of the just previously constructed diffusion $X_{i-1}(t)$. On the other hand while building a DDT, the diffusion $X_i(t)$, $i = 1, \ldots, n$ can follow paths of any of the previously generated diffusions $X_j(t)$, $j = 1, \ldots, i-1$. This states a distinct difference between the two constructions.

We should also mention the differences between the *"jump distribution"* of the LBT and the *"divergence function"* of the DDT. The former must satisfy the condition that it is a well defined distribution and it is also assumed for convenience that we are able to simulate it. On the other hand, Neal (2003, Sec. 3) states that the DDT will produce prior which are continuous (w.p. 1) when the *"divergence function"* $a(t)$ is such that, the $A(t) = \int_0^1 a(t)$ is infinite.

The divergence points $\tau_i, i = 1, \ldots, n-1$ are independent and identically distributed random variables which follow a specified distribution F truncated to the interval $[0, 1]$. F is known as the *"jump distribution"* and its associated parameters (if any exist) as *"jump parameters"* , whilst Neal (2003) calls them *"divergence function"* and *"divergence parameters"* respectively.

5.6 Illustrative Datasets Generated from an LTB

In this section various datasets are simulated via the LBT construction and the nature of their realisations as well as their correlation structure is examined. For illustration, we consider the following *"jump distributions"* :

1. Beta$(1, 20)$;
2. Beta$(20, 1)$;
3. A mixture of these two Beta distributions.

We generate $10,000$ values using the first two distributions Beta$(1, 20)$ and Beta$(20, 1)$. The difference between these two distributions relies on their amount of probability mass around zero (see left plot of Figure 5.5). This is what characterizes the behavior of the correlation structure of the obtained time series (see right plot of Figure 5.5). Beta(20,1) has much more probability mass around one and therefore the chosen diffusions tend to diverge mostly around values close to one which results in a time series with long memory. On the other hand, if we use a Beta(1, 20) as *"jump distribution"* then the obtained realizations are almost i.i.d due to the fact that most of the divergence time points $(\tau_1, \ldots, \tau_{n-1})$ are concentrated around values less than 0.1. For illustration, Figure 5.6 shows the first 1,000 observations of the time series generated by these two *"jump distributions"* .

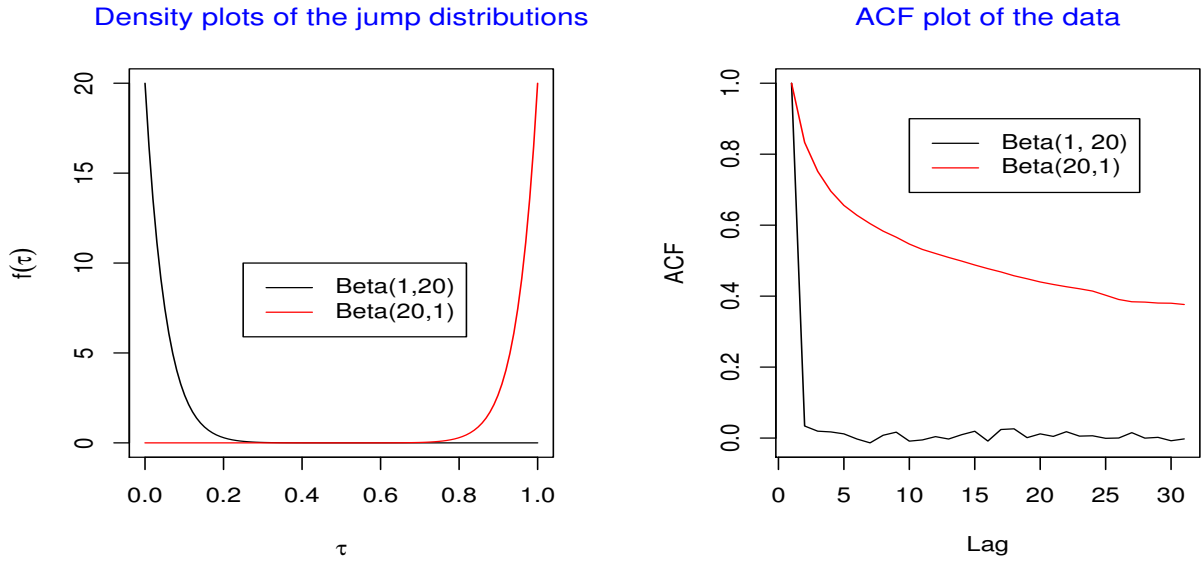

Figure 5.5: Density plots of the Beta distributions (left) and ACF plots for the realizations (right).

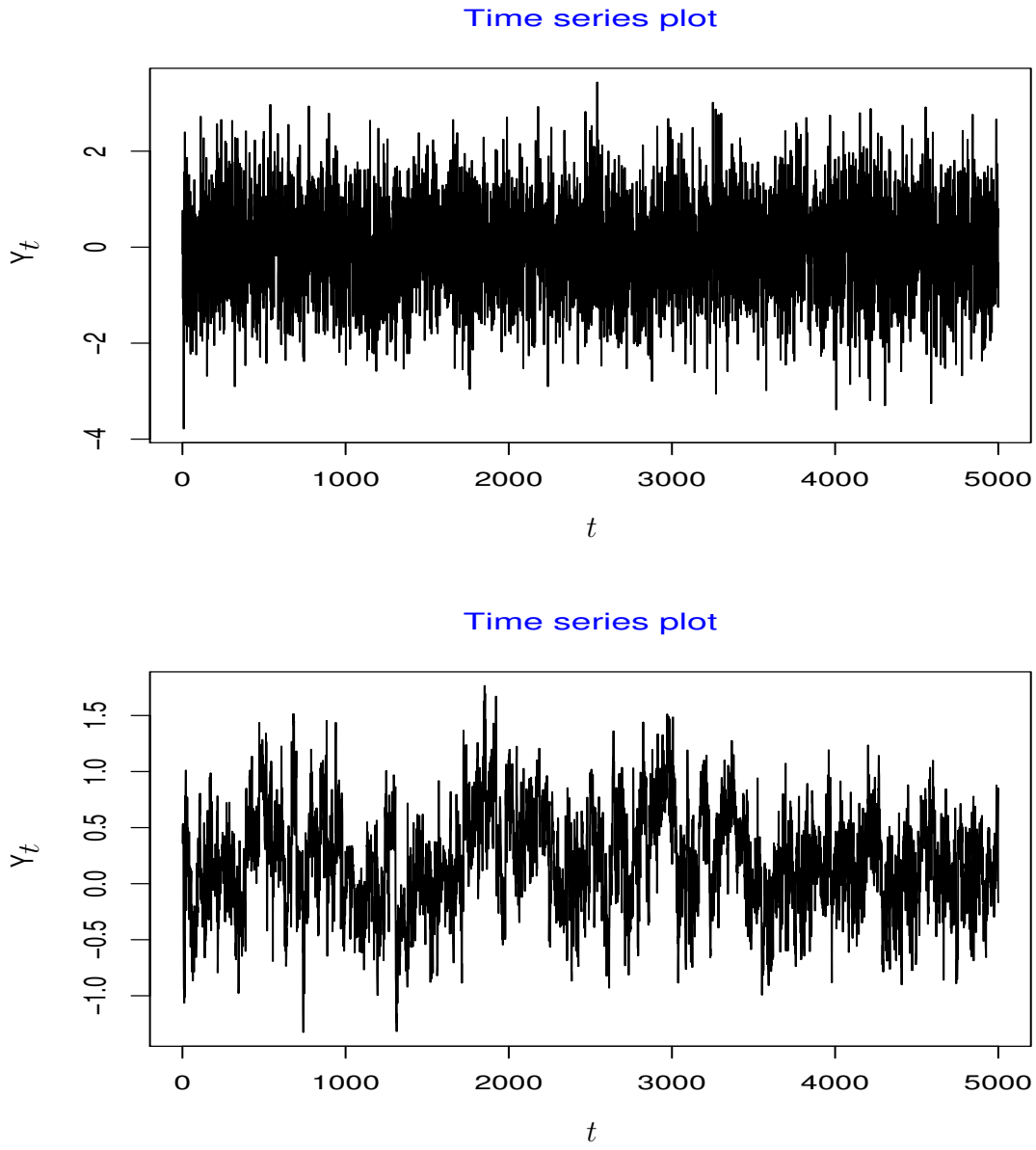

Figure 5.6: The first 5,000 realizations obtained via a LBT using Beta(1,20) (top) and Beta(20,1) (bottom) as *"jump distributions"*

The results obtained so far from these two *"jump distributions"* indicate that a mixture of these two distributions could lead to an interesting nature of realizations. Let us consider the following *"jump distribution"* :

$$f(\tau) = \begin{cases} \text{Beta}(1,20), & \text{with probability } 1-p \\ \text{Beta}(20,1), & \text{with probability } p \end{cases}$$

The actual value of p will specify the structure of the LBT. If p is chosen to be very small, then most of the diffusions will diverge at values of t close to one and much more less times will diverge at values close to zero. The occurrence of small values of the divergence time points τ_i's will "refresh" the tree's memory while the consecutive large values will possibly create clusters of observations due to the result of Theorem 1. Figure 5.7 presents the time series generated using a LBT construction with the above *"jump distribution"* with $p = 0.01$. The small value of p had an effect on the number of clusters created, i.e. $(1 - p) \times N$, where N denotes the number of observations. In our example $N = 5,000$ and $p = 0.01$, therefore we expect to get about 50 clusters.

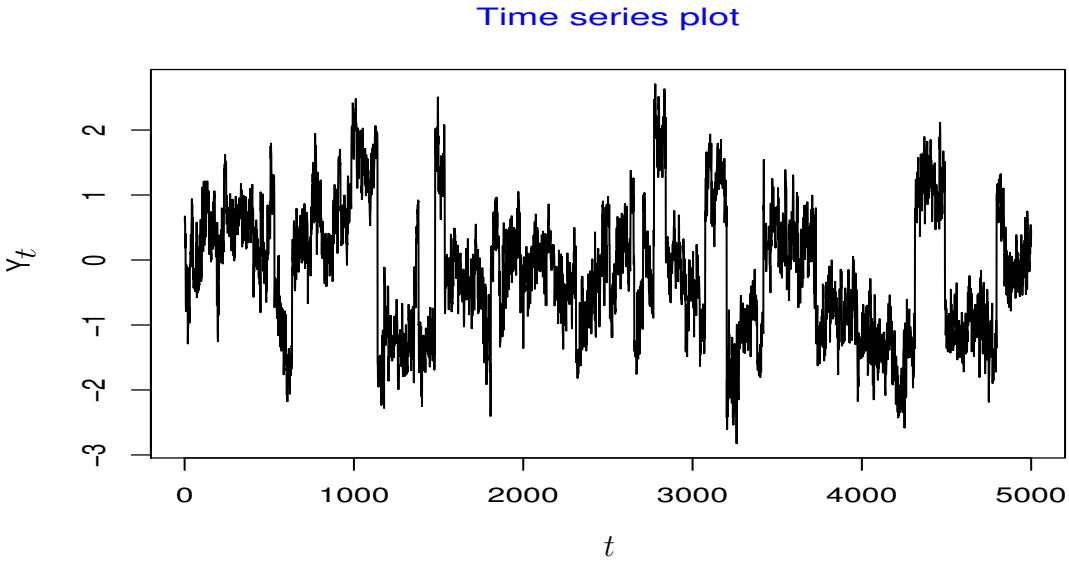

Figure 5.7: An LBT construction using a mixture of Beta distributions as *"jump distribution"*

5.7 Simulation

There exist (at least) two different ways of simulating, n say, data points $Y_1, \ldots, Y_n$, from a latent branching tree; an *approximate* and an *exact*.

The approximate algorithm is fairly easy to implement, because it is mainly based on discretisation of the diffusion paths using Euler approximation. Nevertheless, a significant drawback of such an algorithm is that at each step we need to store the diffusion's path. The more accurate the approximation we want to be, the more costly is to store the path. Therefore, the *approximate* algorithm is not presented, and we propose instead an *exact* algorithm based on *retrospective sampling* techniques (see for example, Beskos et al., 2006).

The main idea behind the exact algorithm relies on the fact that we do not really need to store the full path of each of the diffusions but only their values at the divergence time points $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n-1})$. Because of the property of independent incremements for the Brownian motion we are able to write down the conditional distribution of the value $X(\tau)$ between any two given time points τ_1 and τ_2. Then, we can implement the following algorithm:

Exact Algorithm

1. Draw the $n-1$ divergence points from $t \sim F$;
2. Generate the first data point by simulating from a Normal distribution: $Y_1 \sim N(0,1)$;
3. Draw $X(\tau_1)$ by simulating from a Brownian Bridge between 0 and 1;
4. Generate the second data point (Y_2) by simulating from a Normal distribution: $Y_2 \sim N(X(\tau_1), 1-\tau_1)$;
5. Set $i = 2$, $\tau_0 = 0$ and $\tau_n = 1$;
6. While $(i < n)$ {

   Define the sets $\boldsymbol{L}$ and $\boldsymbol{U}$ as follows:

   - $\boldsymbol{L} = \{j \in [0, i-1] : \tau_j < \tau_i\}$ and $l = \max(\boldsymbol{L})$
   - $\boldsymbol{U} = \{j \in [l+1, i-1] : \tau_j > \tau_l \;\&\; \tau_j > \tau_i\}$. Let $\tau_u = \min(\tau_j)$, $j \in \boldsymbol{U}$.

   i. Simulate the value of the Brownian motion at the i_{th} divergence time point τ_i, $X(\tau_i)$ by drawing from a Brownian bridge between τ_l and τ_u.

   ii. Generate the i_{th} data point by drawing from a Normal distribution: $Y_i \sim N(X(\tau_{i-1}), 1-\tau_{i-1})$;

   iii. $i = i + 1$;

   }

In words the algorithm does the following: i) draw all the divergence time points $(\tau_1, \ldots, \tau_n)$ in advance, ii) draw $Y_1 \sim N(0,1)$, iii) simulate the value of the Brow-

nian motion at the first divergence time point τ_1 by drawing from a conditional Normal distribution (i.e. Brownian bridge in the interval $[0, 1]$), iv) simulate the second data point $Y_2 \sim N(X(\tau_1), 1 - \tau_1)$), v) simulate $X(\tau_2)$ by drawing from a Brownian bridge in the interval $[0, \tau_1]$ or $[\tau_1, 1]$ if $\tau_2 < \tau_1$ or $\tau_2 > \tau_1$ respectively, vi) simulate the third data point $Y_3 \sim N(X(\tau_2), 1 - \tau_2)$). If we repeat the steps (iv-vi) then we can get the desired number of data points.

As already explained, the great advantage of the *exact* algorithm is that is has a much smaller computational cost than the *approximate*. In other words, given the skeleton of a LBT we can fill in the intermediate points with diffusion bridges. Figure 5.8 shows the skeleton of the LBT derived in Section 5.4.

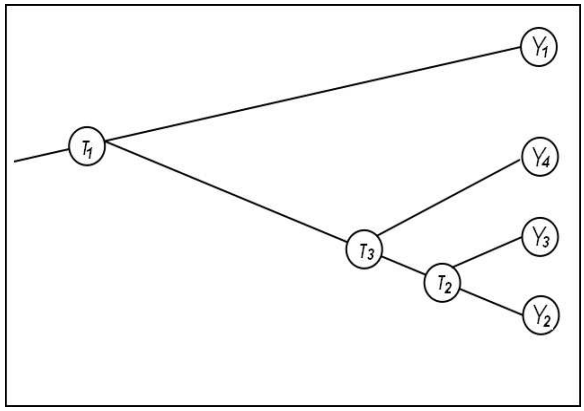

Figure 5.8: A Skeleton of a LBT

5.8 Inference

This section focuses on how to draw inference for the parameters associated with a LBT. In principle, the parameters of interest are the divergence time points $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_{n-1})$. If complete information was available, i.e. the observed data ($\boldsymbol{Y} = Y_1, Y_2, \ldots, Y_n$) and the state of the processes at the divergence time points, $\boldsymbol{X}(\tau) = (X(\tau_1), \ldots, X(\tau_{n-1}))$ then inference for the times $\boldsymbol{\tau}$ would rely on maximization of the likelihood with respect to that vector of unknown parameters. This is usually not an easy task especially when the parameter space is relatively

big.

Furthermore, in practice, the location parameters $\boldsymbol{X}(\tau)$ are also unobserved and therefore inference for them should also be drawn. The maximization becomes even harder in this case. We adopt a fully Bayesian approach where we treat all the unobserved quantities as parameters. Without loss of generality, for illustration we assume that chosen diffusion of the LBT, is a Brownian motion.

The choice of *"jump distribution"* can be viewed as prior for the divergence times, $\boldsymbol{\tau}$ while the parameters associated with it, as *hyperparameter* for which an appropriate (hyper)prior distribution should be specified. Before we show how we make inference for the parameters, we adopt the following notation:

Notation

Let $\phi(x, \mu, \sigma^2)$ be the Gaussian density with mean μ and variance σ^2. Lets denote by $\phi^{BB}(x_b, t_b, t_a, t_c, x_{t_a}, x_{t_c})$ the conditional density of a Brownian bridge in the interval (t_a, t_c) such that $t_a < t_b < t_c$ i.e.:

$$\phi(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{5.2}$$

$$\phi^{BB}(x_b, t_b, t_a, t_c, x_{t_a}, x_{t_c}) = \frac{1}{\sigma_{BB}\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2_{BB}}(x_{t_b}-\mu_{BB})^2\right\} \tag{5.3}$$

where μ_{BB} and σ^2_{BB} are:

$$\mu_{BB} = \frac{x_{t_a}(t_c - t_b) + x_{t_c}(t_b - t_a)}{t_c - t_a} \tag{5.4}$$

$$\sigma^2_{BB} = \frac{(t_c - t_b)(t_b - t_a)}{t_c - t_a} \tag{5.5}$$

Denote by $F(\cdot)$ be the distribution where the τ_i are drawn from and by $f(\cdot)$ the corresponding probability density function. Denote the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)^T$ where: $\boldsymbol{\theta}_1 = (\tau_1, \ldots, \tau_{n-1})^T$ and $\boldsymbol{\theta}_2 = (X(\tau_1), \ldots, X(\tau_{n-1}))^T$, $\boldsymbol{\theta}_3$ the

vector containing the *"jump parameters"* (if any exist) and the data vector by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

We define the following sets:

- $\boldsymbol{L} = \{j \in [0, i-1] : \tau_j < \tau_i\}$
- $\boldsymbol{U} = \{j \in [l+1, i-1] : \tau_j > \tau_l \ \& \ \tau_j > \tau_i\}$.

and denote by $l = \max(\boldsymbol{L})$. Also, we set $\tau_u = \min(\tau_j)$, $j \in \boldsymbol{U}$.

5.8.1 Likelihood

The probability of obtaining a latent branching tree can be calculated once i) the divergence points $(\tau_i, i = 1, \ldots, n-1)$, ii) their locations $X(\tau_i)$, i.e. the state of the process at each time, iii) the data points $Y_i, i = 1, \ldots, n$ are known and the order of these events as well. The likelihood can be expressed as product of two factors, one describing the tree (L_1) and the other the data (L_2). The first factor is the probability of obtaining the (ordered) divergence points (τ_i) and the second is the probability of obtaining the tree given the locations of the divergence points ($X(\tau_i)$) and the final data points (Y_i) given the divergence times.

The likelihood can be expressed in terms of two products as follows:

$$\begin{aligned} f(\boldsymbol{Y}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) &= \prod_{i=1}^{n} f(\tau_i) \times \phi(Y_1, 0, 1) \\ &\times \prod_{i=1}^{n-1} \left(\phi^{BB} \ X(\tau_i), \tau_i, \tau_l, \tau_u, X(\tau_l), X(\tau_u) \cdot \phi(Y_{i+1}, X(\tau_i), 1 - \tau_i) \right) \end{aligned} \tag{5.6}$$

It easy to see that $L_1 = \prod_{i=1}^{n} f(\tau_i)$ and L_2 is a product of normal densities. Since the procedure described in the previous section generates a sequence of n data points, the factor L_2 is most easily obtained in the same way. The appearance of the normal products is due to the choice of the stochastic process, i.e. Brownian

motion. If instead we use a Gamma process, the distributions $\phi^{BB}(\cdot)$ for the intermediate points are substituted by Gamma densities.

Recall, the example presented in Section 5.4. The likelihood of observing the data Y_1, Y_2, Y_3, Y_4 with this order, given the the divergence time points τ_1, τ_2, τ_3 and the corresponding states of the processes, $X(\tau_1), X(\tau_2), X(\tau_3)$ can be derived as follows:

$$
\begin{aligned}
L_1 &= \prod_{i=1}^{4} f(\tau_i) && (5.7)\\
L_2 &= \phi(Y_1, 0, 1) \\
&\times \phi^{BB}(X(\tau_1), \tau_1, 0, 1, 0, Y_1) \times \phi(Y_2, X(\tau_1), 1-\tau_1) \\
&\times \phi^{BB}(X(\tau_2), \tau_2, \tau_1, 1, X(\tau_1), Y_2) \times \phi(Y_3, X(\tau_2), 1-\tau_2) \\
&\times \phi^{BB}(X(\tau_3), 0, \tau_1, 0, X(\tau_1)) \times \phi(Y_4, X(\tau_3), 1-\tau_3). && (5.8)\\
f(\boldsymbol{X}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) &= L_1 \times L_2 && (5.9)
\end{aligned}
$$

An alternative way of rewriting L_2 is the following:

$$
\begin{aligned}
L_2 &= \phi(X(\tau_1), 0, \tau_1) \times \phi(Y_1, X(\tau_1), 1-\tau_1) \\
&\times \phi(X(\tau_3), X(\tau_1), \tau_3 - \tau_1) \times \phi(Y_4, X(\tau_3), 1-\tau_3) \\
&\times \phi(X(\tau_2), X(\tau_3), \tau_2 - \tau_3) \times \phi(Y_3, X(\tau_2), 1-\tau_2) \\
&\times \phi(Y_2, X(\tau_2), 1-\tau_2)
\end{aligned}
$$

The above expression shows that the likelihood can be expressed as the product of the densities of all the segments of the tree. The density of each segment depends on the choice of the diffusion. For instance, in this case where a Brownian motion is used, the segments are Normally distributed. On the other hand, note that L_1 depends only on the *"jump distribution"* and is independent of of the vector $\boldsymbol{X}(\tau)$.

5.8.2 Posterior Distribution

The likelihood and the prior are combined using Bayes theorem and we get up to proportionality the full posterior distribution of the parameters which has the following density:

$$\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3 | \boldsymbol{Y}) \propto L_1 \times L_2 \times \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \tag{5.10}$$

where $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ can also be written as $\pi(\boldsymbol{\theta}_1) \times \pi(\boldsymbol{\theta}_2) \times \pi(\boldsymbol{\theta}_3 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Markov Chain Monte Carlo (MCMC) methods need to be applied in order to draw samples from the full posterior distribution of the parameters. The following algorithm can be implemented:

MCMC Algorithm - Centered I

(Repeat the following steps)

1. Start the chain with initial values $\tau_1^0, \ldots, \tau_n^0$, $X^0(\tau_1), \ldots, X^0(\tau_n)$, $\boldsymbol{\theta}_3^0$;
2. Choose one (or more) of the divergence parameters j, $1 \leq j \leq n-1$ and update τ_j (individually) using Metropolis Hastings;
3. Update each of the location parameters $X(\tau_j)$, $1 \leq j \leq n-1$ using Gibbs sampler;
4. Update $\boldsymbol{\theta}_3$ using either Metropolis Hastings or Gibbs sampler

We discuss the various issues regarding the implementation of the above centered MCMC algorithm. For illustration we derive full conditional distributions of some of the parameters involved in the example described in Section 5.4.

- **Update the divergence time points ($\boldsymbol{\theta}_1$).** The full conditional distribution of $[\tau_i|\boldsymbol{\tau}_{-i}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3]$ is not of standard form and therefore a Metropolis Hastings algorithm is needed. Apart from a random walk Metropolis (RWM), an independence sampler can be used. Optimal proposals can be chosen according to the prior assumed for the times $\boldsymbol{\tau}$. If for instance, a uniform prior is chosen then a Uniform(0,1) can be applied. On the other hand, if a Beta distribution is assumed then this Beta distribution can be used as proposal.

  The conditional distribution of τ_1 given the rest is given as follows:

$$\begin{aligned}
\pi(\tau_1|\ldots) &\propto \phi(X(\tau_1), 0, \tau_1) \times \phi(Y_1, X(\tau_1), 1-\tau_1) \times \phi(X(\tau_3), X(\tau_1), \tau_3 - \tau_1) \\
&\propto \frac{1}{\sqrt{\tau_1}} \exp\left\{-\frac{1}{2}\frac{(X(\tau_1))^2}{\tau_1}\right\} \times \frac{1}{\sqrt{1-\tau_1}} \exp\left\{-\frac{1}{2}\frac{(X(\tau_1)-Y_1)^2}{1-\tau_1}\right\} \\
&\times \frac{1}{\sqrt{\tau_3-\tau_1}} \exp\left\{-\frac{1}{2}\frac{(X(\tau_3)-X(\tau_1))^2}{\tau_3-\tau_1}\right\}
\end{aligned} \tag{5.11}$$

- **Update the location parameters ($\boldsymbol{\theta}_2$).** Because of the choice of the Brownian motion as the driving diffusion of the LBT, the conditional distribution of each of the location parameters $X(\tau_i)$ given the rest, $[\boldsymbol{X}(-\tau_i), \boldsymbol{\tau}, \boldsymbol{\theta}_3]$, is Normally distributed with mean μ_i and variance σ_i^2. It is easy to check that μ_i is a linear combination of the rest $\boldsymbol{X}(-\tau_i)$, and actually it only depends on the the three adjacent states $X(\tau_j), j \sim i$ of $X(\tau_i)$.

  We derive the full conditional distribution of $[X(\tau_1)|\boldsymbol{Y}, \boldsymbol{\tau}, X(\tau_2), X(\tau_3)]$:

$$\pi(X(\tau_1)|\ldots) \propto \phi(X(\tau_1), 0, \tau_1) \times \phi(Y_1, X(\tau_1), 1-\tau_1) \times \phi(X(\tau_3), X(\tau_1), \tau_3 - \tau_1)$$

$$\propto \exp\left\{-\frac{1}{2}\frac{(X(\tau_1))^2}{\tau_1}\right\} \times \exp\left\{-\frac{1}{2}\frac{(X(\tau_1)-Y_1)^2}{1-\tau_1}\right\}$$

$$\times \exp\left\{-\frac{1}{2}\frac{(X(\tau_3)-X(\tau_1))^2}{\tau_3-\tau_1}\right\}$$

$$\pi(X(\tau_1)|\ldots) \equiv N(\mu_1, \sigma_1^2) \tag{5.12}$$

where

$$\mu_1 = \left(\frac{Y_1}{1-\tau_1} + \frac{X(\tau_3)}{\tau_3-\tau_1}\right)\left(\frac{1}{\tau_1} + \frac{1}{1-\tau_1} + \frac{1}{\tau_3-\tau_1}\right)^{-1}$$

and

$$\sigma_1^2 = \left(\frac{1}{\tau_1} + \frac{1}{1-\tau_1} + \frac{1}{\tau_3-\tau_1}\right)^{-1}$$

- **Update the jump parameters** ($\boldsymbol{\theta}_3$)**.** A Gibbs step or a Metropolis may be needed in order to update θ_3 depending on the form of the *"jump distribution"* and also on its prior. For the example we are looking at, such parameter does not exist.

5.9 MCMC Strategies

It might be the case that the standard MCMC algorithm described in the previous section does not offer a well mixing Markov chain due to various problems. For instance, dependence between the missing data $(\boldsymbol{\tau}, \boldsymbol{X}(\tau))$ and the model parameters $\boldsymbol{\theta}_3$. In this section we describe alternative MCMC algorithms and an alternative parameterisation which can be applied in order to draw samples from the posterior distribution of the parameters given the observed data.

5.9.1 Block Update of Location Parameters ($\boldsymbol{\theta}_2$)

The 3rd step of the standard MCMC algorithm states that we can apply a Gibbs sampler to update sequentially (*deterministic scan*) the location parameters $\boldsymbol{X}(\tau_i)$, $i = 1, \ldots, n-1$. Apart from this approach, we are able to update the location parameters as a block by drawing from a multivariate Normal distribution. First we show how we can draw from such a distribution and then we discuss when such an alternative approach is of any practical benefit.

Consider the vector $\boldsymbol{Z} = (\boldsymbol{X}(\tau), \boldsymbol{Y})^T = (X(\tau_1), X(\tau_2), \ldots, X(\tau_{n-1}), Y_1, \ldots, Y_n)^T$ of length $2n-1$. Since:

$$\pi(\boldsymbol{X}(\tau), \boldsymbol{Y}|\boldsymbol{\tau}) \sim N\left(0, Q^{-1}\right) \tag{5.13}$$

where Q can expressed as follows:

$$Q = \begin{bmatrix} Q_{xx} & Q_{xy} \\ Q_{xx} & Q_{xy} \end{bmatrix}$$

It is easy to construct the matrix Q. For its off-diagonal elements, q_{ij}, where $i \neq j$:

$$q_{ij} = \begin{cases} 0, & \text{if } Z_i \text{ and } Z_j \text{ are not connected} \\ -\frac{1}{d_{ij}}, & \text{if } Z_i \text{ and } Z_j \text{ are connected} \end{cases} \tag{5.14}$$

where d_{ij} denotes the length of the branch of the tree which involves the elements of $\boldsymbol{Z}$. The diagonal elements are given from the following form:

$$q_{ii} = -\sum_{j=1, j\neq i}^{2n-1} q_{ij} + \frac{1}{\tau_i} \times \mathbf{1}_{\tau_i} \tag{5.15}$$

where

$$\mathbf{1}_{\tau_i} = \begin{cases} 1 \quad \text{if} \quad \tau_i = \min\left(\tau_j, j = 1, \ldots, n-1\right) \\ \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases}$$

Note that the diagonal elements are equal to the inverse of the conditional variance of each of the components:

$$q_{ii} = (\text{var}(Z_i|\boldsymbol{Z}_{-i}))^{-1} = (\text{var}(X(\tau_i)|\boldsymbol{X}_{-\tau_i}, \boldsymbol{\tau}, \boldsymbol{Y}))^{-1} \tag{5.16}$$

Once the matrix Q is obtained, then is relatively easy to obtain the conditional distribution of $\boldsymbol{X}(\tau)$ given the observed data ($\boldsymbol{Y}$) and the divergence times $\boldsymbol{\tau}$ (Chatfield and Collins, 1980).

$$\pi(\boldsymbol{X}(\boldsymbol{\tau})|\boldsymbol{Y}, \boldsymbol{\tau}) \sim N\left(-Q_{xx} \cdot Q_{xy} \cdot \boldsymbol{Y}, Q_{xx}^{-1}\right) \tag{5.17}$$

Therefore, Equation (5.17) allows us to update the location parameters as a block by drawing from an $n-1$ dimensional Normal distribution and we can then implement the following algorithm:

MCMC Algorithm - Centered II

(Repeat the following steps)

1. Start the chain with initial values $\tau_1^0, \ldots, \tau_n^0$, $X^0(\tau_1), \ldots, X^0(\tau_n)$, $\boldsymbol{\theta}_3^0$;
2. Choose one (or more) of the divergence parameters j, $1 \leq j \leq n-1$ and update τ_j (individually) using Metropolis Hastings;
3. Update location parameters $X(\tau_j)$, $1 \leq j \leq n-1$ as a block by drawing from a multivariate Normal Distribution using Equation (5.17);
4. Update $\boldsymbol{\theta}_3$ using either Metropolis Hastings or Gibbs sampler depending on the parameter.

The question which remains to be answered is whether block updating of the location parameters can improve the mixing or not. Roberts and Sahu (1997) derive rates of convergence of the Gibbs sampler for Gaussian target distributions. The authors discuss various updating strategies such as *deterministic scan* and block schemes and they also study the effect of dimensionality and correlation structure on the convergence rates using different schemes. The following theorem is taken from Roberts and Sahu (1997):

Theorem 2 *If all partial correlations of a Gaussian target density are non-negative, i.e. all the off diagonal-elements of Q are non-positive, then the block update scheme has faster rate of convergence than deterministic scan.*

In other words, this theorem states that if $q_{ij} < 0$ for all $i \neq j$, then block update of the location parameters can only improve mixing. Within a LBT framework,

the form of Equation (5.14) assures that all the off-diagonal elements of matrix Q_{xx} are non-positives since d_{ij} is defined to be strictly positive. Therefore, this suggests that block update of the missing data is worthwhile.

The matrix Q can be obtained for example presented in Section 5.4:

$$Q_{xx} = \begin{bmatrix} q_{11} & 0 & -1/(\tau_3 - \tau_1) \\ 0 & q_{22} & -1/(\tau_2 - \tau_3) \\ -1/(\tau_3 - \tau_1) & -1/(\tau_2 - \tau_3) & q_{33} \end{bmatrix}$$

where

- $q_{11} = 1/\tau_1 + 1/(\tau_3 - \tau_1) + 1/(1 - \tau_1)$,
- $q_{22} = 1/(\tau_2 - \tau_3) + 1/(1 - \tau_2) + 1/(1 - \tau_2)$,
- $q_{33} = 1/(\tau_3 - \tau_1) + 1/(\tau_2 - \tau_3) - 1/(1 - \tau_3)$.

$$Q_{xy} = \begin{bmatrix} -1/(1-\tau_1) & 0 & 0 & 0 \\ 0 & -1/(1-\tau_2) & -1/(1-\tau_2) & 0 \\ 0 & 0 & 0 & -1/(1-\tau_3) \end{bmatrix}$$

$$Q_{yx} = \begin{bmatrix} -1/(1-\tau_1) & 0 & 0 \\ 0 & -1/(1-\tau_2) & 0 \\ 0 & -1/(1-\tau_2) & 0 \\ 0 & 0 & -1/(1-\tau_3) \end{bmatrix}$$

$$Q_{yy} = \begin{bmatrix} 1/(1-\tau_1) & 0 & 0 & 0 \\ 0 & 1/(1-\tau_2) & 0 & 0 \\ 0 & 0 & 1/(1-\tau_2) & 0 \\ 0 & 0 & 0 & 1/(1-\tau_3) \end{bmatrix}$$

5.9.2 Integrate the Location Parameters Out ($\boldsymbol{\theta}_3$)

Recall that the joint distribution $\pi(\boldsymbol{X}(\tau), \boldsymbol{Y}|\boldsymbol{\tau})$ gives us the ability to write explicitly the marginal distribution of $\boldsymbol{Y}|\boldsymbol{\tau}$ by integrating out the vector $\boldsymbol{X}(\tau)$:

$$\pi(\boldsymbol{\tau}, \boldsymbol{\theta}_3|\boldsymbol{Y}) = \int_{\boldsymbol{X}(\tau)} \pi(\boldsymbol{\tau}, \boldsymbol{X}(\tau), \boldsymbol{\theta}_3|\boldsymbol{Y}) \, \mathrm{d}\boldsymbol{X}(\boldsymbol{\tau}) \tag{5.18}$$

Using matrix calculation and results from multivariate analysis (see for example, Chatfield and Collins, 1980):

$$\pi(\boldsymbol{Y}|\boldsymbol{\tau}, \boldsymbol{\theta}_3) \sim N\left(\mathbf{0}, \left(Q_{yy} - Q_{yx} \cdot Q_{xx}^{-1} \cdot Q_{xy}\right)^{-1}\right) \tag{5.19}$$

Having obtained the likelihood having location parameters integrated out($\pi(\boldsymbol{Y}|\boldsymbol{\tau}, \boldsymbol{\theta}_3)$), it is easy then to obtain the marginal posterior distribution distribution of $\boldsymbol{\tau}, \boldsymbol{\theta}_3|\boldsymbol{Y}$:

$$\pi(\boldsymbol{\tau}, \boldsymbol{\theta}_3|\boldsymbol{Y}) \propto \pi(\boldsymbol{Y}|\boldsymbol{\tau}, \boldsymbol{\theta}_3) \times \pi(\boldsymbol{\tau}) \times \pi(\boldsymbol{\theta}_3|\boldsymbol{\tau}) \tag{5.20}$$

We can then implement the following centered algorithm having as target distribution $\pi(\boldsymbol{\tau}, \boldsymbol{\theta}_3|\boldsymbol{Y})$.

MCMC Algorithm - Centered III

(Repeat the following steps)

1. Start the chain with initial values $\tau_1^0, \ldots, \tau_n^0$, $\boldsymbol{\theta}_3^0$;
2. Choose one (or more) of the divergence parameters j, $1 \leq j \leq n-1$ and update τ_j (individually) using Metropolis Hastings;
3. Update $\boldsymbol{\theta}_3$ using either Metropolis Hastings or Gibbs sampler depending on the parameter.

5.9.3 Efficient Non-Centered Parameterisations

The dependence between the missing data and the model parameters often can cause problems with the mixing of the Markov chains. Within a LBT framework, once the prior of the divergence times involves some unknown parameters to be estimated, i.e. $\boldsymbol{\theta}_3$ then *a priori* correlation between $\boldsymbol{\tau}$ and $\boldsymbol{\theta}_3$ is induced. Such problems can be overcome by applying a non−centered parameterisation (Papaspiliopoulos, 2003, Papaspiliopoulos et al., 2003) (see also Section 1.9) where we can break the dependence link between missing data and model parameters by introducing a reparameterisation which makes the missing data and the model parameters *a priori* independent (see Figure 5.9).

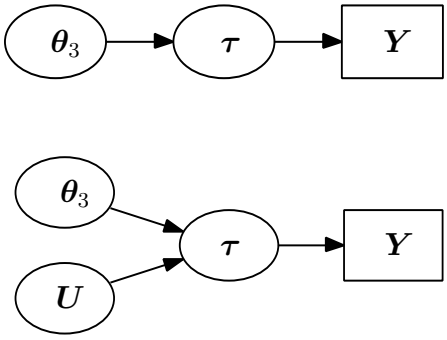

Figure 5.9: Graphical model of the centered (top) and and non−centered hierarchical parameterisation of the model

The break of that dependence link can be done by introducing a change in variables from

$$(\boldsymbol{\tau}, \boldsymbol{X}(\tau), \boldsymbol{\theta}_3) \rightarrow (\boldsymbol{U}, \boldsymbol{X}(\tau), \boldsymbol{\theta}_3)$$

where $U_i = g(\tau_i, \theta_3)$, $i = 1, \ldots, n-1$ and $g(\cdot)$ is a deterministic function. In principle, any one to one function can be used, however an obvious choice is to derive a function using the *inverse cumulative distribution function* (Inverse CDF) method by assigning

$$U_i = F(\tau_i | \boldsymbol{\theta}_3)$$

where $F(\cdot)$ denotes the distribution function of the random variables τ_i, $i = 1, \ldots, n-1$. This leads to the fact that *a priori* each of the transformed variables is uniformly distributed in $[0, 1]$:

$$U_i \sim U(0, 1), \quad i = 1, \ldots, n-1.$$

Once we adopt a NCP we can implement the following MCMC algorithm:

Non-Centered (NC) MCMC algorithm

(Repeat the following steps)

1. `Start the chain with initial values for the parameters:` $\boldsymbol{\tau}^0, \boldsymbol{X}(\tau)^0, {\boldsymbol{\theta}_3}^0$;
2. `Update` $\boldsymbol{X}(\tau)$ `either by deterministic scan, or a block update scheme`;
3. `Obtain a sample of` $(\boldsymbol{\tau}^1, {\boldsymbol{\theta}_3}^1)$ `via a centered algorithm`;
4. `Get` $U_i^1 = g\left(\tau_i^1, {\boldsymbol{\theta}_3}^1\right)$;
5. `Update` $\boldsymbol{\theta}_3$ `using a Metropolis Hastings step by proposing` ${\boldsymbol{\theta}_3}^2 \sim q\left({\boldsymbol{\theta}_3}^1\right)$ `and accept it with probability :`

$$1 \wedge \frac{\pi({\boldsymbol{\theta}_3}^2|\boldsymbol{U}, \boldsymbol{X}(U))}{\pi({\boldsymbol{\theta}_3}^1|\boldsymbol{U}, \boldsymbol{X}(U))} \frac{q({\boldsymbol{\theta}_3}^1, {\boldsymbol{\theta}_3}^2)}{q({\boldsymbol{\theta}_3}^2, {\boldsymbol{\theta}_3}^1)};$$

[Note that by updating $\boldsymbol{\theta}_3$, the divergence times τ_i's are updated as well since ${\tau_i}^2 = g^{-1}(U_i^1, {\boldsymbol{\theta}_3}^2), i = 1, \ldots, n$]

where $q(\cdot)$ denotes the proposal distribution for the Metropolis step of the parameter $\boldsymbol{\theta}_3$. Now we discuss possible proposal distributions which can be used. Denote by ${\boldsymbol{\theta}_3}^c$ the current value and by ${\boldsymbol{\theta}_3}'$ the proposed value.

1. **Random Walk Metropolis**:

$${\boldsymbol{\theta}_3}' \sim N({\boldsymbol{\theta}_3}^c, \sigma_\theta^2)$$

   Depending on the sign of the parameters of interest, a multiplicative random Walk can also be applied.

2. **"Pseudo-Gibbs" Sampler**: Suppose that the full conditional distribution

of $\boldsymbol{\theta}_3$ is of a standard (closed) form. We can then use this distribution as a proposal:

$$\boldsymbol{\theta}_3{}' \sim \pi(\boldsymbol{\theta}_3{}'|\boldsymbol{\tau}^c, \boldsymbol{X}(\tau)^c)$$

3. **Adaptive Sampling**. If we run one of the centered algorithms for an adequate number of observations it is possible that a good approximation of the marginal posterior distribution of the *"jump parameters"* can be derived. If we are able approximate this distribution with one of a standard form (i.e. Gamma, Normal or others), then such a distribution can be used as a proposal.

   When such proposal is used, much care should be taken. If the approximation is bad, then this will lead to an inefficient proposal. It might also be the case that the proposal distribution will have lighter tails than the target and therefore we will not explore the tails of the target distribution appropriately. It is then always better to allow for a proposal with heavier tails even if this causes reduction to the acceptance probability of the independence sampler.

4. **Normal Approximation** If the divergence times are known, then it may be relatively easy to derive maximum likelihood estimators (MLEs) for the parameters $\boldsymbol{\theta}_3$ by differentiation of the likelihood. Denote by $\widehat{\boldsymbol{\theta}}_3$ and $\widehat{\sigma}(\boldsymbol{\theta}_3)$ the MLEs and their corresponding standard errors. We can use the following proposal:

$$\boldsymbol{\theta}_3{}' \sim N\left(\widehat{\boldsymbol{\theta}}_3, \epsilon\widehat{\sigma}(\boldsymbol{\theta}_3)^2\right)$$

   for some $\epsilon > 1$ so as to allow the proposal to do big jumps and avoid similar problems regarding the tails of the target distribution as mentioned in the previous proposal.

Apart from a fully non−centered parameterisation (NCP) it is also possible to construct a partially non−centered parameterisation (PNCP) where a percentage

of the missing data is parameterized as centered and the rest as non−centered. Papaspiliopoulos et al. (2003) conclude that when the missing data are more informative about the model parameter, then a NCP is preferred than a CP. Therefore, an algorithm which will adjust for the proportion of the information presented in the data and "switching" between a CP and NCP, could out perform the CP or the NCP. Such a parameterisation is called "partially non−centered" and be implemented as follows:

The set of the divergence times $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n-1})$ is partitioned in two groups: $\mathcal{C}$ and $\mathcal{U}$. Let $\boldsymbol{\tau}^{\mathcal{C}}$ and $\boldsymbol{\tau}^{\mathcal{U}}$ denote the divergence times which belong in the groups $\mathcal{C}$ and $\mathcal{U}$ respectively. For the times which belong to $\mathcal{U}$, let

$$U_j = g(\tau_j, \boldsymbol{\theta}_3), \quad j \in \mathcal{U}.$$

In other words we propose a change in variable

$$(\boldsymbol{\tau}, \boldsymbol{X}(\tau), \boldsymbol{\theta}_3) \rightarrow (\tau_i, \boldsymbol{X}(\tau), U_j, \boldsymbol{\theta}_3) \text{ for } i \in \mathcal{C}, j \in \mathcal{U}$$

It easy to see that if $\mathcal{U} = \emptyset$, then we get the centered parameterisation. Furthermore, if for $1 \leq i \leq n-1$ we let:

$$Z_i = \begin{cases} 1 & \text{with probability } \mu_i \\ 0 & \text{with probability } 1 - \mu_i \end{cases} \tag{5.21}$$

then set $\mathcal{C} = \{i : Z_i = 1\}$ and $\mathcal{U} = \{i : Z_i = 0\}$. Once the posterior distribution of $\boldsymbol{\theta}_3$ is derived by taking into account the Jacobian for the transformation of the divergence times points. Then following PNCP algorithm can be implemented:

Partially Non−Centered MCMC Algorithm

(Repeat the following steps)

1. `Start the chain with initial values:` $\boldsymbol{\tau}^0, \boldsymbol{X}(\tau)^0, \boldsymbol{\theta}_3$;
2. `Update` $\boldsymbol{X}(\tau)^1$ `using either deterministic scan or block update;`
3. `Obtain a sample of` $(\boldsymbol{\tau}^1, \boldsymbol{\theta}_3{}^1)$ `using a centered algorithm ;`
4. `Update` $\boldsymbol{Z}$ `and hence` $\mathcal{C}$ `and` $\mathcal{U}$;
5. `Get` $U_j^1 = g\left(\tau_j^1, \boldsymbol{\theta}_3{}^1\right),\ j \in \mathcal{U}$;
6. `Update` $\boldsymbol{\theta}_3$ `using Metropolis Hastings Algorithm.`

[Note that by updating $\boldsymbol{\theta}_3$, the divergence times τ_j's are updated as well since ${\tau_j}^2 = g^{-1}(U_i^1, \boldsymbol{\theta}_3{}^2), j \in \mathcal{U}$]

Apart from the presented PNCP, other exist as well. Following Papaspiliopoulos et al. (2003) instead of non−centering some of the divergence times and some not, we can partially non−center each of the times τ_i. An example of such a PNCP is given in Section 5.10. Section 5.10 considers various examples where such reparameterisations are applied.

5.10 Applications on Simulated Data Sets

We are interested in assessing the performance of models obtained via a LBT construction and also the efficiency of the different MCMC strategies presented in Section 5.9. Therefore, in this section, we will simulate some datasets using the *exact* algorithm (see Section 5.7) considering different *"jump distributions"* (JD). Then we draw inference for the parameters of interest, such as the divergence

times τ_i, $i = 1, \ldots, n-1$ and also for the model parameters associated with them. Comparisons between the different MCMC strategies are also given.

In order to assess model's efficiency to capture the *true "jump distribution"* we look at the *empirical cumulative distribution function* (ECDF) of the resultant samples of the posterior distribution of $\boldsymbol{\tau}$ and compare it with the *true* CDF. In addition, another measure of efficiency is to compare the average divergence time $\overline{\tau}$ obtained from the posterior samples with the *true* one. The average divergence time can be calculated as follows:

Denote by K the number of samples obtained via MCMC and by $\boldsymbol{t}_i$ the posterior sample of the ith divergence time. Then we easily obtain:

$$\overline{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{K} \sum_{j=1}^{K} t_{ij} \right) \tag{5.22}$$

5.10.1 JD: Uniform

We simulate a dataset which consists of $n = 200$ points where the divergence times points $\tau_1, \ldots, \tau_{n-1}$ are uniformly distributed. Such a *"jump distribution"* does not involve any parameters, and therefore we are interested in the posterior distribution of the times given the observed data, i.e. $\pi(\tau_i|\boldsymbol{Y})$, $i = 1, \ldots, n-1$. We apply the standard centered algorithm (see Section 5.8.2). The Metropolis Hastings step needed to update the times is performed by using an independence sampler having a Uniform(0,1) as proposal distribution.

Figure 5.10 shows a posterior sample of the obtained CDF of the divergence time points $\boldsymbol{\tau}$ and reveals a pretty good fit. The average divergence times is equal to 0.497, close to the *true* value (equal to 0.5).

Figure 5.11 shows the posterior distribution for three different divergence times, $\tau_{12}, \tau_{20}, \tau_{22}$. Recall that in general, τ_i refers to the time at which the $(i+1)$th diffusion has diverged in order to obtain the data point Y_{i+1}. The actual observations related to these three time points we are focusing at are given below:

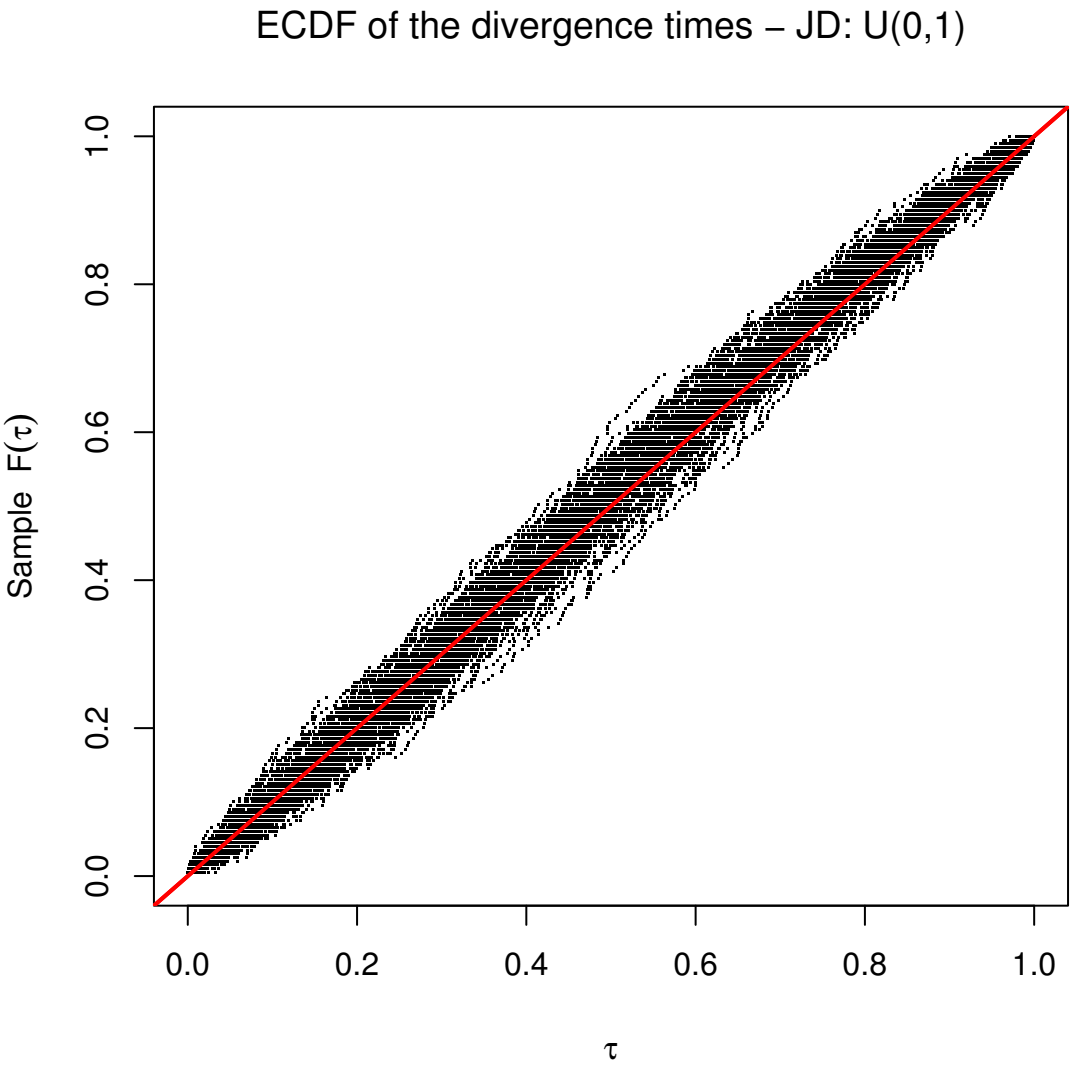

Figure 5.10: ECDF of times for the Uniform *"jump distribution"* - Red line shows the true CDF

$$(Y_{12}, Y_{13}, Y_{20}, Y_{21}, Y_{22}, Y_{23}) = (-2.043, -1.067, -1.607, -1.683, -2.032, 1.056)$$

Intuitively, we would expect for consecutive observations whose difference $Y_{i+1} - Y_i$ is small, for instance Y_{20}, Y_{21}, the posterior distribution of the related divergence time, has most of its probability mass around values close to one. On the other hand, if such a difference is relatively big, for instance Y_{22}, Y_{23}, then it is more likely that the posterior distribution will be concentrated around values closer to zero than to one. If this difference is neither very big nor very small, then it is harder be very confident about the divergence time point and therefore we would expect a rather flatter distribution. Figure 5.11 indicates the fairly well capability of the model to capture the *true* and *unknown* Uniform *"jump distribution"* and the posterior distribution of the various divergence time points.

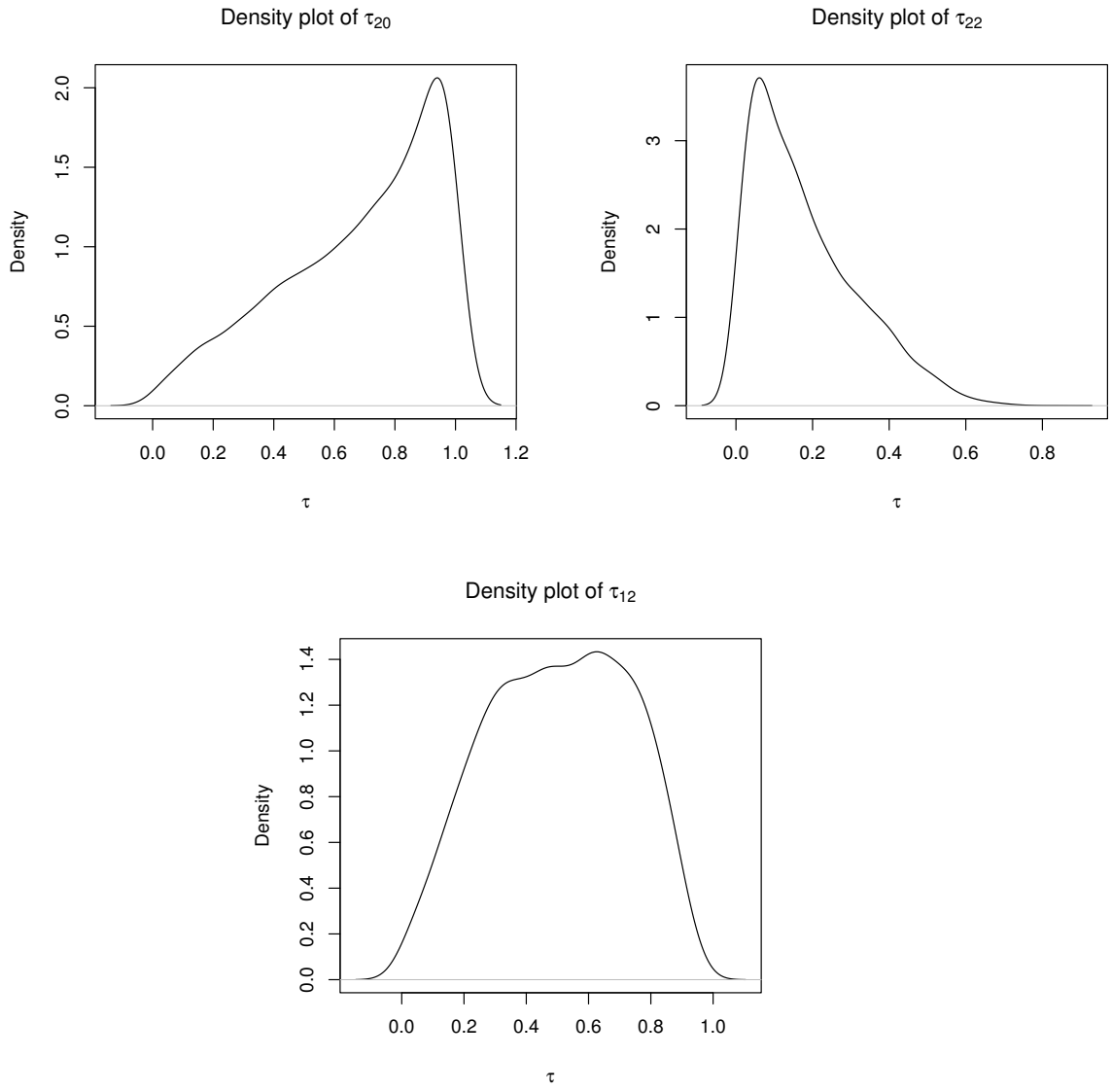

Figure 5.11: Posterior distributions of the divergence times $\tau_{20}, \tau_{22}, \tau_{12}$

Regarding the mixing of the MCMC algorithm it turns out to be relatively good despite the large dimension of the parameter's space which need to be imputed (see Figure 5.12 which shows the ACF plot of the average divergence time). Other possible strategies can also be considered such those described in Section 5.9 to obtain posterior samples for the parameters of interest.

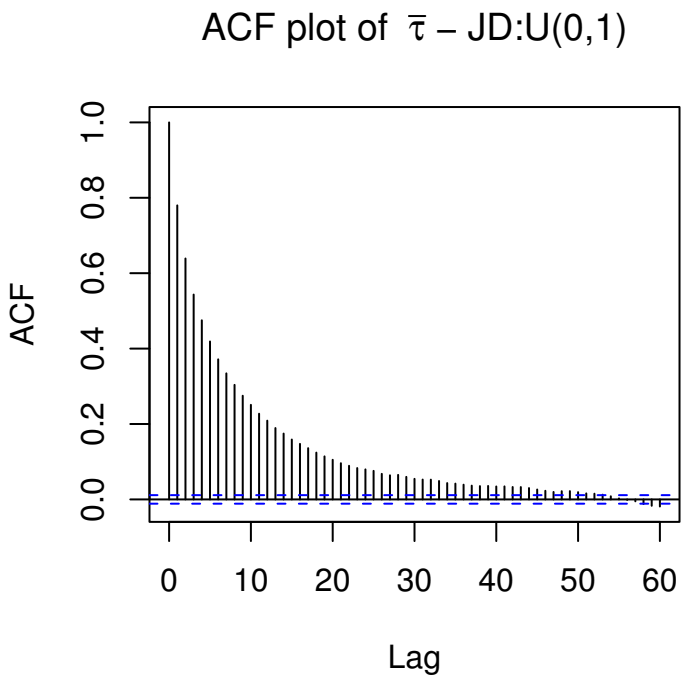

Figure 5.12: ACF plot of the average divergence time point - JD:U$(0, 1)$

5.10.2 JD: Fréchet

Table 5.2 shows that if a Fréchet *"jump distribution"* is chosen, whose density is given by:

$$f(\tau_i) = \frac{1}{\tau_i^2} \exp\left\{1 - \frac{1}{\tau_i}\right\}, \qquad 0 \leq < \tau_i \leq 1$$

then the realizations obtained from a LBT are very heavily correlated. We apply the standard centered MCMC algorithm and derive the posterior distributions of the divergence times. A Uniform proposal is used to update the times (as before). Nevertheless, one could also propose a new value for the parameters $\boldsymbol{\tau}$ by drawing values from the prior, i.e. the Fréchet distribution using inverse CDF method.

We simulate again $n = 200$ observations using a LBT where the divergence times follow a Fréchet distribution. Assuming such a prior over the times, $\pi(\boldsymbol{\tau})$, we obtain posterior samples, $\pi(\boldsymbol{\tau}_i|\boldsymbol{Y})$, $i = 1, \ldots, n-1$, for each of the parameters. We derive a sample from posterior distribution of the CDF of the times and again

the model performs really well (see Figure 5.13). Apart from this, the estimated average divergence, $\overline{\tau}$, time (using Equation 5.22) is equal to 0.592, fairly close to the true value (0.595). In addition the mixing of $\overline{\tau}$ seems to be relatively good (see Figure 5.14) taking into account the large dimension of the parameter space. As before, block update of the locations parameters, or integrating them out is also feasible.

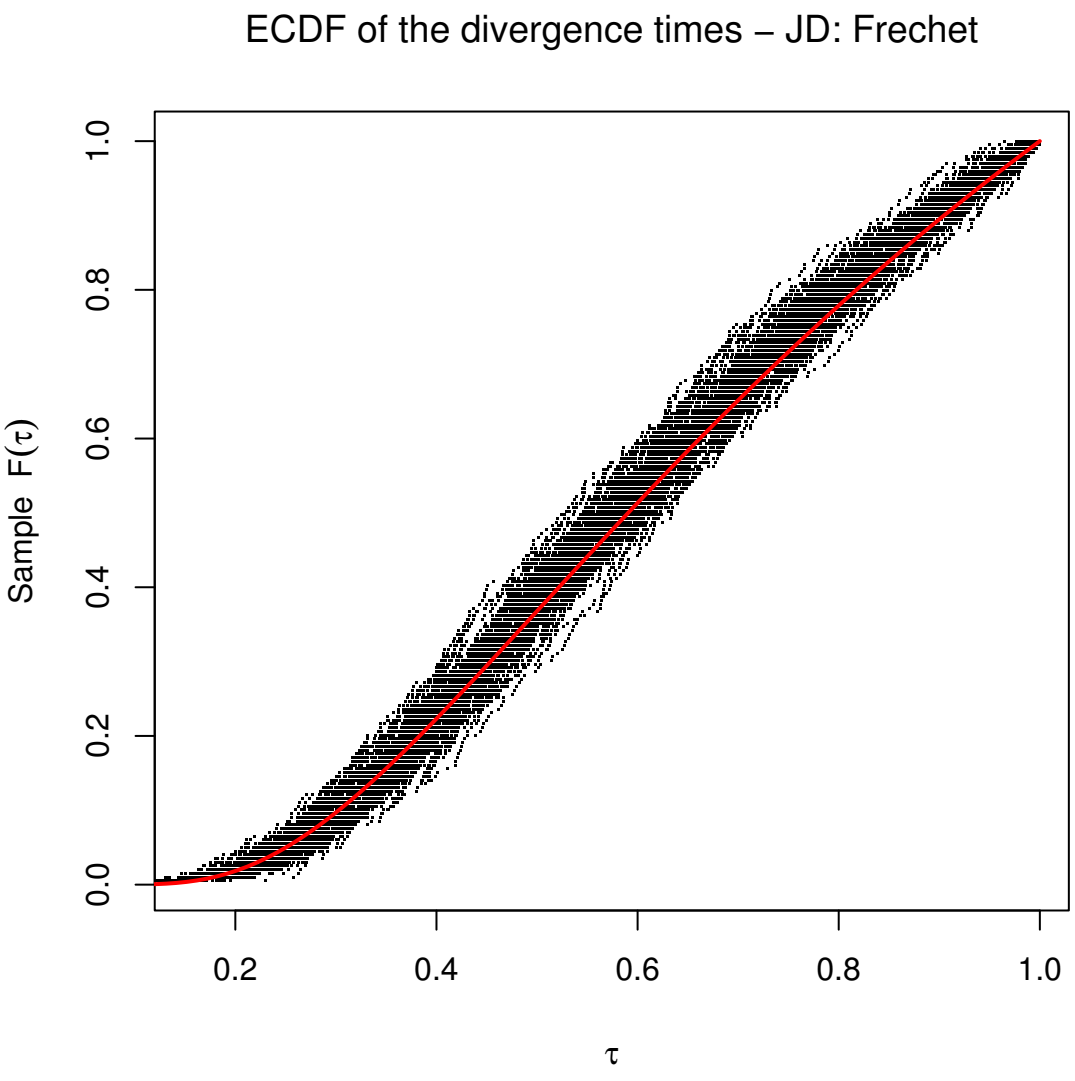

Figure 5.13: ECDF of times for the Fréchet *"jump distribution"* - Red line shows the true CDF

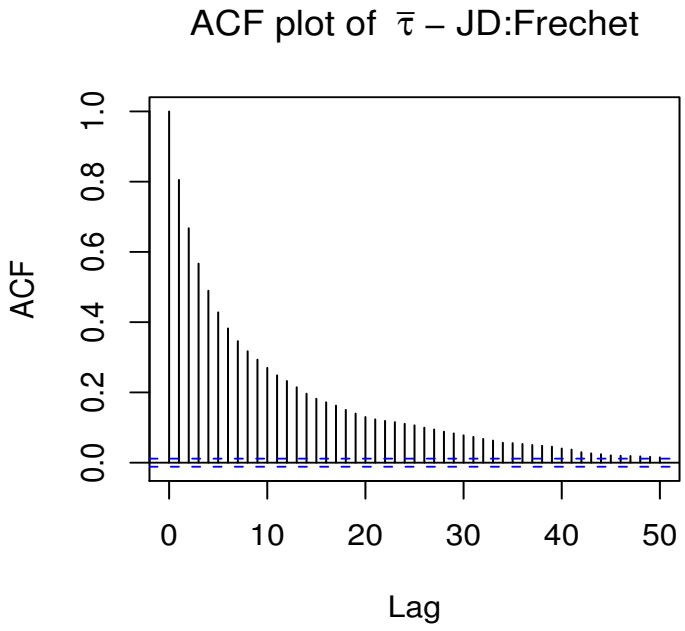

Figure 5.14: ACF of the average divergence time point - JD: Fréchet

5.10.3 JD: Beta$(\alpha, 1)$

So far, the examples we have considered did not involve any model parameters ($\boldsymbol{\theta}_3$). We now turn our attention to a simulated dataset from a LBT by assuming the following Beta$(\alpha, 1)$ *"jump distribution"* :

$$f(\tau_i) = \alpha \tau_i^{\alpha-1}, \quad 0 \leq \tau_i \leq 1, \;\; a > 0$$

In this case apart from drawing the posterior distribution of each of the divergence times, we are also focusing on making inference for the *"jump parameters"* as well, i.e. $\boldsymbol{\theta}_3 = \alpha$. The likelihood term L_1 is equal to

$$\begin{aligned} L_1 &= \prod_{i=1}^{n-1} f(\tau_i) \\ &= \prod_{i=1}^{n-1} \alpha \tau_i^{\alpha-1} \end{aligned}$$

$$
\begin{aligned}
&= \alpha^{n-1} \prod_{i=1}^{n-1} \exp\left\{\log\left(\tau_i^{\alpha-1}\right)\right\} \\
&= \alpha^{n-1} \exp\left\{(\alpha-1)\sum_{i=1}^{n-1}\log\left(\tau_i\right)\right\} \qquad (5.23)
\end{aligned}
$$

The other likelihood term L_2 refers to the tree structure and is given in Equation 5.8. Once the Beta$(\alpha, 1)$ prior is assigned, then a prior for the hyper-parameter α also needs to be adopted. Since that parameter is strictly positive we choose a conjugate Gamma prior with parameters λ_α and ν_α:

$$\pi(\alpha) \equiv Ga(\lambda_\alpha, \nu_\alpha).$$

The full posterior distribution of the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \alpha)^T$, where $\boldsymbol{\theta}_1 = \boldsymbol{\tau}$ and $\boldsymbol{\theta}_2 = \boldsymbol{X}(\tau)$ is given in Equation

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) \propto \alpha^{\lambda_\alpha - 1} \exp\left\{-\alpha\left(\nu_\alpha - \sum\log\left(\tau_i\right)\right)\right\} \times L_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \qquad (5.24)$$

The full conditional posterior distribution of the *jump parameter* α is then easily derived due to the chosen conjugate prior:

$$\pi(\alpha|\boldsymbol{X}(\tau), \boldsymbol{\tau}) \equiv \left(n + \lambda_\alpha - 1, \nu_\alpha - \sum_{i=1}^{n-1}\log(\tau_i)\right) \qquad (5.25)$$

We construct an MCMC algorithm which is based on the centered algorithm described in Section 5.8.2 and is shown below:

Centered Algorithm - JD : $\tau \sim Beta(\alpha, 1)$

(Repeat the following steps)

1. `Start the chain with initial values:` α^0, τ_j^0, $X^0(\tau_j)$, $j = 1, \ldots, n-1$;

2. `Choose uniformly one (or more) divergence points` τ_j `and update it (them) using Metropolis Hastings algorithm;`

3. `Update each of the location parameters` $X(\tau_j)$, $1 \leq j \leq n-1$ `using Gibbs sampler;`

   or

   `update them as a block;`

4. `Update` α `by using Gibbs Sampler and drawing from its conditional distribution (5.25).`

The form of the conditional distribution of α allow us to perform a Gibbs step to update it. On the other hand, a Metropolis-Hastings step is needed to update the times $\boldsymbol{\tau}$. First, a Uniform proposal was tried, but due to its inefficiency, a Beta$(\alpha^c, 1)$ was chosen which turned out to be perform rather better than the Uniform proposal. In detail, Step 2, was implemented as follows:

2.1 `Choose uniformly one (or more) divergence points` τ_j, $j = 1, \ldots, n-1$;

2.2 `Propose` τ_j':

$$\tau_j' \sim \texttt{Beta}(\alpha^c, 1);$$

2.3 `Accept` τ_j`, with probability:`

$$1 \wedge \frac{\pi(\boldsymbol{\tau}'|\boldsymbol{X}(\tau), \boldsymbol{Y}, \alpha)}{\pi(\boldsymbol{\tau}|\boldsymbol{X}(\tau), \boldsymbol{Y}, \alpha)} \frac{\tau_j^{\alpha^c - 1}}{\tau_j'^{\alpha^c - 1}}$$

where $\boldsymbol{\tau}' = (\tau_1, \tau_2, \ldots, \tau_{j-1}, \tau_j', \tau_{j+1}, \ldots, n-1)$. Note that α^c denotes the current value of parameter α. The algorithm states that we should choose uniformly one of the τ_j's and propose to update it. It turns out that the more τ_j's we choose to update then better the mixing of the Markov chain is.

Integrate α out

In addition, because of the closed form of the conditional posterior distribution of the model parameter, $\pi(\alpha|\boldsymbol{\tau}, \boldsymbol{X}(\tau), \boldsymbol{Y})$, we can integrate it out and have as a target distribution:

$$\pi(\boldsymbol{\tau}, \boldsymbol{X}(\tau)|\boldsymbol{Y}) = \int_\alpha \pi(\alpha, \boldsymbol{\tau}, \boldsymbol{X}(\tau)|\boldsymbol{Y}) \, \mathrm{d}\alpha \tag{5.26}$$

The Step 2 of the MCMC algorithm has to be modified in order to be applied properly. This is because we cannot propose $\tau_j' \sim \text{Beta}(\alpha^c, 1)$ since α does not exist in the parameter space any more. However, we can substitute α^c either with a fixed value which may obtained from a pilot study. Apart from such a choice, we could also update the divergence times by using a maximum likelihood estimator of α given the current values of the Markov chain:

2.1 `Evaluate the MLE of` α `given the current values of the divergence times` $\boldsymbol{\tau}^c$`,`

$$\alpha_{ML}^c = \frac{1-n}{\sum_{i=1}^{n-1} \log(\tau_i^c)};$$

2.2 `As the Step 2.1 of the previous algorithm;`

2.3 `As the Step 2.2 of the previous algorithm;`

2.4 `As the Step 2.3 of the previous algorithm;`

The true value of α is 4.0 and a rather non-informative prior is adopted – a Gamma$(0.1, 0.1)$ i.e. mean equal to 1 and variance equal to 100. First we apply, the standard MCMC algorithm and apart from concentrating on the divergence time τ_i's, $i = 1, \ldots, n-1$, we also look at the model parameter α. In terms of model fit, we derive a posterior sample from the CDF of the divergence time points and the posterior density of α. Figures 5.15 and 5.16 show a pretty good fit since the mode of the posterior distribution of α is around value 4.0.

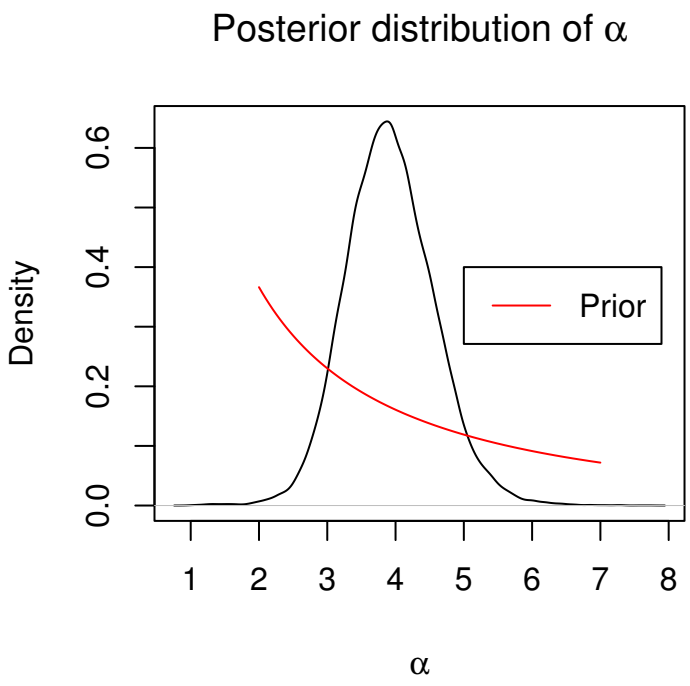

Figure 5.15: Posterior distribution of α obtained via the Centered MCMC algorithm

Nevertheless, the mixing of "trace" of parameter α is very slow and this can lead to inaccurate inference (see Figure 5.17). Such a poor mixing can be explained by looking at a scatter plot between the missing data ($\boldsymbol{\tau}$) and the model parameter (α). Figure 5.18 indicates a high correlation between the average divergence time $\overline{\boldsymbol{\tau}}$ and α. If we consider these two as the only parameters of interest, then it is well known for a two-state Gibbs sampler the convergence of the algorithm is linked to the correlation between the parameters (Amit, 1991). Although the situation here

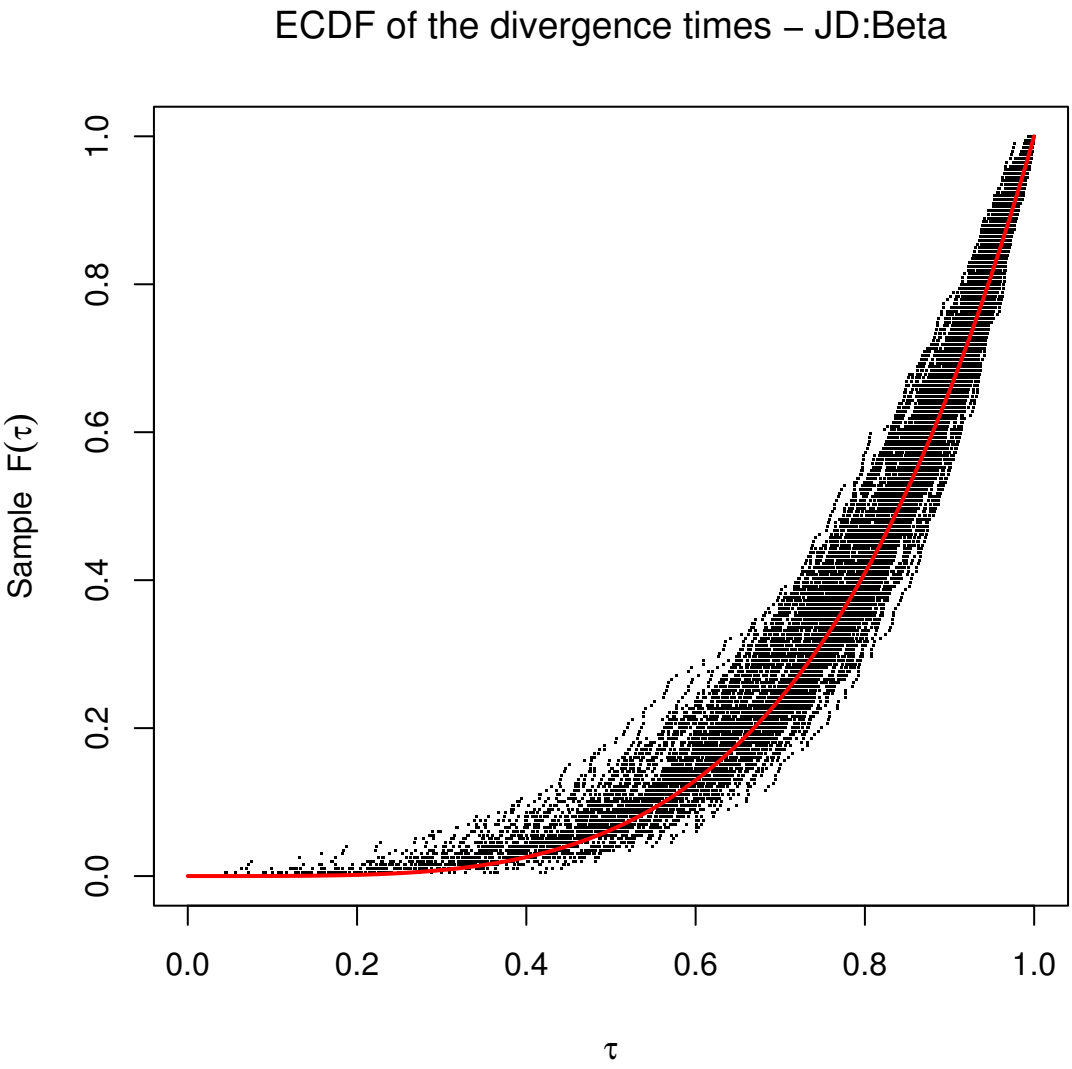

Figure 5.16: ECDF of times for the Beta$(\alpha, 1)$ *"jump distribution"* - Red line shows the true CDF

is far more complicated than a two-state Gibbs sampler, this gives us an intuition why such a problematic mixing occurs. A non−centered reparameterisation gets around this problem of high correlation by breaking down the dependence between the divergence times $\boldsymbol{\tau}$ and α. We show now the techniques described in Section 5.9.3 can be applied for this case-specific Beta$(\alpha, 1)$- *"jump distribution"* .

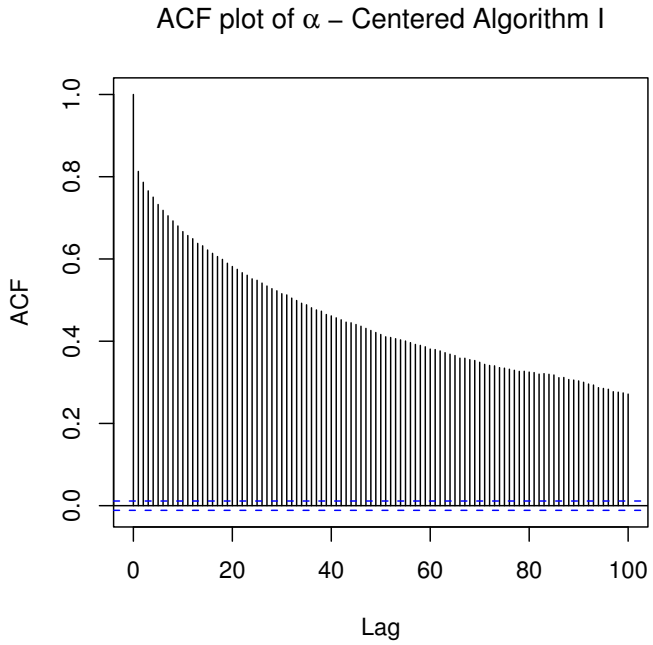

Figure 5.17: ACF plot of the posterior sample obtained via the centered algorithm

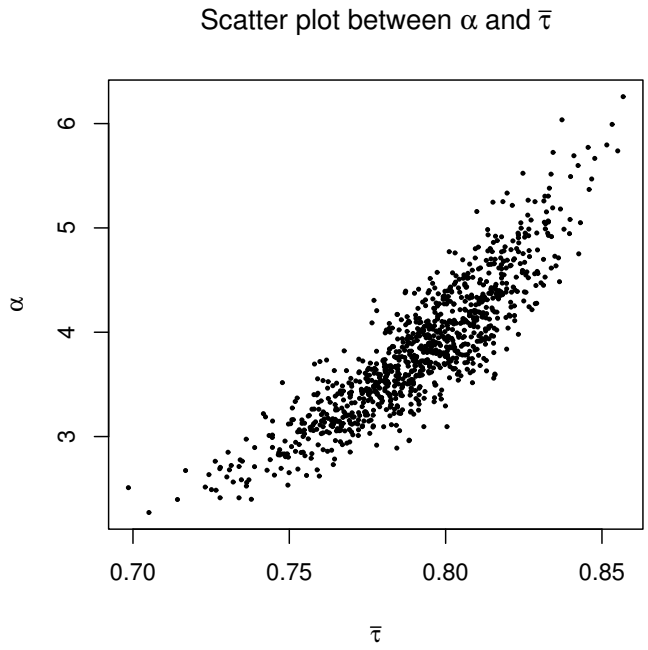

Figure 5.18: Correlation plot between missing data and model parameter

Non−Centered Parameterisations

Figure 5.19 shows a graphical representation of a centered and a non−centered pararameterisation of a LBT model with a Beta$(\alpha, 1)$ prior over the divergence times.

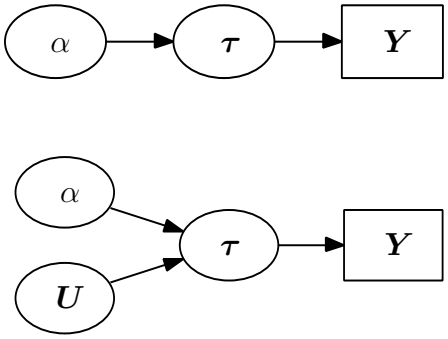

Figure 5.19: Centered (top) and Non−Centered Parameterisations

The first step needed for the implementation a NCP is to choose a function $g(\cdot)$ which will make the missing data and the parameter of interest *a priori* inde-

pendent. As suggested already (see Section 5.9.3), the cumulative distribution function of each of the missing data can be used. We can introduce the following random variables:

$$U_i = \tau_i^{\alpha}$$

such that *a priori*, $U_i \sim U(0,1)$, $i = 1, \ldots, n-1$. The Jacobian of such transformation is:

$$|\mathcal{J}| = \alpha^{n-1} \prod_{i=1}^{n-1} U_i^{1-\frac{1}{\alpha}}$$

Once a change in variables has been introduced, then the full posterior distribution with respect to the *new* variables must be derived:

$$\pi(\alpha, \boldsymbol{X}(U), \boldsymbol{U}) \propto L_2(\boldsymbol{U}, \boldsymbol{X}(U)) \cdot \alpha^{\lambda_\alpha - 1} \exp\{-\nu_\alpha \alpha\} \tag{5.27}$$

where $\boldsymbol{X}(U)$ denotes the vector location parameters with respect to the introduced variables U_i, $i = 1, \ldots, n-1$ instead of the times τ_i, $i = 1, \ldots, n-1$. Note, that due to the choice of function $g(\cdot)$, the term L_1 (see Equation 5.23) which appears in (5.24) cancels with the Jacobian term above. In addition, the full conditional distribution of α is not of closed form any longer. Therefore, a Metropolis-Hastings step is needed to update it. Using the general algorithm of performing a NCP the following case-specific one can be adopted:

Non-Centered (NC) MCMC algorithm - JD: Beta($\alpha, 1$)

(Repeat the following steps)

1. `Start the chain with initial values for the parameters:` $\boldsymbol{\tau}^0, \boldsymbol{X}(\tau)^0, \alpha^0$;
2. `Update` $\boldsymbol{X}(\tau)$ `either by deterministic scan, or a block update scheme`;
3. `Obtain a sample of` $(\boldsymbol{\tau}^1, \alpha^1)$ `via a centered algorithm`;
4. `Get` $U_i^1 = g\left(\tau_i^1, \alpha^1\right)$;
5. `Update` α `using a Metropolis Hastings step by proposing` $\alpha^2 \sim q\left(\alpha^1\right)$ `and accept it with probability :`

$$1 \wedge \frac{\pi(\alpha^2|\boldsymbol{U}, \boldsymbol{X}(U))}{\pi(\alpha^1|\boldsymbol{U}, \boldsymbol{X}(U))} \frac{q(\alpha^1, \alpha^2)}{q(\alpha^2, \alpha^1)};$$

[Note that by updating α, the divergence times τ_i's are updated as well since ${\tau_i}^2 = U_i^{1^{(1/\alpha^2)}}, i = 1, \ldots, n$]

Based on the methods described in Section 5.9.3 we can update α using the following proposals:

- **Random Walk Metropolis**: $q(\cdot) \equiv N(\alpha, \sigma_\alpha^2)$. Then, the acceptance rate becomes:

$$1 \wedge \frac{\pi(\alpha^{'}|\boldsymbol{X}(U), \boldsymbol{U})}{\pi(\alpha|\boldsymbol{X}(U), \boldsymbol{U})}.$$

  Alternatively, a multiplicative random walk can be used.

- **"Pseudo-Gibbs" Sampler:** $q(\cdot) \equiv Ga\left(n + \lambda_\alpha - 1, \nu_\alpha - \sum_{i=1}^{n-1} \log \tau_i\right)$. The proposed value is accepted with probability equal to:

$$1 \wedge \frac{\pi(\alpha^{'}|\boldsymbol{X}(U),\boldsymbol{U})}{\pi(\alpha|\boldsymbol{X}(U),\boldsymbol{U})}\frac{q(\alpha^{'},\alpha)}{q(\alpha,\alpha^{'})}$$

- **Normal Approximation:** Denote by $\widehat{\alpha} = \frac{1-n}{\sum_{i=1}^{n-1}\log(\tau_i^c)}$ and by $\sigma(\widehat{\alpha}) = \frac{\widehat{\alpha}^2}{n-1}$. Propose a new value as follows:

$$\alpha' \sim N\left(\widehat{\alpha}, \epsilon\sigma\left(\widehat{\alpha}\right)\right)$$

  where $\epsilon > 1$ and $\tau_i^c, i = 1\ldots, n-1$ denote the current value of the divergence times.

- **Adaptive Sampling**: can be applied very straightforward (for more details, see Section 5.9.3).

Apart from a fully non−centered reparameterisation (NCP), a partially non−centered parameterisation (PNCP) exists as well which can be implemented in various ways. The first has already been described in Section 5.9.3 by non−centering some of the divergence times. On the other hand, another way of non−centering is to decide at each MCMC step, whether to center or to non−center all the the times $\boldsymbol{\tau}$.

Figure 5.20 shows the significant improvement in efficiency gained by using a NCP in comparison to the centered algorithm (see Figure 5.17). Note that the available options for updating α were tried and all gave similar results. It should be mentioned that Step 3, where a sample of $(\boldsymbol{\tau}, \alpha)$ via centered algorithm is needed, is implemented by having as target, the posterior distribution having α integrated out, i.e. $\pi(\boldsymbol{\tau}, \boldsymbol{X}(\tau))$ since it makes the algorithm more efficient. The integrated autocorrelation time (IAT) was computed for both algorithms showing a reduction of a factor of 4.5 for the non−centered algorithm compared to the centered.

Concluding, we realize that once high correlation between the missing data and the model parameter exists, a NCP seems to be essential in order improve the rate of convergence of the standard (CP) MCMC algorithm. Such a parameterisation is needed even more when the dimension of parameter space (i.e missing data,

model parameters) since dependence gets even larger.

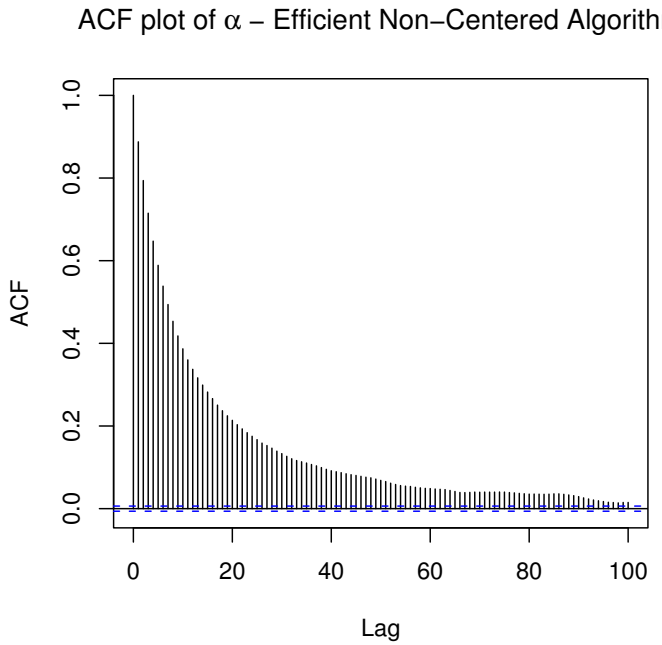

Figure 5.20: ACF plot of the posterior sample obtained via the non−centered algorithm

5.11 An Application on Genome Scheme Data

The simulation study performed in Section 5.10 showed a very good performance of the LBT framework. In this section, we are concerned with analysing some genome scheme data. First, before performing any sophisticated analysis we describe the nature of the data and explain why such a dataset could be modelled within our proposed framework.

5.11.1 Isochores

The availability of the human genome draft sequence offers an exceptional opportunity to bring sequence patterns into line with the chromosome structures revealed by modern molecular cytogenetics. The human genome is a mosaic of *isochores*,

which are long DNA segments (more than 300 kb on average) which are relatively homogeneous in G+C (above a size of 3kb) when compared to the pronounced heterogeneity throughout the entire genome. Higher, Lower and Medium-density genomic segments are respectively called *H*, *L* and *M* isochores. The isochore concept has been considered a "fundamental level of genome organisation" (Eyre-Walker and Hurst, 2001).

The reason why the isochores are relevant for genome biology is based on observations of gene and short interspersed repetitive elements (SINE) densities, as well as recombination frequency, which are all higher in (G+C)-rich isochores, whereas long interspersed repetitive elements (LINE) are denser in (G+C)-poor isochores (see for example, Bernardi, 2000, Oliver et al., 2004, and the references therein). Moreover, Human isochores were first identified by density-gradient ultracentrifugation of bulk DNA, and differ in important features, e.g. genes are found predominantly in the GC-richest isochores. In addition, the density of genes has been shown to be higher in *H* isochores than in *L* ones. Genes in *H* isochores are more compact with a smaller proportion of intronic sequences and code for shorter proteins than do genes in *L* isochores (see Melo de Lima et al., 2005, and the references therein).

5.11.2 Existing Methods

The recent availability of the draft human genome sequence allowed for a direct test of the isochore model and it was hoped that isochores could be identified at the sequence level. Since then, the existence of isochores in the human genome has been the object of an active debate. Therefore, different approaches have been developed for isochore prediction. Homogeneous regions such as isochores are currently ascertained by plain statistics on moving windows of arbitrary length. In other words, a long DNA sequence is divided in windows of size 3,000 say, and the numbers of C+Gs within each window is counted. Therefore, usually predictions

of isochore boundaries are based on this moving-window plot of C+G content. Nevertheless, it has been argued that the basic model seems inappropriate for various reasons such as long-range correlations (Smith and Lercher, 2002), correlation between segment means and length (Oliver et al., 2002, Bernaola-Galvan et al., 2002) and the existence of multimodal distributions of means (Bernardi, 2000). On the other hand, a LBT framework does not require such assumptions and therefore it would be of interest what kind of information can be drawn from such datasets.

5.11.3 The Data

The major histocompatibility complex (MHC) is one region of DNA sequence of an organism's genome, which contains H3 and L2 isochores. The available dataset consists of a DNA-sequence of length $3,675,000$. Windows of length of $3,000$ were chosen and the number of C+Gs within each of them is shown in Figure 5.21. Figure 5.21 reveals a pattern similar to the one shown in Figure 5.1 with less distinct "clusters". Is is interesting to see that marginally the data are Normally distributed (see left hand plot of Figure 5.23) with mean and standard error around 1382 and 193 respectively. Moreover, the data appear to have a long-range correlation (see Figure 5.22). The dataset is divided into two subsets according to the 820th observation which can be considered as a change-point. Although they are short time series, the ACFs for both series show similar long-range dependence behavior. Figure 5.24 suggests that the long-range dependence of the whole series is not due simply to non-stationarity. We transform the data to a $N(0,1)$ distribution and in association with the characteristics described above we assume that the time series falls within a LBT class.

5.11.4 A Fully Bayesian Analysis

Assuming that the data follow within a LBT structure, we will adopt a Bayesian approach in order to draw inference for the parameters of interest, i.e. the di-

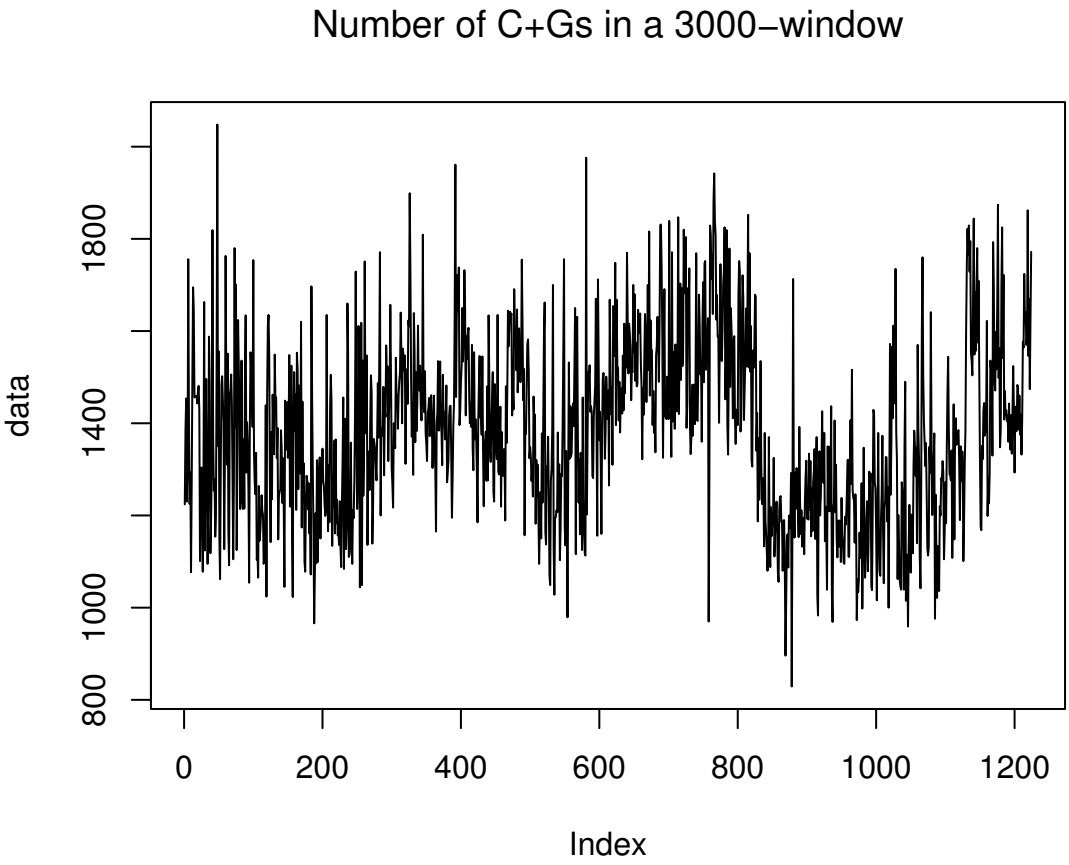

Figure 5.21: Number of C+Gs included in each window of 3,000 from the DNA sequence

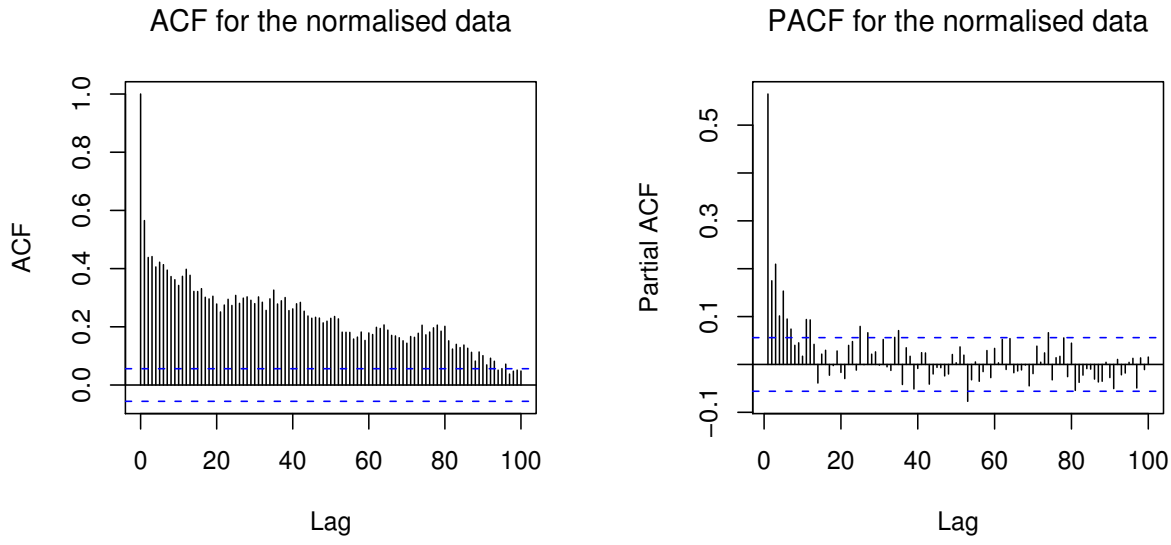

Figure 5.22: ACF and PACF plot of the normalised data

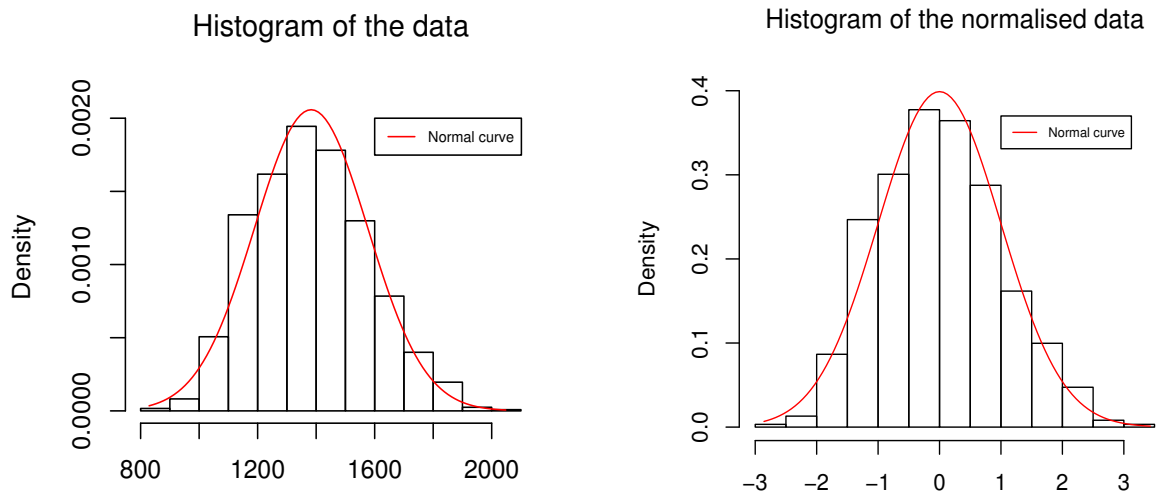

Figure 5.23: Histogram of the non-normalised (left) and normalised (right) data. Red lines reveals the Normal curve.

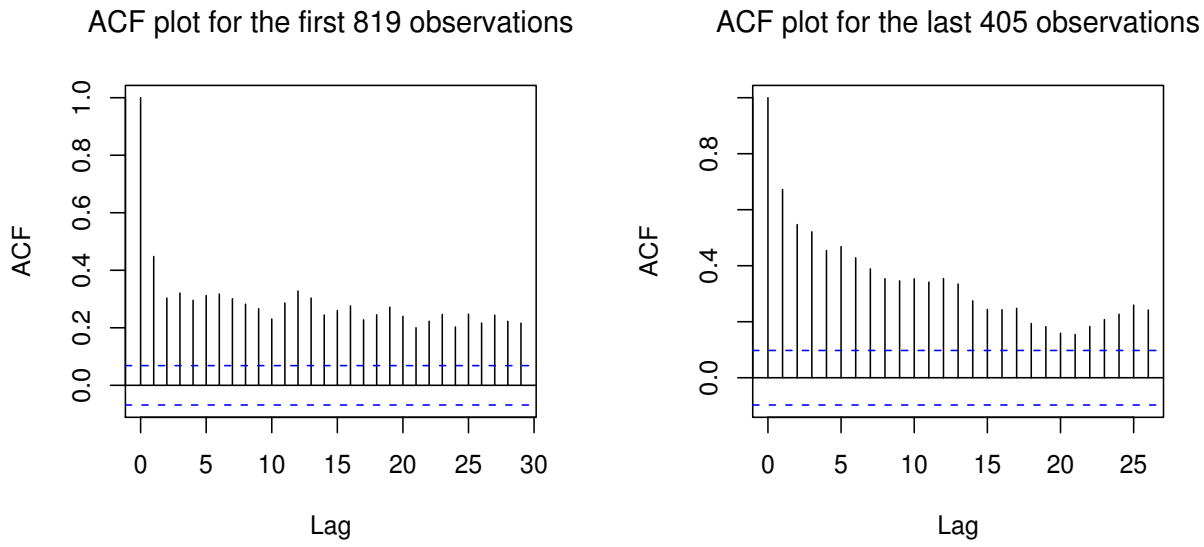

Figure 5.24: ACF plots for two subsets of the data

vergence times ($\boldsymbol{\tau}$) and the location parameters ($\boldsymbol{X}(\tau)$). We follow the approach which was described in detail in Section 5.8.

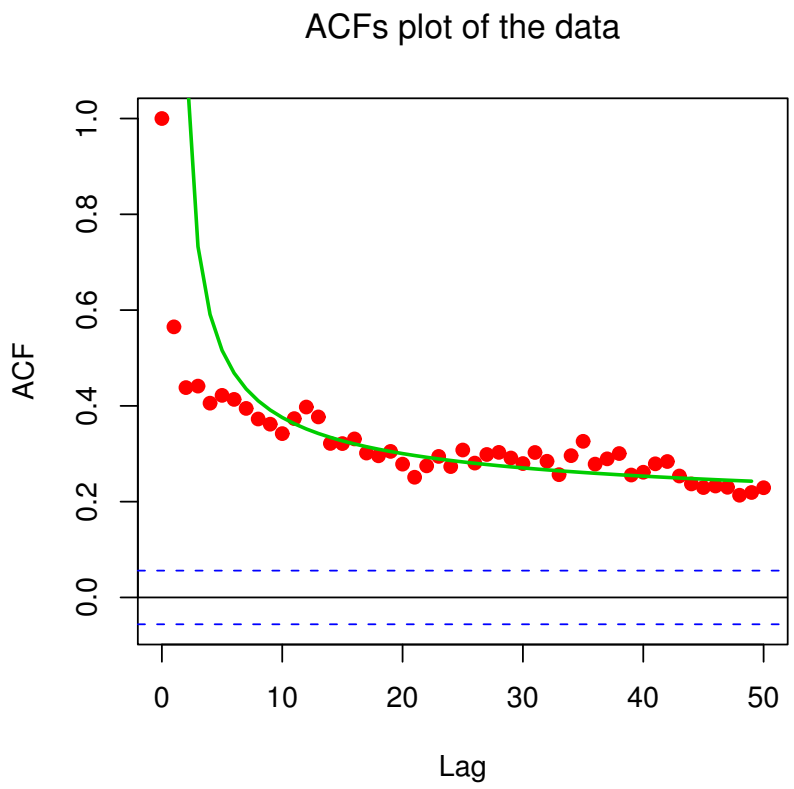

Figure 5.25: ACF plot of the data. The green line indicates the rate of decay of the covariance function according to an assumed Fréchet distribution

First, a prior over the distribution of the divergence times should be assigned. A first attempt is to assume a Uniform distribution so as to let the data to modulate the shape of the *"jump distribution"* . However, the significantly high correlation between the observations indicates that a *"jump distribution"* similar to Fréchet distribution might be also appropriate. Figure 5.25 shows that such an assumption seems very sensible as it appears to have a pretty good fit with the estimated ACF. Therefore, we decide to perform the Bayesian analysis by considering these two rather different prior distributions and compare the results.

$$\text{Prior I}: \quad \tau_i \sim f(\cdot) \quad \text{where} \quad f(\tau_i) = \tau^{-2} \exp\left\{1 - \frac{1}{\tau}\right\}$$

$$\text{Prior II}: \quad \tau_i \sim f(\cdot) \quad \text{where} \quad f(\tau_i) = U(0,1)$$

for $i = 1, \ldots, N-1$ and let $N = 1224$ denote the number of observations. It is straightforward to apply the following (centered) algorithm:

MCMC Algorithm

(Repeat the following steps)

1. Start the chain with initial values $\tau_1^0, \ldots, \tau_n^0$, $X^0(\tau_1), \ldots, X^0(\tau_n)$;
2. Choose one (or more) of the divergence parameters j, $1 \leq j \leq n-1$ and update τ_j (individually) using Metropolis Hastings algorithm;
3. Update each of the location parameters $X(\tau_j)$, $1 \leq j \leq n-1$ using Gibbs sampler;

We have chosen an independence sampler to update the divergence times by proposing from the corresponding prior, i.e. if a Uniform prior is chosen (Prior I):

2.1 Propose $\tau_j^{'} \sim U(0,1)$;

2.2 Accept $\tau_j^{'}$ with probability

$$1 \wedge \frac{\pi\left(\tau_j^{'} | \boldsymbol{\tau}_{-j} | \boldsymbol{X}(\tau)\right)}{\pi\left(\tau_j^{'} | \boldsymbol{\tau}_{-j} | \boldsymbol{X}(\tau)\right)}$$

Alternatively if the second prior (Prior II) is used then:

2.1 Propose $\tau_j^{'} \sim$ Fréchet;

2.2 Accept $\tau_j^{'}$ with probability

$$1 \wedge \frac{\pi\left(\tau_j^{'} | \boldsymbol{\tau}_{-j} | \boldsymbol{X}(\tau)\right)}{\pi\left(\tau_j^{'} | \boldsymbol{\tau}_{-j} | \boldsymbol{X}(\tau)\right)} \frac{\tau_j^{-2} \exp\left\{1 - \frac{1}{\tau_j}\right\}}{\tau_j^{'-2} \exp\left\{1 - \frac{1}{\tau_j^{'}}\right\}}$$

The location parameters are updated via Gibbs sampler as shown in Section 5.9. Note that neither Prior I nor Prior II are associated with *"jump parameters"* .

5.11.5 Results

We run the aforementioned MCMC algorithms to obtain samples from the posterior distribution of the divergence time points:

$$\pi(\tau_i|\boldsymbol{Y}), \quad i = 1, \ldots, 1224$$

which actually contain all the information needed to infer about the covariance structure of this dataset. Note that we have assumed two different priors about the *"jump distribution"* and therefore it would be very interesting to see the effect of both to the posterior distribution of the parameters of interest. The first 15 observations of the dataset are shown below:

```
-0.819757754 -0.004661498  0.366774770 -0.788804732 -0.463797997
 1.924743563  0.144944777 -0.814598917 -0.458639160 -1.578106802
-0.566974738  0.650510809  1.610054503  1.357271487  0.377092444
```

In a similar manner to the simulation study performed in Section 5.10 we are interested in examining the posterior distribution of a divergence time τ_j, $j = 1, \ldots, N-1$. Recall, that each of the times represents the time at which the $(j+1)_{st}$ diffusion diverged to another independent path to generate the $(j+1)_{st}$ observation Y_{j+1}. Figure 5.26, 5.27, 5.28 show the obtained posterior distributions for some arbitrarily chosen divergence times. We decide to visualise the posteriors via histograms instead of kernel density estimates to avoid problems of the latter showing probability mass for regions outside the interval $[0, 1]$.

Figure 5.26 shows that most of the probability mass of the marginal posterior distribution of τ_4, is concentrated around values close to one regardless of the choice of the prior. However, the effect of the Fréchet prior which places much

more probability mass at large values than the Uniform is transparent. Such a result is consistent with the structure of the LBT since Y_4 and Y_5 are relatively close.

On the other hand, the difference between Y_5 and Y_6 is much bigger. Therefore the posterior distribution of τ_5 obtained using the uninformative (Uniform) prior has located most of its probability mass at values close to zero (see top plot in Figure 5.27). This is not the case where the Fréchet prior is assumed and the mode of the posterior distribution is around 0.4, clearly affected by such an informative prior. This also holds for the divergence time τ_6 which is related with observations Y_6 and Y_7. Similarly, the informative prior "drags" the mode of $\pi(\tau_6|\boldsymbol{Y})$ to 0.4 while the Uniform prior lets it to be close to zero (see Figure 5.28). Posterior inference is drawn for the other divergence times as well and they all shared a common structure which is relation with the properties of LBT.

Summarizing, the clear message from this analysis is that extra carefulness is needed when deciding which prior to assume over the divergence times. Nevertheless, if an non−informative prior is used, such as a Uniform, then information about the parameters is based only on the likelihood.

Model Adequacy

We now empirically investigate whether an infinite order model based on a LBT framework seems suitable for this specific dataset or a simpler Markov model could be used instead.

Having obtained the posterior distribution of each of the divergence times τ_i, $i = 1, \ldots, n-1$, we simulate 1,000 realisations from the model (i.e. observations, $Y_1, Y_2, \ldots, Y_n$). Apart from the the partial autocorrelation function we also obtain a smoothed spectrum of each of the series. Moreover, a realisation of the model has been produced by treating the averages of the posterior distributions $\pi(\tau_i|\cdot)$ as point estimates for each of the divergence time points.

We first simulate a realisation of the fitted model using the posterior means of the divergence times. Figure 5.29 shows that the obtained path has a very similar pattern to the one obtained from the real data (See Figure 5.21).

Figure 5.30 shows the pacf and the smoothed spectrum plot for each of the different realisations. We should bring to attention that the plots based on the posterior means of the distribution of the divergence times look quite different from the ones which were obtained from the random samples of the distributions. This is especially the case when we look at low frequencies or small lags for the spectrum and the pacf respectively. This is due to the fact that most of the posterior distributions $\pi(\tau_i|\cdot)$ are highly skewed and therefore the posterior mean seems inappropriate as a location measure for such distributions.

We then compare the plots of the pacf and the smoothed spectrum for the simulated realisations with the corresponding plots for the observed data. Although there is some evidence for lack of fit for short lags and low frequencies (see Figure 5.30) which indicates that our model does not really capture very well the short-range dependence, on the other the hand, for high frequencies and long lags the fit significantly improves. Overall, both the smoothed spectrum and the plots of the pacf (obtained by simulating realisations) of the fitted model are in a reasonable agreement with the ones obtained from the real data. We can conclude then that a model with long-term dependence behaviour seems reasonable in order to analyse this dataset.

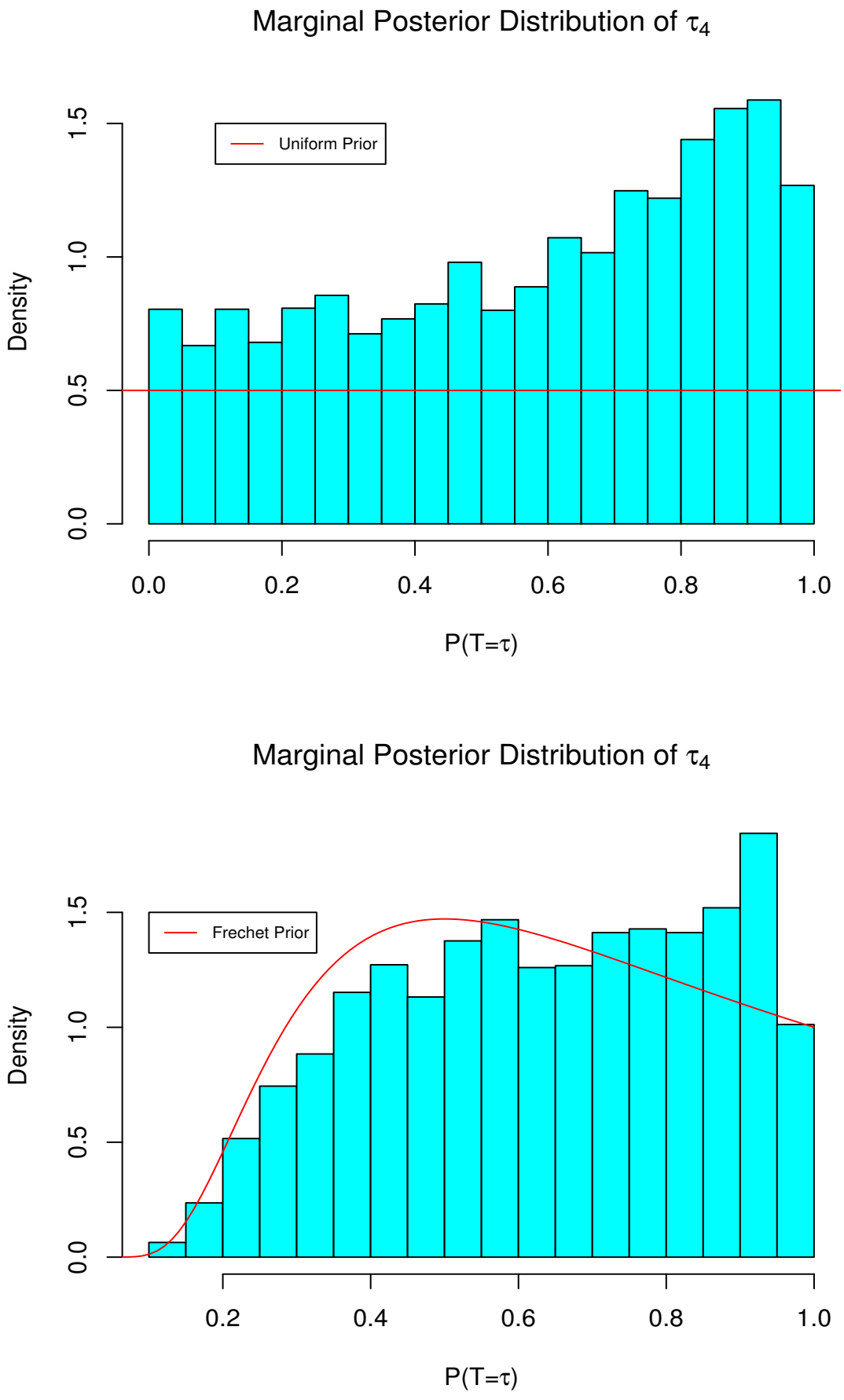

Figure 5.26: Posterior distribution for the 4_{th} divergence time points assuming a Uniform (top) and a Fréchet (bottom) prior for the vector $\boldsymbol{\tau}$.

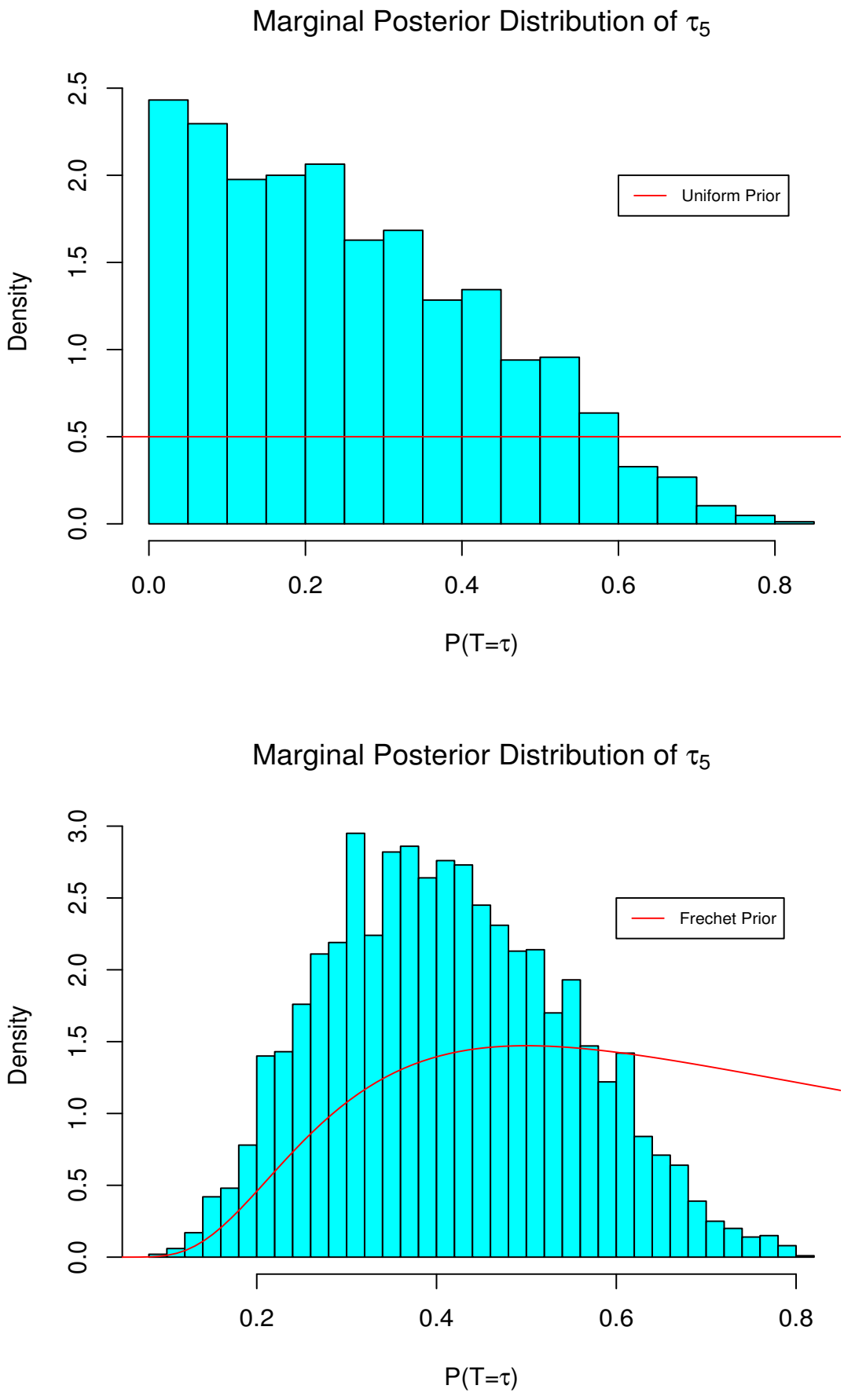

Figure 5.27: Posterior distribution for the 5_{th} divergence time points assuming a Uniform (top) and a Fréchet (bottom) prior for the vector $\boldsymbol{\tau}$.

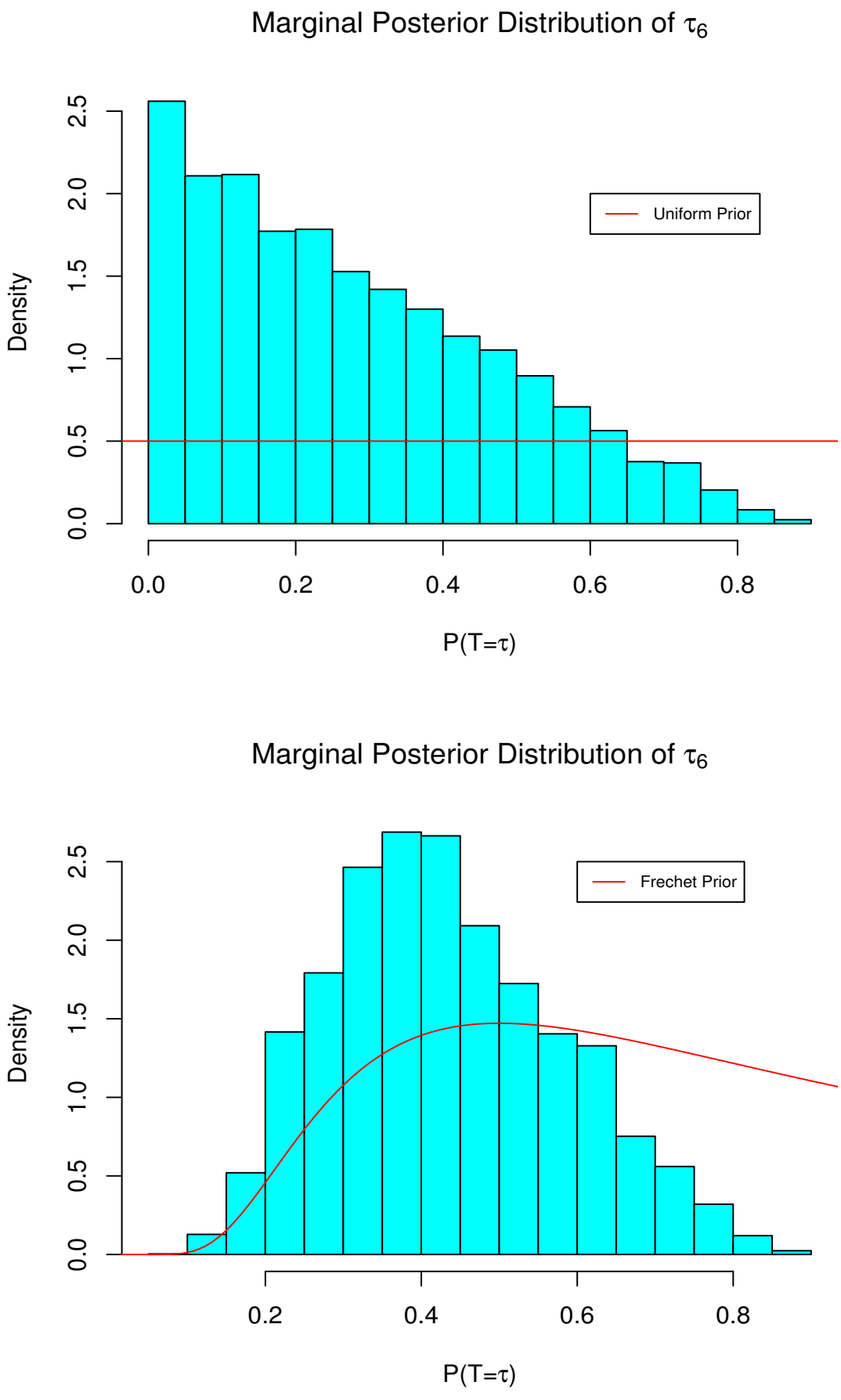

Figure 5.28: Posterior distribution for the 6_{th} divergence time points assuming a Uniform (top) and a Fréchet (bottom) prior for the vector $\boldsymbol{\tau}$.

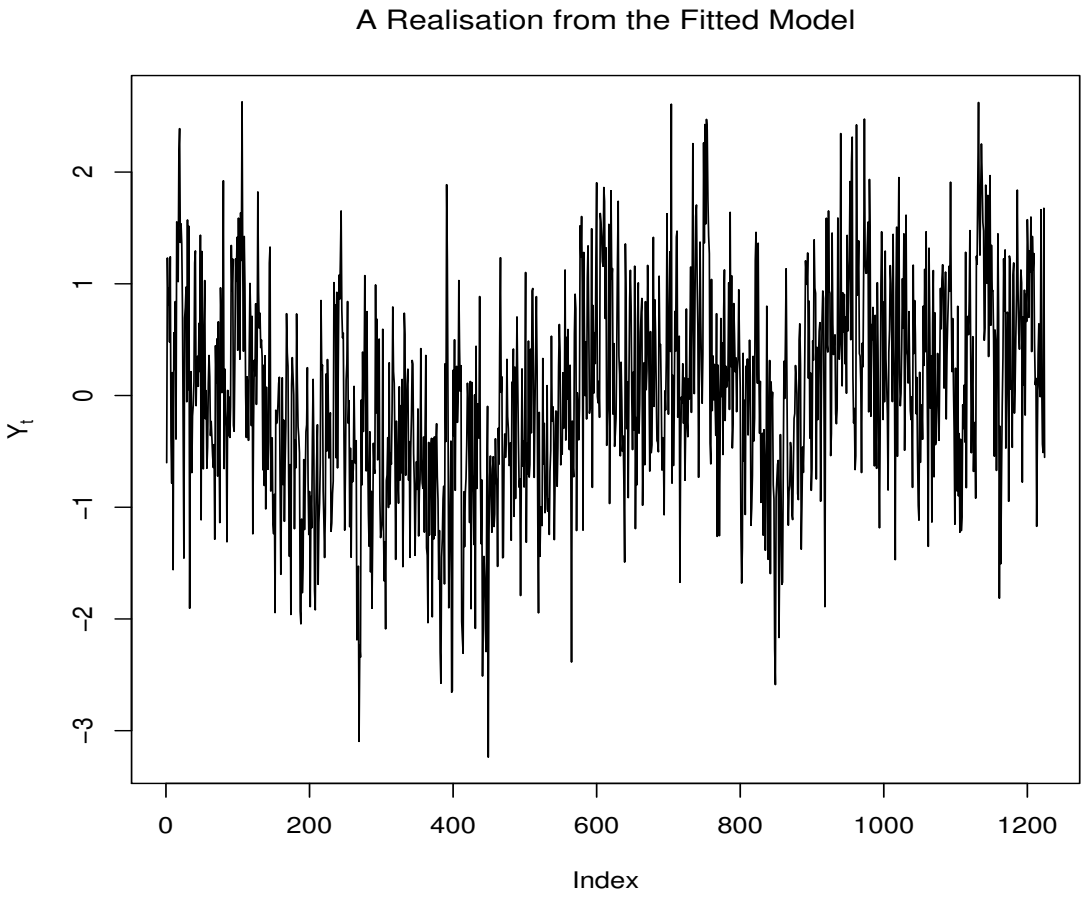

Figure 5.29: A simulate realisation from the fitted model using the posterior mean of the divergence times

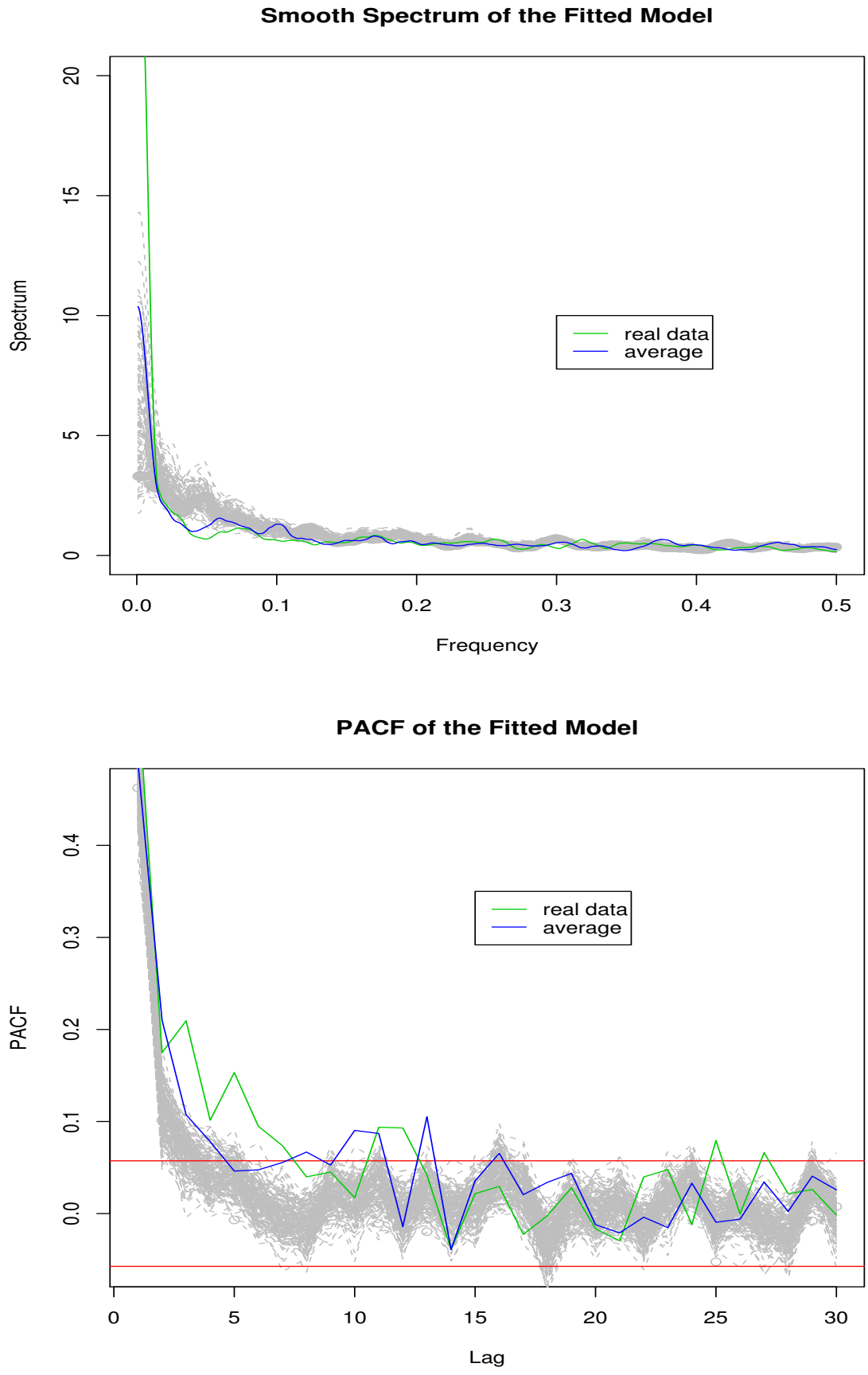

Figure 5.30: Smoothed Spectrum and PACF plots of 1,000 realisations of the fitted model using samples from the posterior distribution of the divergence time points $\boldsymbol{\tau}$. The blue line is obtained by simulating a realisation of the model using the posterior means the posterior distributions. The green line refers to the plots obtained from the actual real data.

5.12 Discussion

In this chapter, we have presented a novel methodology for constructing a class of semi-parametric time series models where the observations have fixed (and pre specified) margins with a rich collection of dependence structure. The construction of such a class of models is based on an underlying stochastic process, termed as a "latent branching tree" (LBT), via which the nature of the generated realisations is characterized; recall Section 5.4 for the construction of the LBT. A LBT offers a very flexible way of determining a rich collection of dependence structure between the observations by allowing them have margins which fall in a variety of distributions, including Normal, Exponential, Gamma and Poisson.

The construction of a LBT requires the simulation of diffusions. However, we have shown how the discretisation of the time can be avoided for some specifically chosen diffusions, by applying retrospective sampling techniques. That is, there is no need to store the full path of any diffusion but just their values at the times when they diverged, as well as the value of diffusion at these times. We have also provided methods for drawing Bayesian inference via MCMC methods. This resulted in obtaining the posterior distribution, of the divergence time points ($\boldsymbol{\tau}$), the location parameters, ($\boldsymbol{X}(\tau)$), and any other hyper parameters associated with the prior assumed over the *"jump distribution"* .

In some circumstances, especially when hyper parameters are involved, the standard MCMC algorithms do not offer well mixing Markov chains and therefore, the need for more efficient algorithms is essential. Strategies for block update of the location parameters or integrating them out and efficient non−centered parameterisations (ENC) have been presented. Within the context of a particular example, the efficiency gained by using an ENCP instead of a CP was transparent. Nevertheless, when choosing an alternative MCMC strategy, the computational cost and the cpu time need needed to run the algorithm should always be taken into account and decide which to use, on the basis of their relative efficiency.

We should bring to attention that the implementation of any of the MCMC algorithms mentioned in this chapter is complicated and care is necessary so as to provide reliable results. As the dimension of the dataset increases, and in consequence the dimension of the parameter space, block updating or integrating out the location parameters can be dramatically very time consuming. On the other hand, due to the construction of the tree, routinely used techniques for sparse matrices in other contexts such as Markov random fields (see for example, Knorr-Held and Rue, 2002, Rue et al., 2004), can be easily accommodated within this framework. This observation is very important especially when integrating the location parameters out or updating them as a block, offers significantly faster, in terms of convergence, Markov chains.

Examples with simulated datasets showed that the performance of the LBT is very good since it manages to capture the *true* unobserved and underlying stochastic process. However, in applications with real datasets, it is not always obvious what prior distributions shall be assumed over the divergence times. An alternative (and simpler) approach is to assume a Uniform prior and then let the data indicate the *true "jump distribution"* (see Section 5.11).

In conclusion, the general modelling framework presented in this chapter has much to offer in the area of time series where the dependence structure is assumed not to be of a standard form, eg. AR-type and also when observations have marginal distributions outside the Gaussian context.

5.13 Further Work

In this section we will briefly refer to some work we will be interested in doing in the future regarding generalisations of the existing methodology of constructing a LBT and also extensions which are motivated by applications which consist of real datasets.

5.13.1 Methods

Marginals outside the Gaussian context

In Section 5.4 we described in detail the construction of a LBT. It has been noted, that in principle any diffusion can be used instead. For instance, if we are interested in modelling data which are Gamma distributed we could have used a Gamma process. On the other hand, even with the existing methodology this is feasible. That is, by choosing a Brownian motion as the driving diffusion and then an extra level to the hierarchy by transforming the Normally distributed variables (as shown in Table 5.1). The problem becomes more interesting, when we focus on integer-valued distributions such as Poisson where a Poisson process can be used.

Prediction

A very important aspect in the area of time series is the prediction of the future. Within our context this can be done very straightforward due to the Bayesian approach adopted. Therefore, once we obtain posterior samples for the unknown parameters, say $\boldsymbol{\theta}$, we can (forward) simulate a LBT (see Section 5.7). Note, that with this approach, the uncertainty about the parameters is taken into account and is being integrated out. The prediction's performance of a LBT should be examined and investigate whether more robust predictions could be obtained via such a framework than standard time series modelling.

General Proposals for the MCMC Algorithms

An important issue while constructing an MCMC algorithm which involves a Metropolis-Hastings step, is the choice of the proposal distribution. Although no particular problems occurred throughout the examples described so far, we bring to attention the following result. For any two independent Normally distributed

random variables, say Z_1, Z_2, i.e.

$$\begin{aligned} Z_1 &\sim N(0,1) \\ Z_2 &\sim N(0,1) \end{aligned}$$

if we define $Z = Z_1 - Z_2$, then

$$\begin{aligned} Z &\sim N(0,2) \\ \mathbb{P}(|Z| < z) &= \mathbb{P}(-z < Z < z) \\ &= \mathbb{P}\left(-\frac{z}{\sqrt{2}} < \frac{Z}{\sqrt{2}} < \frac{z}{\sqrt{2}}\right) \\ &= \Phi\left(\frac{z}{\sqrt{2}}\right) - \Phi\left(-\frac{z}{\sqrt{2}}\right) \\ &= \Phi\left(\frac{z}{\sqrt{2}}\right) - \left(1 - \Phi(\frac{z}{\sqrt{2}})\right) \\ &= 2\Phi\left(\frac{z}{\sqrt{2}}\right) - 1 \end{aligned} \tag{5.28}$$

where $\Phi(\cdot)$ denotes the distribution function a random variable following a standard Normal(0,1) distribution. Equation (5.28) states the probability of the absolute difference between two consecutive observations (obtained via a LBT) being less than a certain value. Figure 5.31 shows the cumulative distribution function (CDF) for $|Z|$ and the corresponding complementary CDF. Although (5.28) refers to standard Normal random variables, it is straightforward to generalize it for any Normal random variable with mean μ and variance σ^2.

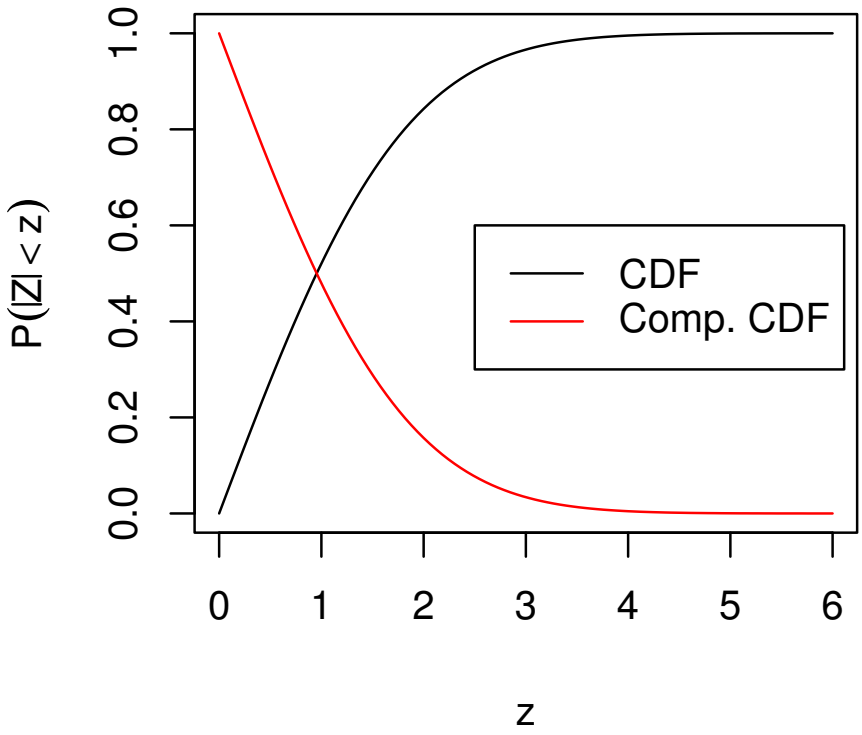

Figure 5.31: The cumulative distribution function of the random variable $|Z|$, where $Z = Z_1 - Z_2$, with $Z_i \sim N(0, 1)$, for $i = 1, 2$

It is of practical interest to construct a general proposal which will take into account the absolute difference of the two observations, Y_i and Y_{i+1}. For instance, if the observed difference $|Y_i - Y_{i+1}|$ is very unlikely to have been occurred by chance (according to (5.28)), then a proposal with most of its probability mass around values close to one should be used. Note that the same argument holds regardless of whether the difference is relatively big or small.

Bayesian Non-Parametric

Within the current framework, a parametric form of the *"jump distribution"* has to be specified in advance before constructing the LBT. Note that this often leads to the need of drawing inference for the hyperparameters and this could be prob-

lematic. A further extension of the current work is to relax this condition and to become fully non-parametric by considering a Dirichlet distribution for the divergence times ($\boldsymbol{\tau}$).

Multivariate Time Series

Throughout this chapter we were mainly concerned with constructing and drawing inference for univariate time series. There exist applications where we are interested in modelling multivariate time series. Within our framework, extensions for Normally distributed data can be done as follows. While constructing the diffusion paths, at each of the divergence time point, instead of drawing from a univariate N(0,1), we could generate realisations from a multivariate Normal of a pre specified dimension. Nevertheless, more consideration is needed when we are analysing time series where the observation have distributions outside the Gaussian family.

Concluding, we should bring to attention that any extension which is mentioned above must be in agreement with the general structure of the tree. In other words, if another diffusion is chosen and if it has independent increments, we should be able to write explicitly its density such that the likelihood is easy to handle and make the inference feasible.

5.13.2 Applications

Motivated by the genome scheme data which were analysed in Section 5.11, a particular form a *"jump distribution"* which could be assumed:

$$\mathbb{P}(\tau) = \begin{cases} \tau_1, & \text{with probability } 1-p \\ \tau_2, & \text{with probability } p \end{cases}$$

Such a discrete *"jump distribution"* will allow for the Brownian motions either to diverge at time τ_1 or τ_2. By an appropriate choice of these two divergence time points and the "jump probability" p, say, $\tau_1 = 0$,$\tau_1 = 0.9$ and $p = 0.9$ then most

of diffusions will diverge at times close to 1 and a few times close to 0. Due to the fundamental property of a LBT which states that if the stochastic process diverges at times close to zero (or close to one) then the observations Y_k, Y_{k+1}, $k \in (1, \ldots, n-1)$ will be less (much) correlated, approximately $(1-p) \times N$ clusters will be created. This reminds us of a standard change-point model with one change-point (at time τ_1). A natural extension to multiple change point model (with fixed number of change points) can easily done by allowing the following *"jump distribution"*

$$P(\tau) = \begin{cases} \tau_1, & \text{with probability } p_1 \\ \tau_2, & \text{with probability } p_2 \\ \vdots & \\ \tau_k, & \text{with probability } p_k \end{cases}$$

The available flexibility on determining the *"jump distribution"* of a LBT allows us to perform a further step of generalization and consider the following prior over the divergence times:

$$f(\tau) = \begin{cases} U(\tau_1 - \alpha, \tau_1 - \alpha), & \text{with probability } 1-p \\ U(\tau_2 - \alpha, \tau_2 - \alpha), & \text{with probability } p \end{cases}$$

In other words, instead of having fixed divergence points around some values, for instance at time τ_1 or τ_2, the Brownian motions diverge in one of the two "bands", each of the having length α. Obviously, such a *"jump distribution"* can be extended to have k bands with different lengths (α_k) as follows:

$$f(\tau) = \begin{cases} U(\tau_1 - \alpha_1, \tau_1 - \alpha_1), & \text{with probability } p_1 \\ U(\tau_2 - \alpha_2, \tau_2 - \alpha_2), & \text{with probability } p_2 \\ \vdots & \\ U(\tau_k - \alpha_k, \tau_k - \alpha_k), & \text{with probability} p_k \end{cases}$$

Moreover, other distributions apart from Uniforms can be used. However, the more complicated the form of the *"jump distribution"* is, the harder the estimation of any of the parameters will be. Note that although usually the length of the chosen band can be fixed, the most interesting parameters to draw inference for, are the probabilities p_i, $i = 1, \ldots, k$.

Preliminary work on such kind of *"jump distributions"* showed that the centered algorithm as shown in Section 5.8.2 leads to very slow mixing Markov chains. Therefore the need of better mixing algorithms is necessary. The strategies described in Section 5.9 improve the standard algorithms, especially when the location parameters, $\boldsymbol{X}(\tau)$ are integrated out. Nevertheless, we have already discussed that such an approach is very costly computationally when the size of the dataset increases. It would be very interesting to compare the performance of such an MCMC strategy by integrating the location parameters $\boldsymbol{X}(\tau)$ and the introduction of a non−centered parameterisation between π and τ_i. Summarizing, such generalisations of the *"jump distribution"* can be seen as an alternative way of defining a change-point model. Within the segments we allow for an additional level of covariance structure as defined via the LBT. A comparison between the results obtained from a standard change-point model and the above formulation would be of interest.

Appendix A

Appendix for Part II

A.1 On Minima of Random Variables

This section investigates the behavior of a random variable which is the minimum of n independent and identically distributed random variables. We focus on the distribution of the minimum and in particular on its expectation since this mainly describes the nature of a LBT.

We adopt the following the notation: Denote by $F_X(x)$ the distribution function of $X_1, \ldots, X_n$ and $f_X(x) = \partial F(x)/\partial x$ the corresponding probability density function. Let:

$$Z = \min(X_1, \ldots, X_n) = \min_{i=1}^{n} X_i$$

and we are interested in $F_Z(z)$ and $E[Z]$. Before considering any specified distribution F we require the following proposition:

Proposition 2 (Probability density function of minima) *Let $X_1, X_2, \ldots, X_n$, independent and identically distributed variables which follow the distribution function F. The distribution function of the random variable Z, where $Z = min(X_1, \ldots, X_n)$ is given by*

$$F_Z(z) = 1 - (1 - F_X(x))^n.$$

Proof:

$$\begin{aligned} F_Z(z) &= \Pr[\min(X_i) \leq z] = 1 - \Pr[\min(X_i) > z] = 1 - \prod_{i=1}^{n} (1 - \Pr[X_i > z]) \\ &= 1 - (\Pr[(X_i > z)])^n = 1 - (1 - F_X(x))^n \end{aligned}$$

□

The probability density function (pdf) of Z is derived by $\frac{\partial F_Z(z)}{\partial z}$:

$$f_Z(z) = n \cdot (1 - F_X(z))^{n-1} \cdot f_X(z) \tag{A.1}$$

A.1.1 Minimum of Uniform r.v. $[X \sim U(a, b)]$.

Let $X \sim U(a, b)$. Then $f_X(x) = 1/(b-a)$ and $F_X(x) = (x-a)/(b-a)$. Using the above proposition:

$$f_Z(z) = \frac{n}{(b-a)^n}(b-z)^{n-1} \tag{A.2}$$

For the special case of a standard Uniform distribution $a = 0, b = 1$ then:

$$f_Z(z) = n(1-z)^{n-1} \tag{A.3}$$

i.e. $Z \sim Beta(1, n)$ and therefore $E[Z] = 1/(n+1)$.

A.1.2 Minimum of Beta r.v. $[X \sim Beta(a, b)]$

We consider the special case where $a > 0, b = 1$, i.e. $X \sim Beta(a, 1)$ and $F_X(x) = x^a$ and $f_X(x) = ax^{a-1}$. Proposition A.1 shows that:

$$f_Z(z) = n \cdot a(1 - x^a)^{n-1} x^{a-1} \tag{A.4}$$

We need to calculate the expectation $E[Z] = \int_0^1 f_Z(z)dz$. This integral is evaluated by substitution ($v = x^a$) and therefore:

$$E[Z] = n \cdot B\left(\frac{a+1}{a}, n\right) = \frac{n}{a} \cdot \frac{(n+1/a)!}{(n-1)!}$$

It is easy to see that if $U_i \sim U(0,1)$ then the random variable $X_i = U_i^{1/a}$ follows a Beta distribution with parameters a and 1, i.e. $X_i \sim Beta(a,1)$. Therefore:

$$E[\min(X_1, \ldots, X_n)] = E[\min(U_1^{1/a}, \ldots, U_n^{1/a})] = E[(\min(U_1, \ldots, U_n)^{1/a}] = E[Z^{1/a}] \tag{A.5}$$

Making use of Jensen's inequality (see for example Ross, 1996, page 40) which states the following: *If the function f is convex then $E[f(x)] \geq f(E[x])$*, we get:

$$\begin{aligned} \text{If } a > 1 \quad &: \quad E[Z^{1/a}] \geq (E[Z])^{1/a} \geq \left(\frac{1}{n+1}\right)^{1/a} \\ \text{If } 0 < a < 1 \quad &: \quad E[Z^{1/a}] \geq (E[Z])^{1/a} \leq \left(\frac{1}{n+1}\right)^{1/a} \end{aligned}$$

Although (A.5) gives us the explicit form of E[Z] which can be useful, we are also interested in calculating the rate of its decay with respect to n, i.e $\lim_{n\to\infty} E[Z]$. We make use of Stirling's approximation formula (see for example Ross, 1996, page 144)

$$E[Z] \quad = \quad nB(1+1/a, n) = n\frac{\Gamma(n)\Gamma(1+1/a)}{\Gamma\left(\frac{a+1}{a}+n\right)} \propto \frac{n\Gamma(n)}{\Gamma\left(\frac{a+1}{a}+n\right)} \tag{A.6}$$

- $n\Gamma(n) = (n-1)! = \sqrt{2\pi} \cdot n^{n+1/2}e^{-n}$
- $\Gamma\left(n + \frac{a+1}{a}\right) = \sqrt{2\pi} \cdot \left(n + 1 + \frac{1}{a}\right)^{n+1+\frac{1}{a}-\frac{1}{2}} \cdot e^{-(n+1+\frac{1}{a})}$

Therefore we get:

$$E[Z] \quad \approx \quad \frac{n^{n+\frac{1}{2}} \cdot e^{-n}}{\left(n+1+\frac{1}{a}\right)^{n+\frac{1}{2}+\frac{1}{a}} \cdot e^{-n} \cdot e^{-1-\frac{1}{a}}} = \frac{n^{n+1/2}}{(n+1+1/a)^n} \cdot \frac{1}{n+1+\frac{1}{a}}^{\left(\frac{1}{2}+\frac{1}{a}\right)}$$

$$\approx \left(\frac{n}{n+\frac{a+1}{a}}\right)^{n} \cdot \left(\frac{n}{n+\frac{a+1}{a}}\right)^{1/2} \cdot \left(\frac{1}{n+\frac{a+1}{a}}\right)^{1/a}$$

$$\approx \left(\frac{1}{1+\left(\frac{a+1}{a}\right)/n}\right)^{n} \cdot \left(\frac{1}{1+\left(\frac{a+1}{a}\right)/n}\right)^{1/2} \cdot \left(\frac{1}{n+\frac{a+1}{a}}\right)^{1/a}$$

The first term:

$$\lim_{n\to\infty} \left(\frac{1}{1+\left(\frac{a+1}{a}\right)/n}\right)^{n} = e^{-(1+1/a)}$$

The second:

$$\lim_{n\to\infty} \frac{1}{1+\left(\frac{a+1}{a}\right)/n} = 1$$

Therefore, the rate of decay to zero of $E[Z]$ depends on the third term and that's implies that $E[Z]$ decays with rate:

$$O\left(\frac{1}{n^{1/a}}\right)$$

Note that the *Delta method* gives a similar result:

$$E[g(Z)] \approx g(E[Z]) = \left(\frac{1}{n+1}\right)^{1/a}$$

A.1.3 Minimum of Exponential r.v. $[X \sim Exp(\lambda)]$

Let $X_1, \ldots, X_n$ be i.i.d. $\sim Exp(\lambda)$ random variables truncated from 0 to 1. The probability density function (pdf) of such a random variable X_i and the cumulative distribution function are given below:

$$f_X(x) = \frac{\lambda}{1-e^{-\lambda}} \cdot e^{-\lambda x},\ 0 < x < 1$$

$$F_X(x) = \frac{1-e^{-\lambda x}}{1-e^{-\lambda}}\ 0 < x < 1$$

Using basic principles we realize that $f_Z(z)$ is not a well known distribution as

it was for the minimum of Beta$(a, 1)$ and moreover calculating the $E[Z]$ is not easily intractable. Instead, the inverse method (Ripley, 1987) is used to generate samples of an Exponential distribution using Uniforms. In other words,

$$\text{If } U_i \sim U(0,1), \text{ then } X_i = F_U^{-1}(u)$$

Therefore:

$$X_i = -\frac{1}{\lambda}\log\left(1 - U_i(1 - e^{-\lambda})\right) \sim Exp(\lambda),\ 0 < x < 1 \tag{A.7}$$

The expectation now becomes more tractable:

$$\begin{aligned} E[\min(X_1, \ldots X_n)] &= E\left[\min_{i=1}^{n}\left(-\frac{1}{\lambda}\log\left(1 - U_i(1 - e^{-\lambda})\right)\right)\right] \\ &= E\left[-\frac{1}{\lambda}\log\left(1 - \min U_i(1 - e^{-\lambda})\right)\right] \\ &= E\left[-\frac{1}{\lambda}\log\left(1 - Z(1 - e^{-\lambda})\right)\right] \end{aligned} \tag{A.8}$$

By making use of the *Delta method*:

$$E[g(Z)] \approx g(E[Z]) = -\frac{1}{\lambda}\log\left(\frac{n + e^{-\lambda}}{n+1}\right) \tag{A.9}$$

A.1.4 Special case

A random variable X with the following probability density and cumulative distribution function is considered:

$$f(x) = x^{-2}e^{1-\frac{1}{x}},\ \text{where } 0 < x < 1$$

$$F(x) = e^{1-1/x}$$

Such a random variable can be generated using the inversion method:

$$\text{If } U_i \sim U(0,1) \text{ then } X_i = \frac{1}{1-\log U_i} \sim F(x)$$

The expectation:

$$\begin{aligned} E[Z] &= E[\min(X_1,\ldots,X_n)] = E\left[\frac{1}{1-\log W}\right], \text{ where: } W \sim Beta(1,n) \\ &\approx \frac{1}{1+\log(n+1)}, \text{ since } E[W] = \frac{1}{n+1} \end{aligned}$$

Therefore, we concluded that the rate of decay of the expected value of Z is $1/\log n$.

A.1.5 Minimum of Bernoulli r.v. $[X \sim Bernoulli(p)]$

Let $X_1,\ldots,X_n$ i.i.d random variables. If we denote Z to be the random variable of the minimum then performing simple calculations like above we end up with the following result:

$$E[Z] = (1-p)^n.$$

Bibliography

Addy, C. L., Longini, I. M., and Haber, M. (1991). A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, 47(3):961–974.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.*, 74(1):47–97.

Alexandersen, S., Zhang, Z., Donaldson, A. I., and Garland, A. J. M. (2003). The pathogenesis and diagnosis of foot-and-mouth disease. *J. Comp. Pathol.*, 129(1):1–36.

Allen, L. J. S. and Burgin, A. M. (2000). Comparison of deterministic and stochastic SIS and SIR models in discrete time. *Math. Biosci.*, 163(1):1–33.

Amit, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multivariate Anal.*, 38(1):82–99.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York.

Anderson, R. M., Fraser, C., Ghani, A. C., Donnelly, C. A., Riley, S., Ferguson, N. M., Leung, G. M., Lam, T. H., and Hedley, A. J. (2004). Epidemiology, transmission dynamics and control of sars: the 2002-2003 epidemic. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 359(1447):1091–1105.

Anderson, R. M. and May, R. M. (1991). *Infectious diseases of humans:Dynamics and Control*. Oxford University Press, New York.

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, volume 151 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 2:1152–1174.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*. Hafner Press [Macmillan Publishing Co., Inc.] New York, second edition.

Bailey, N. T. J. and Thomas, A. S. (1971). The estimation of parameters from population data on the general stochastic epidemic. *Theroritical Pop. Biol.*, 2:253–270.

Ball, F. (1983). The threshold behaviour of epidemic models. *J. Appl. Probab.*, 20(2):227–241.

Ball, F. (1986). A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. in Appl. Probab.*, 18(2):289–310.

Ball, F. and Donnelly, P. (1995). Strong approximations for epidemic models. *Stochastic Process. Appl.*, 55(1):1–21.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Ann. Appl. Probab.*, 7(1):46–89.

Ball, F. G. and Lyne, O. D. (2002). Optimal vaccination policies for stochastic epidemics among a population of households. *Math. Biosci.*, 177/178:333–354. Deterministic and stochastic modeling of biointeraction (West Lafayette, IN, 2000).

Ball, F. G. and Lyne, O. D. (2006). Statistical inference for epidemics among a population of households. *In preparation.*

Bartlett, M. S. (1949). Some evolutionary stochastic processes. *J. Roy. Statist. Soc. Ber. B.*, 11:211–229.

Becker, N. (1979). An estimation procedure for household disease data. *Biometrika*, 66(2):271–277.

Becker, N. G. (1989). *Analysis of infectious disease data*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(2):287–307.

Becker, N. G., Britton, T., and O'Neill, P. D. (2003). Estimating vaccine effects on transmission of infection from household outbreak data. *Biometrics*, 59(3):467–475.

Becker, N. G. and Hasofer, A. M. (1997). Estimation in epidemics with incomplete observations. *J. Roy. Statist. Soc. Ser. B*, 59(2):415–429.

Bennett, K., Phillipson, J., Lowe, P., and Ward, N. (2001). *The impact of Foot and Mouth crisis on rural firms: A survey of microbusinesses in the North east of England*. Research Report. University of Newcastle.

Beran, J. (1992). Statistical methods for data with long range dependence. *Statistical Science*, 7(4):404–427.

Beran, J. (1994). *Statistics for long-memory processes*, volume 61 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York.

Bernaola-Galvan, P., Carpena, P., Roman-Roldan, R., and Oliver, J. L. (2002). Study of statistical correlations in DNA sequences. *Gene*, 300(1-2):105–115.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene*, 241:3–17.

Bernardo, J.-M. and Smith, A. F. M. (1994). *Bayesian theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Beskos, A., Papaspiliopoulos, O., and Roberts, Gareth. O. Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):333–382. with discussion.

Billard, L., Medley, G. F., and Anderson, R. M. (1990). The incubation period of aids virus. In Gabriel, J.-P., Lefevre, C., and Picard, P., editors, *Stochastic processes in epidemic theory. Lecture notes in Biomathematics*, pages 21–35. Springer-Verlag, Berlin.

Box, G. E. P. and Jenkins, G. M. (1970). *Times series analysis. Forecasting and control*. Holden-Day, San Francisco, Calif.

Britton, T. and Becker, N. G. (2000). Estimating the immunity coverage required to prevent epidemics in a community of households. *Biostatistics*, 1(4):389–402.

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

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods*. Springer Series in Statistics. Springer-Verlag, New York, second edition.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.*, 7(4):434–455.

Brooks, S. P. and Roberts, G. O. (1999). On quantile estimation and Markov chain Monte Carlo convergence. *Biometrika*, 86(3):710–717.

Chatfield, C. and Collins, A. J. (1980). *Introduction to multivariate analysis*. Chapman & Hall, London.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.*, 91(434):883–904.

Cox, D. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220. With discussion and a reply by the authors.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Daley, D. and Vere-Jones, D. (1988). *An introduction to the theory of point processes*. Sringer.

Daley, D. J. and Gani, J. (1999). *Epidemic modelling: an introduction*, volume 15 of *Cambridge Studies in Mathematical Biology*. Cambridge University Press, Cambridge.

Dawid, A. P. (1979). Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B*, 41(1):1–31.

Deardon, R., Brooks, S. P., Grenfell, B., Keeling, M. J., Tildesley, M. J. Savill, S. J., Shaw, D., and Woolhouse, M. E. J. (2006). Inference for individual-level models of infectious diseases in large populations. *Submitted*.

Demiris, N. and O'Neill, P. D. (2005). Bayesian inference for epidemics with two levels of mixing. *Scand. J. Statist.*, 32(2):265–280.

Demiris, N. and O'Neill, P. D. (2006). Computation of final outcome probabilities for the generalised stochastic epidemic. *Statistics and Computing*, 16(3):309 – 317.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38. With discussion.

Diekmann, O. and Heesterbeek, J. A. P. (2000). *Mathematical epidemiology of infectious diseases*. Wiley Series in Mathematical and Computational Biology.

John Wiley & Sons Ltd., Chichester. Model building, analysis and interpretation.

Dietz, K. and Schenzle, D. (1985). Mathematical models for infectious disease statistics. In *A celebration of statistics*, pages 167–204. Springer, New York.

Diggle, P. J. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Stat. Methods Med. Res.*, 15(4):325–336.

Ducatez, M. F., Olinger, C. M., Owoade, A. A., De Landtsheer, S., Ammerlaan, W., Niesters, H. G. M., Osterhaus, A. D. M. E., Fouchier, R. A. M., and Muller, C. P. (2006). Avian flu: multiple introductions of h5n1 in nigeria. *Nature*, 442(7098):37.

Enserink, M. (2006). Avian influenza. h5n1 moves into africa, european union, deepening global crisis. *Science*, 311(5763):932.

Ethier, S. N. and Kurtz, T. G. (1986). *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York. Characterization and convergence.

Eyre-Walker, A. and Hurst, L. D. (2001). The evolution of isochores. *Nat. Rev. Genet.*, 2:549–555.

Feller, W. (1971). *An introduction to probability theory and its applications*, volume II. Wiley and Sons, New York.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001a). The foot-and-mouth epidemic in great britain: Pattern of spread and impact of interventions. *Science*, 292(5519):1155–1161.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001b). Transmission intensity and impact of control policies on the foot and mouth epidemic in great britain. *Nature*, 413:542–547.

Ferguson, N. M., Keeling, M. J., Edmunds, W. J., Gani, R., Grenfell, B. T., Anderson, R. M., and Leach, S. (2003). Planning for smallpox outbreaks. *Nature*, 425(6959):681–685.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pages 287–302. Academic Press, New York.

Gaver, D. P. and Lewis, P. A. W. (1980). First-order autoregressive gamma sequences and point processes. *Adv. in Appl. Probab.*, 12(3):727–745.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika*, 82(3):479–488.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398–409.

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 Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov Chain methods. *IMA J. Math. Appl. Med. Biol.*, 15:19–40.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov chain Monte Carlo in practice.* Interdisciplinary Statistics. Chapman & Hall, London.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *J. Time Ser. Anal.*, 1(1):15–29.

Gray, H. L., Zhang, N.-F., and Woodward, W. A. (1989). On generalized fractional processes. *J. Time Ser. Anal.*, 10(3):233–257.

Grenfel, B. and Dobson, A. (1995). *Ecology of infectious diseases in natural populations.* Cambridge University Press, Cambridge.

Hamer, W. H. (1906). Epidemic disease in england. *The Lancet*, 1:733–739.

Hastings, W. (1970). Monte carlo sampling using markov chains and their applications. *Biometrika*, 57:97–109.

Hayakawa, Y., O'Neill, P. D., Upton, D., and Yip, P. S. F. (2003). Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles of several types. *Aust. N. Z. J. Stat.*, 45(4):491–502.

Heesterbeek, J. A. P. and Dietz, K. (1996). The concept of R_0 in epidemic theory. *Statist. Neerlandica*, 50(1):89–110.

Hills, S. E. and Smith, A. F. M. (1992). Parameterization issues in Bayesian inference. In *Bayesian statistics, 4 (Peñíscola, 1991)*, pages 227–246. Oxford Univ. Press, New York.

Hoehle, M. (2003). R_0 estimation by the martingale method. *Danish Institute of Agricultural Sciences, Biometry Research Unit.* Internal Report,.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, 68(1):165–176.

Isham, V. (1991). Assessing the variability of stochastic epidemics. *Math. Biosci.*, 107(2):209 – 224.

Isham, V. (1993). Stochastic models for epidemics with special reference to AIDS. *Ann. Appl. Probab.*, 3(1):1–27.

Isham, V. (2005). Stochastic models for epidemics. In *Celebrating Statistics: Papers in Honour of Sir David Cox on His 80th Birthday*, Oxford Statistical Science Series., pages 1–31. Oxford University Press.

Isham, V. and Medley, G. (1996). *Models for infectious human diseases: their structure and relation to data.* Cambridge University Press, Cambridge.

Jacobs, P. A. and Lewis, P. A. W. (1977). A mixed autoregressive-moving average exponential sequence and point process (EARMA 1, 1). *Advances in Appl. Probability*, 9(1):87–104.

Joe, H. (1996). Time series models with univariate margins in the convolution-closed infinitely divisible class. *J. Appl. Probab.*, 33(3):664–677.

Jørgensen, B. and Song, P. X.-K. (1998). Stationary time series models with exponential dispersion model margins. *J. Appl. Probab.*, 35(1):78–92.

Keeling, M. J. (1999). The effects of local spatial structure on epidemiological invasions. *Proc. Biol. Sci.*, 266(1421):859–867.

Keeling, M. J. (2005). Models of foot-and-mouth disease. *Proc. R. Soc. B*, 272:1195–1202.

Keeling, M. J. and Grenfell, B. T. (1998). Effect of variability in infection period on the persistence and spatial spread of infectious diseases. *Math. Biosci.*, 147(2):207–226.

Keeling, M. J. and Grenfell, B. T. (2000). Understanding the persistence of measles: reconciling theory, simulation and observation. *Proc. Roy. Soc. London*, B269:335–343.

Keeling, M. J., Woolhouse, M. E., May, R. M., Davies, G., and Grenfell, B. T. (2003). Modelling vaccination strategies against foot-and-mouth disease. *Nature*, 421(6919):136–142.

Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., and Grenfell, B. T. (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 McKendrick, A. G. (1927). Contributions to the mathematical theory of epidemics, part i. *Proc. Roy. Soc. London*, A115:700–721.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scand. J. Statist.*, 29(4):597–614.

Koopman, J. S., Simon, C. P., Jacquez, J. A., and Park, T. S. (1989). Selective contact within structured mixing with an application to HIV transmission risk from oral and anal sex. In *Mathematical and statistical approaches to AIDS epidemiology*, volume 83 of *Lecture Notes in Biomath.*, pages 316–348. Springer, Berlin.

Lamperti, J. (1962). Semi-stable stochastic processes. *Trans. Amer. Math. Soc.*, 104:62–78.

Lawrance, A. J. (1980). The mixed exponential solution to the first-order autoregressive model. *J. Appl. Probab.*, 17(2):546–552.

Lawrance, A. J. and Lewis, P. A. W. (1977). An exponential moving-average sequence and point process (EMA1). *J. Appl. Probability*, 14(1):98–113.

Lawrance, A. J. and Lewis, P. A. W. (1985). Modelling and residual analysis of nonlinear autoregressive time series in exponential variables. *J. Roy. Statist. Soc. Ser. B*, 47(2):165–202. With discussion.

Lefèvre, C. and Picard, P. (1993). An unusual stochastic order relation with some applications in sampling and epidemic theory. *Adv. in Appl. Probab.*, 25(1):63–81.

Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M. H., Fisman, D., and Murray, M. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300(5627):1966–1970.

Liu, J. S., Wong, W., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimatos and augmentation schemes. *Biometrika*, 81:27–40.

Lloyd, A. L. (2001). Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proc. Roy. Soc. London*, B268:985–993.

Longini, I. M. and Koopman, J. S. (1982). Household and community transmission parameters from final distributions of infections in households. *Biometrics*, 38(1):115–126.

Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev.*, 10:422–437.

McKendrick, A. G. (1926). Applications of mathematics to medical problems. *Proc. Edin. Math. Soc.*, 44:98–130.

McKenzie, E. (1982). Product autoregression: a time-series characterization of the gamma distribution. *J. Appl. Probab.*, 19(2):463–468.

Melo de Lima, C., Gueguen, L., Piau, D., and Gautier, C. (2005). Prediction of human isochores using a hidden markov model. *JOBIM*.

Meng, X.-L. and van Dyk, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B*, 59(3):511–567. With discussion and a reply by the authors.

Meng, X. L. and van Dyk, D. (2001). The art of Data Augmentation. *Journal of Computational and Graphical Statistics*, 10:1–50.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–191.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial point processes*, volume 100 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL.

Mollison, D. (1995). *Epidemic models: their structure and relation to data.* Cambridge University Press, Cambridge.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc.*, 78(381):47–65. With discussion.

Morris, M. (1991). A loglinear modelling framework for selective mixing. *Math. Biosci.*, 107:349–377.

Morris, M. (1996). Behavior change and non−homogenenous mixing. In Isham, V. and Medley, G., editors, *Models for infectious human diseases: their structure and relation to data*, pages 239–252. Cambridge University Press, Cambridge.

Morris, R. S., Wilesmith, J., Stern, M. W., Sanson, R. L., and Stevenson, M. A. (2001). Predictive spatial modelling of alternative control strategies for the foot-and-mouth disease epidemic in great britain, 2001. *Vet. Rec.*, 149:137–145.

Neal, P. and Roberts, G. (2005). A case study in non-centering for data augmentation: stochastic epidemics. *Stat. Comput.*, 15(4):315–327.

Neal, P. and Roberts, G. (2006). Optimal scaling for partially updating MCMC algorithms. *Ann. Appl. Probab.*, 16(2):475–515.

Neal, P. J. and Roberts, G. O. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, 5(2):249–261.

Neal, R. M. (2003). Density modeling and clustering using Dirichlet diffusion trees. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 619–629. Oxford Univ. Press, New York.

Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimisation. *Computer Journal*, 7:308–313.

Oliver, J. L., Carpena, P., Hackenberg, M., and Bernaola-Galvan, P. (2004). Isofinder: computational prediction of isochores in genome sequences. *Nucleic. Acids. Res.*, 32:287–292.

Oliver, J. L., Carpena, P., Roman-Roldan, R., Mata-Balaguer, T., Mejias-Romero, A., Hackenberg, M., and Bernaola-Galvan, P. (2002). Isochore chromosome maps of the human genome. *Gene*, 300(1-2):117–127.

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M., and Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. C*, 49(4):517–542.

O'Neill, P. D. and Becker, N. G. (2001). Inference for an epidemic where susceptibility varies. *Biostatistics*, 2(1):99–108.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A*, 162:121–129.

Papaspiliopoulos, O. (2003). *Non-centered parametrisations for hierarchical models and data augmentation.* PhD thesis, Department of Mathematics and Statistics, Lancaster University, Lancaster.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 307–326. Oxford Univ. Press, New York. With a discussion by Alan E. Gelfand, Ole F. Christensen and Darren J. Wilkinson, and a reply by the authors.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60:607–612.

Rida, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *J. Roy. Statist. Soc. Ser. B*, 53(1):269–283.

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L. M., Lam, T. H., Thach, T. Q., Chau, P., Chan, K. P., Lo, S. V., Leung, P. Y., Tsang, T., Ho, W., Lee, K. H., Lau, E. M.,

Ferguson, N. M., and Anderson, R. M. (2003). Transmission dynamics of the etiological agent of sars in hong kong: impact of public health interventions. *Science*, 300(5627):1961–1966.

Ripley, B. D. (1987). *Stochastic simulation*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B*, 59(2):291–317.

Roberts, G. O. and Tweedie, R. (2006). *Understanding MCMC*. Springer-Verlag.

Ross, R. (1916). An application of the theory of probabilities to the study of *a priori* pathometry, i. *Proc. Roy. Soc. London*, A92:204–230.

Ross, R. (1917a). An application of the theory of probabilities to the study of a *priori* pathometry, ii. *Proc. Roy. Soc. London*, A93:212–225.

Ross, R. (1917b). An application of the theory of probabilities to the study of a *priori* pathometry, iii. *Proc. Roy. Soc. London*, A93:215–240.

Ross, S. M. (1996). *Stochastic processes*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

Rue, H., Steinsland, I., and Erland, S. (2004). Approximating hidden Gaussian Markov random fields. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(4):877–892.

Sahu, S. and Roberts, G. (1999). On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing*, 9(1):55–64.

Sanson, R. L. (1993). *The developemnt of a decision support system for an animal disease emergency*. PhD thesis, Massey University, New Zealand.

Sanson, R. L. (1994). The epidemiology of foot-and-mouth disease: implications for New Zealand. *N. Z. Vet. J.*, 42(2):41–53.

Sanson, R. L., Morris, R. S., and Stern, M. W. (1999). Epiman-fmd: a decision support system for managing epidemics of vesicular disease. *Rev Sci Tech*, 18(3):593–605.

Sanson, R. L., Struthers, G., King, P., Weston, J. F., and Morris, R. S. (1993). The potential extent of transmission of foot-and-mouth disease: astudy of the movement of animals and materials in southland, new zealand. *N Z Vet J*, 41(1):21–28.

Sato, K. (1999). *Lévy processes and Inifinitely Divisible Distributions*. Cambridge University Press.

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E., Brooks, S. P., and Grenfell, B. T. (2006). Topographic determinants of foot and mouth disease transmission in the UK 2001 epidemic. *BMC Vet Res*, 2:3.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, 55(1):3–23.

Smith, N. G. C. and Lercher, M. J. (2002). Regional similarities in polymorphism

in the human genome extend over many megabases. *Trends Genet.*, 18(6):281–283.

Sokal, A. D. (1996). Monte carlo methods in statistical mechanics: Foundations and new algorithms. *Lecture at the Cargése Summer School on "Functional Integration: Basics and Applications"*.

Streftaris, G. and Gibson, G. J. (2004). Bayesian inference for stochastic epidemics in closed populations. *Stat. Model.*, 4(1):63–75.

Tanner, M. A. (1996). *Tools for statistical inference*. Springer Series in Statistics. Springer-Verlag, New York, third edition. Methods for the exploration of posterior distributions and likelihood functions.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.*, 82(398):528–550. With discussion and with a reply by the authors.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9.

Tildesley, M. J., Savill, N. J., Shaw, D. J., Deardon, R., Brooks, S. P., Woolhouse, M. E., Grenfell, B. T., and Keeling, M. J. (2006). Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the uk. *Nature*, 440(7080):83–86.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Whittle, P. (1955). The outcome of a stochastic epidemic—a note on Bailey's paper. *Biometrika*, 42:116–122.

Williams, T. (1971). An algebraic proof of the threshold theorem for the general stochastic epidemic (abstract). *Advances in Applied Probability*, 3:223.

Wilson, E. B. and Burke, M. H. (1942). The epidemic curve. I. *Proc. Nat. Acad. Sci. U. S. A.*, 28:361–366.

Wilson, E. B. and Burke, M. H. (1943). The epidemic curve. II. *Proc. Nat. Acad. Sci. U. S. A.*, 29:43–48.

Yu, P., Habtemariam, T., Wilson, S., Oryang, D., Nganwa, D., Obasa, M., and Robnett, V. (1997). A risk-assessment model for foot and mouth disease (FMD) virus introduction through deboned beef importation. *Prev. Vet. Med.*, 30:49–59.