Bayesian Inference for Stochastic Epidemic Models using Markov chain Monte Carlo Methods

by Nikolaos Demiris, BSc, MSc

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy, January 2004

Στους γονείς μου

Contents

1 Introduction to Stochastic Epidemic Models, Bayesian Statistics and Inference from Outbreak Data **1**

1.1 Introduction 1

1.1.1 Epidemic Modelling 2

1.1.2 Bayesian Inference 4

1.1.3 Inference from Outbreak Data using Epidemic Models . . 6

1.2 Stochastic Epidemic models 7

1.2.1 Generalised Stochastic Epidemic model 7

1.2.2 Epidemic models with two levels of mixing 11

1.3 Bayesian Statistical Inference 15

1.3.1 Basic theory 15

1.3.2 Bayesian Computation 20

1.4 Statistical Inference from Outbreak Data 25

1.4.1 The Nature of Infectious Disease Data 25

1.4.2 Previous Work on Epidemic Modelling 27

2 Exact Results for the Generalised Stochastic Epidemic **34**

2.1 Introduction 34

2.2 The Generalised Stochastic Epidemic 36
2.2.1 The Basic Model 36
2.2.2 Final Size Probabilities 37
2.3 Exact Final Size Probabilities 39
2.3.1 Multiple Precision Arithmetic 39
2.3.2 Multiple Precision Final Size Probabilities 40
2.4 Evaluation of limit theorems 41
2.4.1 Probability of Epidemic Extinction 41
2.4.2 Gaussian Approximation 43
2.5 Statistical Inference for the GSE 47
2.5.1 Bayesian Inference 47
2.5.2 MCMC algorithm 48
2.6 Results 49
2.7 Conclusion 56

3 Approximate Bayesian Inference for Epidemics with two levels of mixing 57

3.1 Introduction 57
3.2 Epidemic models with two levels of mixing 58
3.2.1 Stochastic Epidemic model 58
3.2.2 Final outcome of a homogeneous SIR epidemic with outside infection 59
3.2.3 Asymptotic approximations 62
3.3 Data and Augmented Likelihood 67
3.3.1 Final outcome data 67

3.3.2 Augmented Likelihood 68
3.4 Markov chain Monte Carlo algorithm 70
3.5 Application to data 72
3.5.1 Influenza outbreak data 72
3.5.2 Artificial final outcome data 82
3.5.3 Simulated final outcome data 86
3.6 Evaluation for the homogeneous case 91
3.6.1 Exact formula 92
3.6.2 Likelihood comparison for the GSE 93
3.6.3 Inference Comparison 96
3.7 Discussion 97

4 Bayesian Inference for Stochastic Epidemics using Random Graphs 99

4.1 Introduction 99
4.2 Epidemic models and Data 101
4.2.1 Stochastic Epidemic Models 101
4.2.2 Final Outcome Data 103
4.3 Random Graphs and the Likelihood 103
4.3.1 The Random Graph 104
4.3.2 The Graph Construction 106
4.3.3 The Likelihood 118
4.4 Markov chain Monte Carlo algorithm 124
4.4.1 The Independence Sampler 125
4.4.2 The Birth Death Sampler 127

4.5 Application to Data 130
4.5.1 Influenza Outbreak Data 132
4.5.2 Separate and Combined Influenza Data 141
4.5.3 Artificial Data 149
4.6 Evaluation using Exact Results 151
4.6.1 Perfect Data 151
4.6.2 Homogeneous Case 152
4.7 Discussion 154

5 Discussion and Future Work **157**

Bibliography **160**

Abstract

This thesis is concerned with statistical methodology for the analysis of stochastic SIR (Susceptible→Infective→Removed) epidemic models. We adopt the Bayesian paradigm and we develop suitably tailored Markov chain Monte Carlo (MCMC) algorithms. The focus is on methods that are easy to generalise in order to accomodate epidemic models with complex population structures. Additionally, the models are general enough to be applicable to a wide range of infectious diseases.

We introduce the stochastic epidemic models of interest and the MCMC methods we shall use and we review existing methods of statistical inference for epidemic models. We develop algorithms that utilise multiple precision arithmetic to overcome the well-known numerical problems in the calculation of the final size distribution for the generalised stochastic epidemic. Consequently, we use these exact results to evaluate the precision of asymptotic theorems previously derived in the literature. We also use the exact final size probabilities to obtain the posterior distribution of the threshold parameter R_0.

We proceed to develop methods of statistical inference for an epidemic model with two levels of mixing. This model assumes that the population is partitioned into subpopulations and permits infection on both local (within-group) and global (population-wide) scales. We adopt two different data augmentation algorithms. The first method introduces an appropriate latent variable, the *final severity*, for which we have asymptotic information in the event of an outbreak among a population with a large number of groups. Hence, approximate inference can be performed conditional on a "major" outbreak, a common assumption for stochastic processes with threshold behaviour such as epidemics and branching processes.

In the last part of this thesis we use a *random graph* representation of the

epidemic process and we impute more detailed information about the infection spread. The augmented state-space contains aspects of the infection spread that have been impossible to obtain before. Additionally, the method is exact in the sense that it works for any (finite) population and group sizes and it does not assume that the epidemic is above threshold. Potential uses of the extra information include the design and testing of appropriate prophylactic measures like different vaccination strategies. An attractive feature is that the two algorithms complement each other in the sense that when the number of groups is large the approximate method (which is faster) is almost as accurate as the exact one and can be used instead. Finally, it is straightforward to extend our methods to more complex population structures like overlapping groups, small-world and scale-free networks

Acknowledgements

The work presented in this thesis was carried out under the direction of Philip O'Neill. Phil has been extremely patient and encouraging throughout my PhD. Working with him was a great pleasure and I am grateful to him for our discussions about epidemics and a lot more, too many to be mentioned here.

I feel lucky that I met Petros Dellaportas during my MSc studies in Athens. He inspired me to work on statistics and he suggested that I come to the UK where I had the pleasure to meet a lot of very interesting people. Pete Neal was particularly helpful during my introduction to stochastic epidemics and Gareth Roberts gave me a lot of useful advice both about statistics and life in England. Conversations with Frank Ball, Niels Becker, Tom Britton, Gavin Gibson, Owen Lyne and Sergey Utev have been particularly useful and I thank them all.

Many people made my life in Nottingham more enjoyable. My girlfriend Vanessa has been an incredible partner in many ways. I am delighted that I met Bill, Dimos, Fred, George, Jesus, Jon, Juan, Julian, Ilias, Laura, Omiros and Pericle, sharing time with them was always very rewarding.

I would also like to thank my teammates in the Postgrads basketball team, the people in the Mathematics department and in Florence Boot Hall, particularly Bernard and Pat.

Last but not least, my family was extremely supportive throughout this period, despite me being abroad! A big thanks to them.

I am grateful to the UK Engineering and Physical Sciences Research Council for financial support.

List of Figures

2.1 The final size distribution of an epidemic among a population of 800 individuals. 38

2.2 The final size distribution of an epidemic among a population of 1000 individuals and the corresponding Gaussian approximation. 43

2.3 The scaled final size distribution of an epidemic among a population of 1000 individuals and the corresponding Gaussian approximation. 45

2.4 The standardised final size probabilities of an epidemic among a population of 1000 individuals and the corresponding Gaussian approximation. 47

2.5 Posterior density of R_0 for the three different infectious periods when $x = 30$. 51

2.6 Posterior density of R_0 for the three different infectious periods when $x = 60$. 53

2.7 Posterior density of R_0 for the two different priors. 54

2.8 Plot of the autocorrelation function based on the posterior output of R_0. 54

2.9 Trace plot of the posterior output of R_0. 55

3.1 Posterior density of R_* for two different priors. 73
3.2 Posterior trace plot of λ_L. 75
3.3 Plot of the autocorrelation function for λ_L. 75
3.4 Posterior density of λ_G as α varies. 76
3.5 Posterior density of λ_L as α varies. 77
3.6 Posterior density of R_* as α varies. 78
3.7 Scatterplot of λ_L and λ_G as α varies. 79
3.8 Graphical comparison of the likelihood for the three different methods. 96

4.1 The output for the total number of links for the smallpox dataset. 132
4.2 The posterior density of λ_G for the two different priors. 133
4.3 The posterior autocorrelation function of λ_G. 135
4.4 Trace of the posterior density of λ_L from the Random Graph algorithm. 136
4.5 Trace of the posterior density of λ_G from the Random Graph algorithm. 137
4.6 Trace of the posterior density of the total number of local out-degrees. 138
4.7 Trace of the posterior density of the total number of global out-degrees. 139
4.8 The posterior density of λ_L from the Random Graph and the Severity algorithms. 140
4.9 The posterior density of R_* from the Random Graph and the Severity algorithms. 141

4.10 Scatterplot of λ_L and λ_G for the Tecumseh data when the infectious period follows a Gamma distribution. 142

4.11 The Posterior distribution of the mean local and global degree for the Tecumseh data. 143

4.12 Posterior Density of λ_G for the two separate and the combined Tecumseh outbreaks. 145

4.13 Posterior Density of λ_L for the two separate and the combined Tecumseh outbreaks. 146

4.14 Posterior Density of R_* for the two separate and the combined Tecumseh outbreaks. 147

4.15 Posterior Density of λ_G for the two artificial datasets. We denote with Data10 the dataset where all the values are multiplied by 10.148

4.16 Posterior Density of λ_L for the two artificial datasets. We denote with Data10 the dataset where all the values are multiplied by 10.149

4.17 Posterior Density of R_0 using the random graph algorithm and the algorithm based on the multiple precision solution of the triangular equations. 153

List of Tables

2.1 Posterior summary statistics for the three different infectious periods when 30 out of 120 individuals are ultimately infected. . 50

2.2 Posterior summary statistics for the three different infectious periods when 60 out of 120 individuals are ultimately infected. . 51

3.1 The Tecumseh influenza data 73

3.2 Posterior parameter summaries from MCMC algorithm using the Tecumseh dataset in the case that $\alpha = 1$. 79

3.3 Posterior parameter summaries from MCMC algorithm using the Tecumseh dataset in the case that $\alpha = 0.1$. 80

3.4 Posterior parameter summaries from MCMC algorithm using the Tecumseh dataset in the case that $\alpha = 10^{-5}$. 82

3.5 The dataset of the first example 83

3.6 Posterior parameter summaries for the data in Example 1. . . . 83

3.7 The dataset of the second example 84

3.8 Posterior parameter summaries for the data in Example 2 using the a Normal random walk proposal. 85

3.9 The "perfect" dataset 87

3.10 Posterior parameter summaries for the "perfect" data. 88

3.11 The second simulated dataset 88
3.12 Posterior parameter summaries for the second simulated dataset. 89
3.13 The third simulated dataset 90
3.14 Posterior parameter summaries for the third simulated dataset. . 90

4.1 Posterior parameter summaries for the Influenza dataset with households sizes truncated to 5 and with a Gamma distributed infectious period. 133
4.2 The 1977-1978 Tecumseh influenza data 140
4.3 The 1980-1981 Tecumseh influenza data 142
4.4 Posterior parameter summaries for the full 1977-1978 Influenza dataset. 143
4.5 Posterior parameter summaries for the full 1980-1981 Influenza dataset. 144
4.6 Posterior parameter summaries for the combined 1977-1978 and 1980-1981 Influenza datasets with household sizes up to 7. . . . 144
4.7 The Artificial dataset 147
4.8 Posterior parameter summaries for the perfect data divided by 10 and rounded to the closest integer. The true values are $\lambda_L = 0.06$ and $\lambda_G = 0.23$. 151
4.9 Posterior parameter summaries for the smallpox data. 152

Chapter 1

Introduction to Stochastic Epidemic Models, Bayesian Statistics and Inference from Outbreak Data

1.1 Introduction

In this thesis we will describe methods for Bayesian statistical inference for stochastic epidemic models. The focus will be on general methodology for the analysis of an epidemic model where the population is partitioned into groups. However, the approach can often be extended to more complex, and realistic population structures. The different methods are illustrated using both real life and simulated outbreak data.

This chapter serves as an introduction to the main themes that this thesis uses. Stochastic epidemic models are appropriate stochastic processes that can

be used to model disease propagation. Two processes of this kind are presented and their behaviour is outlined. Subsequently we give a brief introduction to Bayesian inference, which is the paradigm we shall follow throughout the thesis, and the modern computational tools used to facilitate the analysis of realistically complex models. The last part of this chapter contains a short review of the analysis of infectious disease data using stochastic epidemic models.

1.1.1 Epidemic Modelling

Stochastic and Deterministic Models

We will focus on homogeneous and heterogeneous stochastic epidemic models. Disease propagation is an inherently stochastic phenomenon and there is a number of reasons why one should use stochastic models to capture the transmission process. Real life epidemics, in the absence of intervention from outside, can either go extinct with a limited number of individuals getting ultimately infected, or end up with a significant proportion of the population having contracted the disease in question. It is only stochastic, as opposed to deterministic, models that can capture this behaviour and the probability of each event taking place. Additionally, the use of stochastic epidemic models naturally facilitates estimation of important epidemiological parameters as will become apparent in the following chapters. Finally, from a subjective point of view, stochastic models are intuitively logical to define, since they naturally describe the contact processes between different individuals. However, the need for realistically complex models has made deterministic models more popular, since it is possible to analyse numerically quite elaborate deterministic models. Hence, it appears reasonable that effort should go towards developing general methods of statistical analysis that can be applied to complex stochastic mod-

els.

Modelling Disease Propagation

In recent years there has been increasing interest in the use of stochastic epidemic models for the analysis of real life epidemics. The need for accurate modelling of the epidemic process is vital, particularly because the financial consequences of infectious disease outbreaks are growing, two important recent examples being the 2001 foot and mouth disease (FMD) outbreak in the UK and the severe acute respiratory syndrome (SARS) epidemic in the spring of 2003. For modelling of these high impact epidemics see Ferguson *et al.* (2001) and Keeling *et al.* (2001) for FMD and Riley *et al.* (2003) and Lipsitch *et al.* (2003) for SARS.

In order to prevent, or at least reduce, infection spread we need models that can accurately capture the main characteristics of the disease in question since understanding disease propagation is vital for the most effective reactive measures. Additionally, if we want to adopt a proactive approach and model vaccination strategies, we need methodology for performing statistical inference for the parameters of epidemiological interest. Hence, it readily becomes apparent that it is vital that epidemic models of general applicability and methodology for their statistical analysis should be developed.

Two Stochastic Epidemics

In this chapter we will describe two stochastic epidemic models. The so-called generalised stochastic epidemic is a rather simple model defined on a homogeneous and homogeneously mixing population. It is called generalised because the infectious period i.e., the time that an infective individual remains infectious, can have any specified distribution. The special case where the infectious

period follows an exponential distribution makes the model Markovian and is known as the general stochastic epidemic (e.g. Bailey (1975) chapter 6). Note that certain non exponential infectious periods (like Gamma with integer shape parameter) can be incorporated in a Markovian model using additional compartments. However, the generalised stochastic epidemic is a unified process, even for infectious period distributions that cannot be written as the sum (or linear combination) of exponentials.

Subsequently, we describe a more complex model where the population is partitioned into groups and infectives have contacts both within and between the group. This model is motivated by a desire for additional realism since it is well known that disease spread is greatly facilitated in groups such as schools and households. The main reference for this so-called *epidemic model with two levels of mixing* is Ball *et al.* (1997). The generalised stochastic epidemic is a special case of the two-level-mixing model when all the households are of size one and we will evaluate our methods for this special case. Methods for statistical inference for epidemics will be reviewed in section four of this chapter. We shall now give a short introduction to Bayesian inference and the modern computational methods used for the implementation of the Bayesian paradigm.

1.1.2 Bayesian Inference

Introduction

Bayesian inference, similarly to likelihood inference requires a sampling model that produces the *likelihood*, the conditional distribution of the data given the model parameters. Additionally, the Bayesian approach will place a *prior* distribution on the model parameters. The likelihood and the prior are then

combined using Bayes' theorem to compute the *posterior* distribution. The posterior distribution is the conditional distribution of the unknown quantities given the observed data and is the object from which all Bayesian inference arises.

We shall now introduce some notation. The model parameters are described with the (potentially multi-dimensional) random variable $\boldsymbol{\theta}$. From the Bayesian perspective, model parameters and data are indistinguishable, the only difference being that we possess a realisation of $\mathbf{X}$, the observed data $\mathbf{X} = \mathbf{x}$. The frequentist and Bayesian approaches, despite arising from different principles do not necessarily give completely dissimilar answers. In fact, they can be connected in a decision-theoretic framework through *preposterior* evaluations (see Rubin, 1984). In this thesis we will adopt the Bayesian paradigm which, while theoretically simple and more intuitive than the frequentist approach, requires evaluation of complex integrals even in fairly elementary problems.

Modern Bayesian Statistics

The use of Bayesian methods in applied problems has exploded during the 1990s. The availability of fast computing machines was combined with a group of iterative simulation methods known as Markov chain Monte Carlo (MCMC) algorithms that greatly aided the use of realistically complex Bayesian models. The idea behind MCMC is to produce approximate samples from the posterior distribution of interest, by generating a Markov chain which has the posterior as its limiting distribution. This revolutionary approach to Monte Carlo was originated in the particle Physics literature in Metropolis *et al.* (1953). It was then generalised by Hastings (1970) to a more statistical setting. However, it was Gelfand and Smith (1990) that introduced MCMC methods to mainstream statistics and since then, the use of Bayesian methods for applied statistical

modelling has increased rapidly.

A comprehensive account of MCMC-related issues and the advances in statistical methodology generated by using this set of computational tools until 1995 is provided in Gilks *et al.* (1996). A contemporary similar attempt would be almost impossible since the use of MCMC has enabled the analysis of many complex models in the vast majority of the statistical application areas. In an introductory technical level, Congdon (2001) describes the analysis of a wide range of statistical models using BUGS, freely available software for Bayesian Inference using MCMC, see Spiegelhalter *et al.* (1996). Many of these models, including generalised linear mixed models, can only be approximately analysed using classical statistical methodology. Conversely, it is straightforward to analyse models of this complexity using routine examples of BUGS.

1.1.3 Inference from Outbreak Data using Epidemic Models

The Need for Epidemic Modelling

The statistical analysis of infectious disease data usually requires the development of problem-specific methodology. There is a number of reasons for this but the main features that distinguish outbreak data are the high dependence that is inherently present and the fact that we can never observe the entire infection process. In many cases the data from the incidence of an infectious disease consist of only the final numbers of infected individuals. Hence, the analysis should take into account all the possible ways that these individuals could be infected. Moreover, even when the data contain the times that the symptoms occur, we cannot observe the actual infection times. Also the true epidemic chain, i.e. who infects who, is typically not observed either.

These reasons suggest that in order to accurately analyse outbreak data, we need a model that describes a number of aspects of the underlying infection pathway. Hence, inference about the data generating process can provide us with an insight about the quantitative behaviour of the most important features of the disease propagation. Additionally, the design of control measures against a disease can be improved through a quantitative analysis based on an epidemic model.

The rest of the chapter is organised as follows. The two stochastic epidemic models and related results that we use throughout the thesis are presented in section 2. In section 3 we first give a short introduction to Bayesian theory while in the remainder of the section we present the main computational tools required for the implementation of the Bayesian paradigm. The chapter concludes with known statistical methodology for inference from infectious disease data.

1.2 Stochastic Epidemic models

1.2.1 Generalised Stochastic Epidemic model

Epidemic model

We describe a simple model for the transmission of infectious diseases where the population is assumed to be closed, homogeneous and homogeneously mixing. We define as closed a population that does not contain demographic changes. Hence, we assume that during the course of the epidemic no births or immigrations occur. We also assume homogeneity of the population in the sense that the individuals belong in the same group and each pair of individuals has the same degree of social contacts with each other. This assumption will be relaxed

later when we will assume that the population is partitioned into groups and individuals will have additional within-group contacts.

The population consists of n individuals out of which m are initially infected and they are able to have *close contacts* i.e., contacts that result in infection, with other individuals of the population. The remaining $n-m$ individuals are assumed to be initially susceptible and can be potentially infected by the m initial infectives. The infectious periods of different infectives are assumed to be independent and identically distributed according to the distribution of a random variable I, which can have any arbitrary but specified distribution.

While infectious, an individual makes contacts with each of the n individuals of the population at times given by the points of a homogeneous Poisson process with intensity $\frac{\lambda}{n}$. The contacts result in immediate infection of the susceptible individual that the infective has contacted. The infectious individual is removed from the infection process once its infectious period terminates. A removed individual can be dead, in case of a fatal disease, or recovered and immune to further infections. The Poisson processes of different individuals are assumed to be mutually independent. The epidemic ceases as soon as there are no infectives present in the population.

Epidemic models of this kind, where an individual is allowed to be in any of the three states, susceptible, infective or removed, are often called S-I-R epidemics. The special (Markovian) case where the infectious period follows an exponential distribution is known as the *general stochastic epidemic*. The assumption of an exponential infectious period is mathematically (and not biologically) motivated since it makes the probabilistic and statistical analysis of the model simpler, see O'Neill and Becker (2002) and Streftaris and Gibson (2004) for applications on diseases with Gamma and Weibull distributed infectious periods respectively. The general stochastic epidemic was originated by

Bartlett (1949) and has received a lot of attention in the probabilistic literature. However, it has been generalised in a large number of ways and we shall describe later in this section exact results for the case with a general infectious period.

Basic reproduction number

The most important parameter in epidemic theory is the *basic reproduction number* R_0 (Dietz, 1993) defined as the expected number of infections generated by a "typical" infected individual in a large susceptible population. In the generalised stochastic epidemic a typical individual can be any of the infectives since the model is homogeneous and homogeneously mixing. In more complicated models the definition of a typical individual is not straightforward and care is required in the definition of an appropriate threshold parameter. We call R_0 a threshold parameter since the value of R_0 determines whether or not a "major" epidemic can occur. Specifically, when $R_0 \leq 1$ the epidemic will die out i.e., in an infinite population only a finite number of individuals will ultimately become infected. In the case that $R_0 > 1$ there is a positive probability that an infinitely large number of individuals will contract the disease in question.

The threshold theorem, described in the previous section, is the most important result in the mathematical theory of epidemics and it was introduced in Whittle (1955), see also Williams (1971) and Ball (1983). We will present in the next section a rigorous derivation of the threshold parameter based on a coupling of the initial stages of the epidemic with a branching process. For the model presented here, $R_0 = \lambda E[I]$. We emphasize that the definition of R_0 as a threshold parameter and the related results are exactly valid only in some asymptotic sense, typically as the population size becomes infinite. How-

ever, this is the most commonly used epidemiological parameter to date and reducing R_0 below unity is typically the aim of programs for epidemic control.

Final size distribution

We shall consider the case where only the final outcome of the epidemic is observed. The *final size* of an epidemic is defined as the number of initially susceptible individuals that ultimately become infected. Let $\phi(\theta) = E(\exp(-\theta I)), \theta > 0$ be the moment generating function of the infectious period I and p_k^n the probability that the final size of the epidemic is equal to k, $0 \leq k \leq n$. Then Ball (1986) proved that

$$\sum_{k=0}^{l} \frac{\binom{n-k}{l-k} p_k^n}{\left[\phi\left(\frac{\lambda(n-l)}{n}\right)\right]^{k+m}} = \binom{n}{l}, \qquad 0 \leq l \leq n. \tag{1.1}$$

The system of equations in (1.1) is triangular and thus, in principle, it is straightforward to calculate the final size probabilities recursively. However, numerical problems appear due to rounding errors even for moderate population sizes of order 50-100. Hence it readily becomes apparent that the calculation of the likelihood, the distribution of the data given a parameter value, requires the development of a different method. In this thesis we will employ two ways to overcome these difficulties. In chapter two we will evaluate the likelihood using augmented precision arithmetic while in chapter four we shall use a random graph that enables the evaluation of the likelihood. We will now describe a more realistic, and complex, model for disease spread in a closed population.

1.2.2 Epidemic models with two levels of mixing

Basic model

In this section we introduce the two-level-mixing model. The statistical analysis of this model will be described in later chapters. In this chapter we will define the model and give an approximation for the early stages of an epidemic in a population with *local* and *global* contacts. The relevant results that are required for inference purposes will be described in chapter three.

Population Structure We consider the model described in Ball *et al.* (1997). The model is defined in a closed population that is partitioned into groups (e.g. households or farms) of varying sizes. Suppose that the population contains m_j groups of size j and let $m = \sum_{j=1}^{\infty} m_j$ be the total number of groups. Then the total number of individuals is $N = \sum_{j=1}^{\infty} j m_j$.

Epidemic Process We will make the S-I-R assumption so that each individual can, at any time $t \geq 0$, be susceptible, infectious or removed. A susceptible individual j may become infectious as soon as he is contacted by an infective and will remain so for a time I_j distributed according to any specified non-negative random variable I. The epidemic is initiated at time $t = 0$ by a (typically small) number of individuals while the rest of the population is initially susceptible. We allow individuals to mix at two levels. Thus, while infective, an individual makes population wide infectious contacts at times given by the points of a Poisson point process with rate λ_G. Each such contact is with an individual chosen uniformly at random from the N initially susceptible individuals. Hence, the person to person rate is $\frac{\lambda_G}{N}$. If the contacted individual has been infected before then the contact has no effect to the state of this individual while if the contacted person is susceptible then he gets infective.

Additionally, each infective individual makes person to person contacts with any given susceptible in its own household according to a Poisson process with rate λ_L. All the Poisson processes (including the two processes associated with the same individual) and the random variables $I_j, j = 1, \ldots, N$, describing the infectious periods of different individuals, are assumed to be mutually independent. Note here that by *contact* we mean the so-called close contacts that result in the immediate infection of the susceptible. At the end of its infectious period the individual is removed and plays no further role in the epidemic spread. The epidemic ceases when there are no infectives present in the population.

Latent Period Note that this model does not assume a latent period for an infected individual. However, the distribution of the final outcome of an SIR epidemic is invariant to fairly general assumptions concerning a latent period, see Ball *et al.* (1997). One way to see this is to consider the infection process in terms of "generations" of infectives. This is not always accurate for the propagation of a disease when temporal data about the epidemic spread are available, but there is no loss of generality when we consider final outcome data. This can be seen by considering the random graph associated with the epidemic and will become more clear when we will consider the construction of the random graph in chapter four. In what follows we describe some results regarding the coupling of the initial stages of the epidemic process with a suitable branching process.

Branching process approximation

The threshold parameter is of considerable practical importance since it gives information about epidemic control through prophylactic measures like vaccination. For stochastic epidemic models threshold parameters are typically defined

as functions of the basic model parameters and the population structure.

The probabilistic properties of the two-level-mixing model are analysed in Ball *et al.* (1997). The authors derive, among other limit theorems, a threshold result using a coupling argument. Specifically, assuming there is a population with infinitely many households, the initial stages of the epidemic are coupled with a branching process (see e.g. Jagers (1975)). The state-space of this branching process is the set of groups, with each group acting as a "superindividual". Thus, the early phase of the epidemic is coupled with a suitable stochastic process for which there is a large amount of theory available. Subsequently the authors prove, as the number of households goes to infinity, that during the early stages of the epidemic the probability of a global infectious contact with a member of an infected household is negligible.

Let us assume that for $n = 1, 2, \ldots$, the proportion $\frac{m_n}{m}$ of groups of size n converges to θ_n as the population size $N \to \infty$. Let also $\hat{g} = \sum_{n=1}^{\infty} n\theta_n$ be the asymptotic mean group size and assume that $\hat{g} < \infty$. Then it is shown in section 3.5 of Ball *et al.* (1997) that there exists, as the number of households goes to infinity, a threshold parameter defined by $R_* = \lambda_G E(I)\nu$. Here $\nu = \nu(\lambda_L) = \frac{1}{\hat{g}} \sum_{n=1}^{\infty}(1 + \mu_{n-1,1}(1))n\theta_n$ is the mean size of an outbreak in a group, started by a randomly chosen individual, in which only local infections are permitted and the initial infective in the group is included in ν. Also $\mu_{n-1,1}(1)$ is the mean final size of an epidemic in a group with a single initial infective and $n-1$ initial susceptibles where only local infections count. This quantity will be evaluated later using equation (3.4).

For simpler models this parameter would typically be the so-called basic reproduction number. However, for complex models it is not always straightforward to define the basic reproduction number. Thus, we will be referring to R_* as the threshold parameter. In the case of a homogeneously mixing popu-

lation, which in the current framework corresponds to all the households being of size 1, the threshold parameter would be $R_0 = \lambda_G E(I)$.

R_* is the threshold parameter that determines the behaviour of the coupled branching process. Hence, by standard branching process theory, if $R_* \leq 1$ the branching process goes extinct, or equivalently, the epidemic will die out with probability 1. The epidemic extinction is defined in the asymptotic sense as mentioned in the previous section. Hence, in an infinite number of households, only a finite number of households will ultimately contain infected individuals. In the case of $R_* > 1$ there is a positive probability that a major epidemic will occur. Thus, in a rigorous treatment of the non-extinction case, out of an infinite number of groups, a positive proportion of them will get infected from outside. The interpretation of the above results in terms of applications is that if one wants to prevent major epidemics using vaccination or other means of control, it will be necessary to keep R_* below unity. Thus, it quickly becomes apparent that the estimation of the transmission parameters of the model is vital.

Approximating the initial stages of the epidemic is one of the two most important types of limit theorems in the epidemic theory. The second type describes results related to a normal approximation for the final size of an outbreak in the event of a major epidemic. Results of this kind for the two-level-mixing model will be described in chapter three, where a central limit theorem will be used for approximate statistical inference for this model.

This thesis utilises the Bayesian approach to statistical inference and in the next section we give a summary of the theory and the tools required for its implementation.

1.3 Bayesian Statistical Inference

1.3.1 Basic theory

In this section we will review the fundamentals of the Bayesian paradigm in a basic non-technical level. For a rigorous and detailed approach see Bernardo and Smith (1994).

Bayes' Theorem

In the Bayesian approach, in addition to specifying the model for the observed data $\mathbf{x} = (x_1, \ldots, x_n)$ given the vector of the unknown parameters $\boldsymbol{\theta}$, in the form of the likelihood function $L(\mathbf{x} \mid \boldsymbol{\theta})$, we also define the *prior* distribution $\pi(\boldsymbol{\theta})$. Inference concerning $\boldsymbol{\theta}$ is then based on its *posterior* distribution, given by

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{L(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto L(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.2}$$

We refer to this formula as *Bayes' Theorem*. The integral in the denominator is essentially a normalising constant and its calculation has traditionally been a severe obstacle in Bayesian computation. We shall demonstrate in the next section how we can avoid its calculation using MCMC methods. The second form in 1.2 can be thought of as "the posterior is proportional to the likelihood times the prior". Clearly the likelihood may be multiplied by any constant (or any function of $\mathbf{x}$ alone) without altering the posterior. Moreover, Bayes' Theorem may also be used sequentially: suppose we have collected two independent data samples, $\mathbf{x_1}$ and $\mathbf{x_2}$. Then

$$\begin{aligned}\pi(\boldsymbol{\theta} \mid \mathbf{x_1}, \mathbf{x_2}) &\propto L(\mathbf{x_1}, \mathbf{x_2} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= L_2(\mathbf{x_2} \mid \boldsymbol{\theta})L_1(\mathbf{x_1} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto L_2(\mathbf{x_2} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \mathbf{x_1}). \end{aligned} \tag{1.3}$$

That is, we can obtain the posterior for the full dataset $(\mathbf{x_1}, \mathbf{x_2})$ by first evaluating $\pi(\boldsymbol{\theta} \mid \mathbf{x_1})$ and then treating it as the prior for the second dataset $\mathbf{x_2}$. Thus, we have a natural setting when the data arrive sequentially over time.

Prior distributions

In this section we briefly present the most popular approaches for the choice of a prior distribution. Additionally to the priors we mention here there exist the so called elicited priors, created using an expert's opinion. However, elicitation methods go beyond the scope of this thesis and we shall not give more details here.

Conjugate priors When choosing a prior from a parametric family, some choices may be more computationally convenient than others. In particular, it can be 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. It is shown in Morris (1983) 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 use of MCMC does not require the specification of conjugate priors. However, they can be computationally convenient and their use is recommended when it is possible and appropriate.

Non-informative priors In many practical situations no reliable prior information concerning $\boldsymbol{\theta}$ exists, or inference based solely on the data is desirable.

In this case we typically wish to define a prior distribution $\pi(\boldsymbol{\theta})$ that contains no information about $\boldsymbol{\theta}$ in the sense that it does not favour one $\boldsymbol{\theta}$ value over another. We may refer to a distribution of this kind as a *noninformative prior* for $\boldsymbol{\theta}$ and argue that the information contained in the posterior about $\boldsymbol{\theta}$ stems from the data only.

In the case that the parameter space is $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_n\}$, i.e., discrete and finite, then the distribution

$$\pi(\theta_i) = \frac{1}{n}, i = 1, \ldots, n,$$

places the same prior probability to any candidate θ value. Likewise, in the case of a bounded continuous parameter space, say $\boldsymbol{\Theta} = [a, b], -\infty < a < b < \infty$, then the uniform distribution

$$\pi(\theta) = \frac{1}{b-a}, a < \theta < b,$$

appears to be noninformative.

For unbounded spaces the definition of noninformative distribution is not straightforward. In the case that $\boldsymbol{\Theta} = (-\infty, \infty)$ a distribution like $\pi(\theta) = c$ is clearly *improper* since $\int \pi(\theta) d\theta = \infty$. However, Bayesian inference is still possible in the special case where $\int Ł(\mathbf{x} \mid \theta) d\theta = D < \infty$. Then

$$\pi(\theta \mid \mathbf{x}) = \frac{L(\mathbf{x} \mid \theta)c}{\int L(\mathbf{x} \mid \theta)cd\theta} = \frac{L(\mathbf{x} \mid \theta)}{D}.$$

There is not however a "default" prior for all cases. The uniform prior is *not invariant* under reparameterisation. Thus, an uninformative prior can be converted, in the case of a different model, to an informative one. One approach that overcomes this difficulty is *Jeffreys prior* given by

$$\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2},$$

where $|\cdot|$ denotes the determinant and $I(\boldsymbol{\theta})$ is the expected Fisher information matrix, having ij-element

$$I_{ij}(\boldsymbol{\theta}) = E_{\mathbf{X}|\boldsymbol{\Theta}}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\mathbf{x} \mid \boldsymbol{\theta})\right].$$

In this thesis we will adopt the view of Box and Tiao (1973, p.23) who suggest that all that is important is that the data dominate whatever information is contained in the prior, since as long as this happens, the precise form of the prior is not important. Hence, we typically employ a few different priors with large variance and as long as the inference results do not change, then we shall consider our inference procedures as "objective".

Inference Procedures

Having obtained the posterior distribution of interest we now 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. Furthermore, we can obtain summaries of our posteriors which can give us all the information that can be obtained using a classical approach to inference plus, in certain cases, additional information. We will mention the most commonly used in practice, point estimation and interval estimation.

Point estimation Point estimation is readily available through $\pi(\boldsymbol{\theta} \mid \mathbf{x})$. The most frequently used location measures are the mean, the median and the mode of the posterior distribution since they all have appealing properties. In the case of a flat prior the mode is equal to the maximum likelihood estimate. For symmetric posterior densities the mean and the median are identical. Moreover, for unimodal symmetric posteriors all the three measures coincide. For asymmetric posteriors the choice is not always straightforward. The median is often preferred since, in the case of one-tailed densities, the mode can be very

close to non-representative values while the mean can be heavily influenced in the presence of outliers. In practice, after visualising the posterior density, or a number of scatterplots in the case of multivariate densities, the evaluation of at least the mean and the median is recommended.

Interval estimation A $100 \times (1-\alpha)\%$ *credibility set* for $\boldsymbol{\theta}$ is a subset C of $\boldsymbol{\Theta}$ such that

$$1 - \alpha \leq P(C \mid \mathbf{x}) = \int_C \pi(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta},$$

where integration is replaced by summation for discrete components of $\boldsymbol{\theta}$. In the case of continuous posteriors the $\leq$ is typically replaced by $=$.

This definition enables appealing statements like "The probability that $\boldsymbol{\theta}$ lies in C given the observed data $\mathbf{x}$ is at least $(1-\alpha)$". This comes in contrast with the usual interpretation of the confidence intervals based on the frequency of a repeated experiment. Probably the most attractive credibility set is the *highest posterior density*, or HPD, set defined as

$$C = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \pi(\boldsymbol{\theta} \mid \mathbf{x}) \geq \xi(\alpha)\},$$

where $\xi(\alpha)$ is the largest constant satisfying $P(C \mid \mathbf{x}) \geq 1 - \alpha$. A credibility set of this kind is appealing because it consists of the most likely $\boldsymbol{\theta}$ values. In a sampling based approach the calculation of the HPD set requires a numerical routine. Hence, it is easier to calculate the *equal tail* credibility set by simply taking the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ which equals the HPD set for symmetric unimodal densities.

There are a number of different approaches to Bayesian model assessment and model choice. We will not consider these issues here since they extend beyond the scope of this thesis. For a discussion of several key ideas in the field, including Bayes factors and model averaging, see Berger (1985). We will now

consider a collection of algorithms that greatly facilitate the implementation of Bayesian modelling known as Markov chain Monte Carlo (MCMC) algorithms.

1.3.2 Bayesian Computation

The main idea behind MCMC is to generate a Markov chain which has as its unique limiting distribution the posterior distribution of interest. It dates back to the seminal paper of Metropolis *et al.* (1953) although the computational power required was not available at the time. The original generation mechanism was generalised by Hastings (1970) in the *Metropolis-Hastings algorithm* that we shall describe in the following section.

The Metropolis-Hastings algorithm

The objective of the Metropolis-Hastings (M-H) algorithm is to generate approximate samples from a density $\pi(\theta)$ known up to a normalising constant. Given a conditional density $q(\theta' \mid \theta)$ the algorithm generates a Markov chain (θ_n) through the following steps:

1. Start with an arbitrary initial value θ_0
2. Update from θ_n to θ_{n+1} $(n = 0, 1, \dots)$ by
   (a) Generate $\xi \sim q(\xi \mid \theta_n)$
   (b) Evaluate $\alpha = min\left\{\frac{\pi(\xi)q(\theta_n|\xi)}{\pi(\theta_n)q(\xi|\theta_n)}, 1\right\}$
   (c) Set

$$\theta_{n+1} = \begin{cases} \xi & \text{with probability } \alpha, \\ \theta_n & \text{otherwise.} \end{cases}$$

The distribution $\pi(\theta)$ is often called the *target* distribution whereas the distribution with density $q(\cdot \mid \theta)$ is the *proposal* distribution. The algorithm

described above will have the correct stationary distribution as long as the chain produced is irreducible and aperiodic. This holds true for an enormous class of proposals and usually it suffices, but is not necessary, that the support of the proposal distribution $q(\cdot \mid \theta)$ contains the support of π for every θ. However, the generality of the theorem suggests that the selection of the proposal can be rather decisive. In practice, a proposal with poor overlap between the high density region of π and $q(\cdot \mid \theta)$ may considerably slow convergence. We will now describe the most popular proposal distributions.

The Independent Case A proposal distribution is called independent if it does not depend on θ. This family of distributions admits the form

$$q(\theta' \mid \theta) = f(\theta').$$

This class of proposals can in theory result in algorithms with satisfactory properties as described in Mengersen and Tweedie (1996). In practice the choice of the actual proposal can affect the mixing of the Markov chain drastically. A proposal that is badly calibrated, i.e., a distribution with little support over the high density region of the target distribution, can have extremely slow mixing. Ideally the proposal should resemble the target density being somewhat more diffuse. Usually MCMC algorithms are not based on an independence sampler alone but make use of a number of proposals. However, it is worth emphasizing that a well calibrated independence sampler can outperform most M-H algorithms. We will now describe the most common choice for $q(\cdot \mid \theta)$, the symmetric *random walk* proposal.

Random Walk Metropolis The natural idea behind the random walk proposal is to perturb the current value of the chain at random and then check whether the proposed value is likely for the distribution of interest. In this

case the proposal has the form $q(\theta' \mid \theta) = f(\|\theta' - \theta\|)$ where $\|\cdot\|$ denotes the absolute value. Thus, the proposed value in the M-H algorithm is of the form

$$\xi = \theta_n + \epsilon,$$

where ϵ is distributed according to a symmetric random variable. For this random walk proposal the acceptance ratio becomes

$$\alpha = min\left\{\frac{\pi(\xi)}{\pi(\theta_n)}, 1\right\}.$$

Hence, the chain will remain longer in points with high posterior value while points with low posterior probability will be visited less often. The most popular choices for the proposal $q(\cdot \mid \theta)$ are the normal, the uniform and the Cauchy distributions. In fact, the Gaussian random walk has been, along with the Gibbs sampler described in the next paragraph, among the most commonly used MCMC schemes to date. The algorithm is widely applicable and the only requirement is the scaling of the variance of the proposal. For the Gaussian case Roberts *et al.* (1997) proved that the optimal scaling of the proposal should result to an acceptance rate of approximately 0.234, at least for high-dimensional situations. We will now turn our attention to the *Gibbs sampler*, the most popular MCMC method, particularly in the years following the paper of Gelfand and Smith (1990).

The Gibbs Sampler The Gibbs sampling approach is a special case of the M-H algorithm directly connected to the target distribution π. The method derives its name from Gibbs random fields, where it was used for the first time by Geman and Geman (1984). The idea is to sample from the joint posterior distribution $\pi(\theta^1, \theta^2, \ldots, \theta^\ell)$ using the one-dimensional full conditional distributions $\pi_1, \pi_2, \ldots, \pi_\ell$. Thus, given the current state of the chain $\theta_n^1, \theta_n^2, \ldots, \theta_n^\ell$ we simulate the next state of the chain by sampling

$$\theta_{n+1}^i \sim \pi_i(\theta^i \mid \theta_{n+1}^1, \theta_{n+1}^2, \ldots, \theta_{n+1}^{i-1}, \theta_n^{i+1}, \ldots, \theta_n^\ell), \quad i = 1, \ldots, \ell.$$

The Gibbs sampler has acceptance probability one. Hence, each sample is a successive realisation from the chain. The θ^i's and the full conditional distributions need not be one-dimensional. In fact, for correlated parameters, *blocking* can improve the convergence of the chain considerably. Moreover, when simulation from a given conditional distribution $\pi_i(\theta^i \mid \theta^j, j \neq i)$ is complicated, possibly due to the absence of a closed-form distributional formula, this simulation can be replaced with a Metropolis-Hastings step having $\pi_i(\theta^i \mid \theta^j, j \neq i)$ as the target distribution. Also sampling from the full conditional distributions is not necessarily done in a systematic way. The *random scan* Gibbs sampler, choosing which full conditional distribution to update at random, can have superior convergence properties in certain cases. These are only the most basic variants of the Metropolis-Hastings algorithm. A vast number of modifications and combinations, leading to *hybrid* samplers, appear in the literature. However, these methods go beyond the scope of this chapter and we shall not pursue these issues further.

Implementation

MCMC methods have generated unlimited applicability of the Bayesian paradigm in nearly every branch of statistics. However, the user should always be cautious since the method is based on asymptotic arguments. Hence, there are two practical issues that need investigation to establish the reliability of the chain outcome. The first of these is the *burn in*, i.e. the number of iterations that need to be discarded from the output.

Burn In The Markov chains produced with the proposal distributions that we described thus far are *ergodic*. This means that the distribution of (θ_n) converges, as n goes to infinity, to $\pi(\cdot \mid x)$ for every starting value (θ_0). However,

the speed of this event i.e., the *rate of convergence* varies depending, among others, on the posterior state-space and the sampler used, see Roberts (1996) for a discussion of these issues. Thus, for k large enough, the resulting (θ_k) is an approximate sample from $\pi(\theta \mid x)$. The problem in practice is to determine what a "large" k means. There is a number of diagnostic tests proposed in the literature that provide us with different indicators on the stationarity of the chain. However, none of these tests can actually *guarantee* convergence. Hence, throughout this thesis we investigate the "trace", a plot of the history, of the chain for very long (typically a few million iterations) runs and all the results reported in this thesis are based on chains that appear to have converged. The second practical concern is that after the burn-in some *thinning* of the chain may be required.

Thinning The sample we obtain, after the initial observations are discarded, does **not** necessarily consist of independent observations. In theory this is not crucial if we are interested in functionals of $\pi(\theta \mid x)$ since the *Ergodic Theorem* implies that the average $\frac{1}{L}\sum_{\ell=1}^{L} f(\theta_\ell)$ converges, as L goes to infinity, to $E_\pi(f(\theta))$. In practice however, some sort of batching may be required. Hence, keeping one sample of the chain out of t iterations, with $t = 20$ or $t = 50$ say, we can achieve approximate independent sampling from $\pi(\theta \mid x)$. Moreover, from the practical point of view, we avoid the creation of unmanageable sample sizes that could potentially hamper the statistical analysis of the output.

1.4 Statistical Inference from Outbreak Data

1.4.1 The Nature of Infectious Disease Data

The Reasons for Modelling

The statistical analysis of infectious disease data is typically not a straightforward problem and as such it requires the development of problem specific methodology. Infectious disease data are usually complicated to analyse and there are a number of reasons that makes their analysis awkward. We shall describe the features of infectious disease data in the following section.

The analysis of outbreak data can be more effective when based on a model for the actual mechanism that generates the data. Moreover, epidemic models provide us with a better understanding of the infection process and also with the epidemiologically important quantities of interest. Finally, there are a number of reasons for the analysis, using epidemic models, of historical incidence data. Analyses of this kind can be very useful for diseases occurring due to both novel and re-emerging pathogens as described in the recent review of Ferguson *et al.* (2003). This is of particular relevance at the moment, not only because of the emergence of the SARS outbreak, e.g., Riley *et al.* (2003) and Lipsitch *et al.* (2003), but also due to the threat of deliberately released pathogens such as smallpox, e.g. Kaplan *et al.* (2002) and Halloran *et al.* (2002). Ferguson *et al.* (2003) argue that there does not exist an epidemic model that can be "truly predictive" in the context of smallpox outbreak planning and consequently that no control method can be *a priori* identified as absolutely optimal. However, they suggest that it is vital that a range of models and a set of control options can be identified. Hence, in the event of an outbreak, the models can be adjusted in order to identify the current optimal control method. We shall now describe in detail the difficulties arising during the statistical analysis of

infectious disease data.

The Features of Infectious Disease Data

One of the complications when analysing infectious disease data is that there are often various levels of inherent dependence that one needs to take into account, particularly in the event of a "major" epidemic. Specifically, despite the fact that stochastic epidemics are typically easy to define, there is often a very large number of ways that can result to the same outcome. The complexity of the models increases enormously as they become more realistic. Hence, assuming biologically plausible distributions for the infectious periods, such as Gamma and Weibull instead of the mathematically convenient constant and exponentially distributed ones, induces an additional level of dependence. These facts come in contrast with the usual independence assumption that underlies many of the standard statistical methods. Moreover, the actual disease incidence data are incomplete in different ways. In particular, a relatively informative dataset consists of the times at which the infectious individuals are detected. Even this level of information however is far from being complete. From the inference viewpoint it would be desirable to observe the times that the individuals did contract the disease, as well as the time that the individuals ended their (potential) latent period and could infect others. Additionally, a significant number of data sets only consist of the numbers of individuals who contracted the disease in question. These data can be important data, verified by clinical measurements, or routinely collected surveillance data. However, when realistically complex models are to be fitted to data of this kind, the likelihood can be analytically and numerically intractable. We shall explore in this thesis a number of imputation methods, i.e., different ways of adding information about the epidemic process, that can aid towards overcoming these

difficulties.

It is this nature of epidemic data that makes the statistical analysis of infectious disease data particularly challenging. In the remainder of this section we shall review the work conducted thus far on statistical inference, Bayesian and classical, from outbreak data and we will complete the section with inference about the epidemic models related to the stochastic epidemic model with two levels of mixing that will be the subject of statistical inference in chapters 3 and 4.

1.4.2 Previous Work on Epidemic Modelling

Monographs on Epidemic Models

There is a vast literature on deterministic and stochastic epidemic modelling. We shall mention here the main books on epidemic modelling. Most of the work on modelling disease transmission prior to 1975 is contained in Bailey (1975). The author presents a comprehensive account of both stochastic and deterministic models, illustrates the use of a variety of the models using real outbreak data and provides us with a complete bibliography of the area.

Becker (1989) presents statistical analysis of infectious disease data. The author uses a number of different models and analyses a large variety of real life outbreak data. The single book that has received most attention recently is Anderson and May (1991). However, the authors only focus on deterministic models, as does the recent monograph by Diekmann and Heesterbeek (2000). A six-month epidemics workshop took place in 1993 in the Isaac Newton Institute in Cambridge. A large part of the outcome of the work conducted in this meeting is summarised in the three volumes edited by Grenfell and Dobson (1996), Isham and Medley (1996) and Mollison (1996) respectively. A recent addition

to the literature of stochastic epidemic modelling is Andersson and Britton (2000) which provides an excellent introduction to stochastic modelling and the authors also mention some basic statistical analysis for stochastic epidemic models. Since the seminal paper of Mollison (1977) on spatial epidemics there has been increasing interest in the applied probability literature for models of this kind. Also, a number of spatial epidemic models based on bond-percolation has been developed since the paper of Kulasmaa (1982), see for example the book by Liggett (1999) and the references therein.

Reviews of Epidemic Models and their Analysis

There does not exist a monograph concerned with the recent progress on the statistical analysis of infectious disease data. Becker and Britton (1999) present a critical review of statistical methodology for the analysis of outbreak data. The authors make an attempt to place emphasis on the important objectives that analyses of this kind should address, as well as suggesting issues where further work is required. Recently, Ferguson *et al.* (2003) conducted a review of epidemic models with reference to planning for smallpox outbreaks. The authors emphasize the importance of epidemic modelling as a useful tool for assessing the threat posed by deliberate release of a known pathogen, as well as dealing with the emergence of a novel virus. We shall now present the statistical analysis of epidemics that are relevant to the models that this thesis will attempt to explore.

Statistical Analysis of Epidemic Models

This section will review methods of parametric inference about the infection rate(s) and the epidemiologically important parameters. See Becker (1989) for a comprehensive account of nonparametric inference methodology based on

martingale methods.

Epidemics in homogeneous populations The first statistical analysis of removal data, based on a continuous-time model, for the purpose of estimating the infection and the removal rate is described in Bailey and Thomas (1971). The authors analysed the general stochastic epidemic using maximum likelihood methods. Rida (1991) derives asymptotic normality results for some estimators of the infection rate and the corresponding basic reproduction number. However, the largest amount of information for inference based on epidemic models defined on a homogeneous population is in Becker (1989). A large number of different approaches are presented including the author's work for parametric as well as non-parametric methods of statistical inference.

As with many application areas of statistics, inference for stochastic epidemic models has benefited considerably from the use of Markov chain Monte Carlo methods. In particular, Gibson and Renshaw (1998) and O'Neill and Roberts (1999) first presented a statistical analysis of S-I-R models based on MCMC methods. O'Neill and Becker (2001) have presented inference procedures for a non-Markovian epidemic model where the infectious period follows a Gamma distribution. Streftaris and Gibson (2003) use MCMC methods in a different extension of the general stochastic epidemic model where the infectious period is distributed according to a Weibull random variable, with particular reference to plant epidemiology. Finally, Hawakaya *et al.* (2003) extend the basic model in two key directions. They allow for a multitype (e.g. different age, sex) model, where the infection rates vary between different types, as well as the actual number of susceptibles being unobserved. The authors derive statistical inference for both the infection rates and the size of the population.

We shall briefly mention three papers that focus more on the statistical

context of inference for epidemics, as opposed to inference for a wider class of epidemic models than those analysed before. A group of parameterisations that can improve the convergence of MCMC algorithms used in the epidemics context is the subject of Neal *et al.* (2003). Specifically, the authors describe algorithms that can be more robust with respect to the mixing of the Markov chain. A method that eliminated the need for assessing convergence of the Markov chain is the perfect simulation algorithm originated by Propp and Wilson (1996). O'Neill (2003) proposes methods of perfect simulation when the infection process is of the Reed-Frost type. Finally, a different statistical method, based on the forward-backward algorithm, that can be used for estimation of the infection and removal rate in the general stochastic epidemic is presented in Fearnhead and Meligkotsidou (2003). Statistical inference is less obvious when the population in question admits a particular structure, e.g. households, and these methods will be described in the next paragraph.

Epidemics in structured populations

Epidemics on independent households Longini and Koopman (1982) consider models in which individuals reside in households and may be potentially infected both from infectives within their household or from individuals outside their household. Their model assumes that the disease within the household progresses independently of the dynamics of the community. This approach is generalised in a model with a general infectious period in Addy *et al.* (1991). The authors extend the work of Ball (1986) on the generalised stochastic epidemic so that individuals can also be infected from the community at large. We will comment further on the approach of Addy *et al.* (1991) and the limitations of their model in chapters 3 and 4. Britton and Becker (2000) use the Longini-Koopman model in order to estimate the critical vaccination

coverage required to prevent epidemics in a population that is partitioned into households. O'Neill *et al.* (2000) use MCMC method to analyse both temporal and final size data from household outbreaks. Finally, a different perfect simulation method is applied in Clancy and O'Neill (2002) where the authors analyse a model related to the Longini-Koopman model where some variation in the probability of individuals from different households being infected from outside is included.

Probably the most important application of epidemic models is in epidemics control, typically using vaccination. A large body of literature exists and we shall only mention a few key references. Becker and Dietz (1995) study the control of diseases among households assuming that once there is an infective in a household everybody contracts the disease. Ball and Lyne (2002) derive the effect of different vaccination policies in a population that is partitioned into households while Becker *et al.* (2003) use an independent households model to estimate vaccine efficacy from household outbreak data, see Halloran *et al.* (1999) for a review of other methods for estimating vaccine efficacy. We shall now mention briefly statistical inference for epidemics with two levels of mixing.

Inference for Epidemics with two levels of mixing Ball *et al.* (1997) briefly consider statistical inference for their model. They mention that, for estimation purposes, their model can asymptotically be approximated by the model of Addy *et al.* (1991). The authors use the basic idea and the results from Addy *et al.* (1991) to examine different vaccination strategies among households. Britton and Becker (2000) also formulate their work in order to estimate the immunity coverage required for preventing an outbreak when the population is partitioned into households, in terms of the two-level-mixing model. However, as mentioned in the previous paragraph, they perform

their statistical inference with respect to an independent households model. A more detailed approach that utilises (pseudo)likelihood inference is presented in Ball and Lyne (2003). The authors describe inference procedures for the multitype version of the model described in Ball and Lyne (2001). The method used is related to the method we describe in chapter 3 and we will comment on specific results in the appropriate sections of chapter 3.

Epidemics with different population structure In real life populations individuals interact with a number of different environments additionally to their household, such as schools and workplaces. However, it would be impossible to capture every aspect of the population structure. In the recent years there has been intense activity in describing the population structure through a *random network* structure. Probably the simplest model of this kind is a Bernoulli random graph where each individuals has social contacts with other individuals in the populations according to a fixed probability. 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 and the disease spreads as a general stochastic epidemic. Neal *et al.* (2003) extend their reparameterisations in this case to offer robust MCMC algorithms for the infection and removal rate as well as the probability of social contact.

These methods can, in principle, be extended to more complicated social structures. It would be particularly interesting to consider statistical inference when the population is assumed to have social contacts according to a complex random graph. There has been intense recent interest in sociology and statistical mechanics since the pioneering work of Watts and Strogatz (1998) on "small-world" networks, see for example the review by Strogatz (2001). Albert and Barabàsi (2002) present an extensive review of the different models for the structure of the community and other networks such as the internet and vari-

ous biological networks, including the so-called scale free networks. A number of algorithms have been developed for the detection of these structures and it would be interesting to combine our approach with an approach of the kind described in Newman (2003). Finally, there has been some interest in statistical inference for spatio-temporal epidemic models, see for example Gibson (1997) and Marion *et al.* (2003).

Chapter 2

Exact Results for the Generalised Stochastic Epidemic

2.1 Introduction

This chapter contains methods for the numerical solution of the set of the triangular equations describing the final size probabilities of the generalised stochastic epidemic. This simple stochastic process was described in the previous chapter, although no attempt was made to describe statistical analysis of the model. The probabilities of the final size can be derived from a well known set of recursive equations, see Ball (1986). However, practical implementation of attempts to solve these equations is frequently hampered by numerical problems, see for example Andersson and Britton (2000). These problems arise even for moderate population sizes, and the accuracy of the solution also depends on the parameter values. We shall explain the reason for these numerical problems in the next section but essentially the problem stems from the nature of the distribution of the final size probabilities and the recursive method by which they are obtained. The recursive nature of the triangular equations means that

in order to evaluate the "interesting" probabilities that are typically the object of inference, we first need to calculate probabilities that are very close to zero. The problem gets worse as the population size increases since the number of the probabilities that are numerically negligible becomes large.

Potential solutions to this problem can, in principle, arise from the numerical analysis literature since the final size distribution is essentially obtained by solving a linear system of equations. Previously derived numerical methods offer a number of approximate solutions for similar problems. However, our problem is of a different nature. In particular, the space of the solution in our case is constrained to the positive real vectors that sum to unity. Hence, the assumptions that typically underly the approximate methods break down.

In this chapter we shall utilise a different approach to the solution of this linear system of equations. Our method involves using multiple precision arithmetic. In particular, each number is stored in the computer as a long vector. In fact, the amount of memory that is allocated to each multiple precision number can be set by the user. Hence, by increasing the length of the vector we can achieve arbitrarily high accuracy. In practice, the price we pay is the rapid increase in the computation time when we compare it to the usual double precision that most computers use in order to perform real number calculations. To offer a quantitative idea of the extended precision arithmetic it is worth recalling that modern computers allocate 8 bytes of memory for a double precision number and 16 bytes of memory for quadruple precision real numbers. In the algorithms we have used in this chapter we have allocated up to 580 bytes of memory for each multiple precision variable used for the computation.

An immediate advantage due to the use of multiple precision arithmetic is the exact evaluation of the final size probabilities for any different population size, initial number of infectives, final size of the epidemic and distribution of

the infectious period. This solution can be used to assess the different limit theorems that appear in the epidemics literature. We have already seen in the previous chapter a branching process approximation for the initial stages of an epidemic as well as a normal approximation for the distribution of the final size in the event of a major epidemic. Additionally, we can perform both Bayesian and likelihood inference since solving the triangular equations immediately provides us with the likelihood function. The remainder of the chapter is organised as follows. Relevant results for the generalised stochastic epidemic are summarised in the next section while section 3 contains the solution, using multiple precision arithmetic, of the triangular equations that determine the final size distribution. Using this distribution we assess two widely used limit theorems for the generalised stochastic epidemic in section 4. In section 5 we describe the MCMC algorithm used for statistical inference while section 6 contains the corresponding results and the chapter ends with some concluding remarks.

2.2 The Generalised Stochastic Epidemic

2.2.1 The Basic Model

In this section we shall briefly recall the notation and the key features of the epidemic process. The epidemic propagates through a population of $n+m$ individuals out of which m are initially infected and the n remaining individuals are susceptible to the disease in question. The infectious periods of different individuals are assumed to be independent and identically distributed according to a random variable I, having an arbitrary but specified distribution. While infectious, an individual makes infectious person-to-person contacts at the points of a time homogeneous Poisson process with intensity $\frac{\lambda}{n}$. The Poisson processes

of different individuals are assumed to be mutually independent.

2.2.2 Final Size Probabilities

Let $\phi(\theta) = E[\exp(-\theta I)]$ be the moment generating function of the infectious period I and p_k^n be the probability that the final size of the epidemic is equal to k, $0 \leq k \leq n$. Then

$$\sum_{k=0}^{l} \frac{\binom{n-k}{l-k} p_k^n}{\left[\phi\left(\frac{\lambda(n-l)}{n}\right)\right]^{k+m}} = \binom{n}{l}, \qquad 0 \leq l \leq n. \tag{2.1}$$

The system of equations in (2.1) is triangular. Linear systems of this kind are typically considered to be straightforward to solve, e.g. Higham(1989). Hence it appears easy to calculate the final size probabilities recursively.

For illustration, in figure 2.1 we present the final size probabilities for an epidemic in a population of 800 individuals. The epidemic was initiated with 1 initial infective, the infectious period I follows a Gamma distribution with mean $E(I) = 4.1$ and variance $Var(I) = 8.405$ being the sum of two exponential random variables with mean 2.05. The infection rate parameter is set to $\lambda = 1$. Hence, in this particular case $R_0 = E(I) = 4.1$. Then we solved equations (2.1) using multiple precision arithmetic, as described later, and the outcome is presented in figure 2.1.

The bimodal shape of the final size distribution is obvious. In particular, the probability mass is split between the initial part, corresponding to the case that the epidemic goes extinct, and the "normal" part around the number of individuals that ultimately get infected once an epidemic has taken off. Hence, even for $R_0 = 4.1$ the probability that the the epidemic will die out quickly is not negligible. We can approximate this probability by the probability that the corresponding branching process dies out and we shall assess this approximation in the fourth section of this chapter. In the case of non-extinction, the

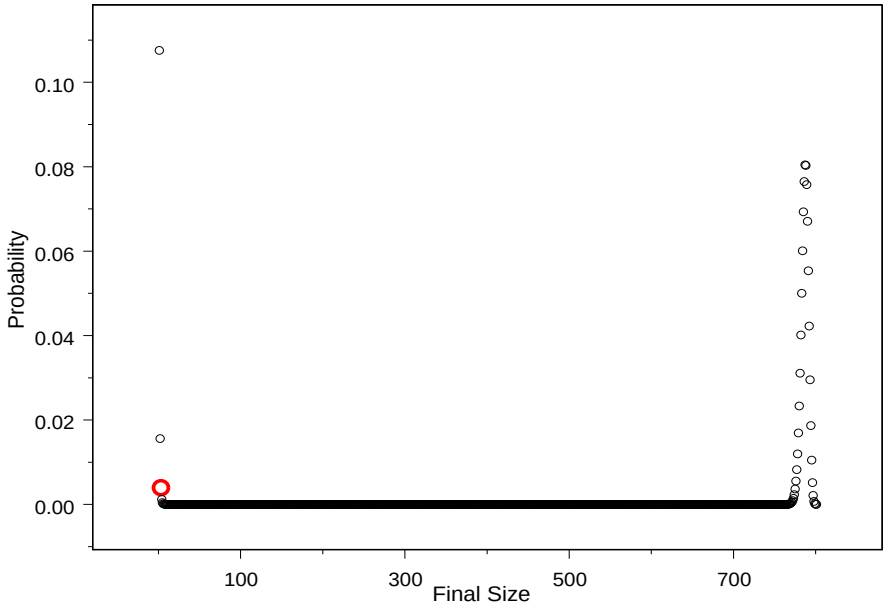

Figure 2.1: The final size distribution of an epidemic among a population of 800 individuals.

largest part of the probability mass is concentrated around the normal part with the most likely final size being approximately 785. We would expect a severe epidemic of this kind since the basic reproduction number is relatively large. Note that the "normality" of most likely sizes in the event of a major epidemic would be even more clear for an epidemic among a smaller population.

Numerical Problems

When using standard (double) precision arithmetic, numerical problems appear due to rounding errors even for moderate population sizes of order 50-100. The exact final size where negative probabilities arise depends on the actual final

size as well as λ and ϕ. In particular, when the population size is, say, 100 the solution, for a given λ and ϕ, appears to be stable when the final size varies between 40 and 60. Therefore, the equations are more stable in the case where the epidemic has taken off. However, when the population size is greater than 100, negative probabilities will appear for any final size calculations.

The Cause of the Instability

The actual reason for numerical problems is the final size probabilities in the range between final sizes that correspond to epidemic extinction and final sizes that comprise of a large proportion of the population. These probabilities represent final sizes that appear to be very unlikely for this particular value of R_0. It is unfortunate that we need to calculate all the intermediate probabilities in order to evaluate the probabilities of the typically most useful phase of the epidemic, close to the "normally distributed" part of the final size distribution. We shall now briefly explain how multiple precision arithmetic operates and how it can be used for solving the triangular equations of the kind presented in (2.1).

2.3 Exact Final Size Probabilities

2.3.1 Multiple Precision Arithmetic

There has been increasing demand in scientific computation for augmented precision arithmetic. The applications vary from simply evaluating mathematical constants (Shanks and Wrench (1962)) with very high accuracy, to solving problems with great financial implications, such as using finite element methods for the design of aeroplanes and cars. The most widely used package of

this kind has been Brent's MP package (Brent (1978)), implemented in Fortran with great functionality and efficiency.

We used Smith's FMLIB package (Smith (1991)) which is a more modern set of multiple precision routines, implemented in Fortran. FMLIB gives comparable speed to Brent's MP routines at low precision and greater speed when higher precision is required. This increase in speed comes mainly from the use of improved algorithms for computing the elementary functions in multiple precision, see Smith (1989). The solution of the triangular equations in (2.1) requires only elementary calculations. Hence, we would expect improved performance using the FMLIB package. We shall not give here details of the design of the multiple precision arithmetic or how the required accuracy and efficiency are obtained as these details go beyond the scope of this thesis. The interested reader is referred to Smith (1991) for the technical details of the implementation of the FMLIB package. We shall now describe briefly the evaluation, using multiple precision arithmetic, of the final size probabilities.

2.3.2 Multiple Precision Final Size Probabilities

For the purposes of solving a triangular system of the kind described in (2.1) it is convenient to rewrite the equations as

$$\sum_{k=0}^{l} \frac{\binom{n-k}{l-k} p_k^n}{\binom{n}{l}\left[\phi\left(\frac{\lambda(n-l)}{n}\right)\right]^{k+m}} = 1, \qquad 0 \le l \le n. \tag{2.2}$$

Then (2.2) can be written as $\mathbf{AP} = \mathbf{1}$ where $\mathbf{A}$ is the $(n+1) \times (n+1)$ lower triangular matrix with elements $a_{kl} = \frac{\binom{n-k}{l-k}}{\binom{n}{l}\left[\phi\left(\frac{\lambda(n-l)}{n}\right)\right]^{k+m}}, l = 0, \ldots, n \quad k = 0, \ldots, l$, $\mathbf{P} = \{p_0, p_1, \ldots, p_n\}^T$ and $\mathbf{1}$ is the vector with all its elements being equal to 1. Then it is straightforward to obtain $\mathbf{P}$ by

$$p_k = \frac{1 - s_k}{a_{kk}}, \quad s_k = \sum_{i=0}^{k-1} a_{ki} p_i, \quad k = 0, \ldots, n. \tag{2.3}$$

Therefore, we only need to define **A**, **P** and 1 as multiple precision and after solving (2.3) we can convert **P** to double precision and proceed as usual. The price we pay is that this solution is far more computer intensive and time consuming compared to one that could be obtained using double precision arithmetic. It does work with very high accuracy for virtually any population size but it can be rather infeasible for real life applications with a very large population size when it is part of another computer intensive algorithm such as Markov chain Monte Carlo. However, this does not have to be the case. In fact, it is straightforward and relatively quick to reconstruct, for a given final size, the likelihood function over a grid of λ values and this can be the basis of a worthwhile alternative to Bayesian inference procedures.

2.4 Evaluation of limit theorems

In this section we shall use the final size distribution, as obtained by solving the triangular equations in (2.3), to assess some aspects of the two most widely used limit theorems in epidemic theory namely, evaluating the probability that the epidemic goes extinct, based on coupling the initial stages of the epidemic with a suitable branching process and, in the case of non-extinction, approximating the distribution of the final size with a Gaussian distribution.

2.4.1 Probability of Epidemic Extinction

The evaluation of the extinction probability for a stochastic epidemic model can be achieved using the branching process approximation for the initial stages of the epidemic process as described in page 9. We recall from Andersson and Britton (2000) theorem 3.1, that, when $R_0 > 1$, an epidemic becomes extinct with probability q^m where m is the number of initial infectives and q is the

smallest root of the equation

$$\phi(\lambda(1-\theta)) = \theta. \tag{2.4}$$

In (2.4) ϕ is the moment generating function of the infectious period I. We evaluated the final size distribution for an epidemic among 1000 individuals, starting with one initial infective, where the infectious period follows an exponential infectious period with rate 1 and the contact rate was set to $\lambda = 1.5$. The solution can be seen in figure 2.2 and will be referred to as the "exact" one. Since the probabilities of final sizes that correspond to epidemic extinction account for a relatively large proportion of unity, the probabilities of the "likely" final sizes, conditionaly upon non-extionction are relatively small. Note that for this model we have $R_0 = 1.5$.

It is straightforward to solve (2.4) for both epidemics and compare with the exact number as obtained by solving the triangular equations since the solution given by (2.4) becomes exact when the epidemic propagates in an infinite population. In the first epidemic when $N = 800$ and $R_0 = 4.1$ solving (2.4) yields $q = 0.1289$. The corresponding number based on the final size distribution was $\sum_{j=1}^{i} p_j^N = 0.12904$. In this case we used the probabilities of final sizes up to $i = 50$. However, the results do not change much (up to the fourth decimal point) for any i in the interval $(30, 700)$. Hence, we see that the coupling of the initial stages of an epidemic with a branching process yields a reasonable approximation for the extinction probability of a supercritical epidemic in a population of 800, at least for the parameter values considered.

For the second epidemic when $N = 1000$ and $R_0 = 1.5$ solving (2.4) yields $q = 0.667$. The corresponding number based on the final size probabilities was $\sum_{j=1}^{200} p_j^N = 0.67205$. Again, varying the upper bound of the sum between 100 and 300 has little influence on the actual extinction probability. Thus, even when there is a large probability that the epidemic goes extinct, the coupling

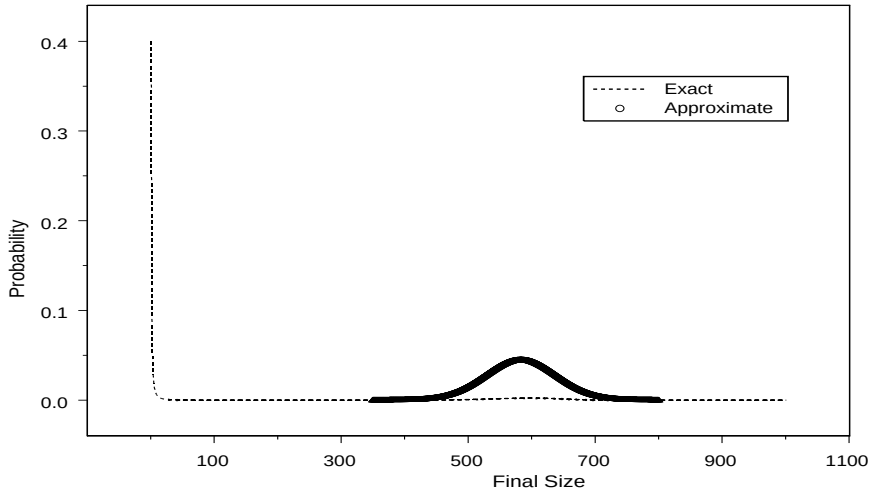

Figure 2.2: The final size distribution of an epidemic among a population of 1000 individuals and the corresponding Gaussian approximation.

of the epidemic process with a suitable branching process provides a reasonable estimate of the extinction probability. In the following section we attempt to assess how well a Gaussian approximation can describe the probabilities of the "likely" final sizes in the event of a major epidemic i.e., when $R_0 > 1$.

2.4.2 Gaussian Approximation

We shall use the final size probabilities, as obtained by solving (2.3) to assess the Normal approximation for the distribution of the final size described in Andersson and Britton (2000) theorem 4.2, which holds conditional on the occurrence of a major epidemic. Let us denote by T_N the final size of the epidemic under consideration and by τ the proportion of the individuals that get ultimately infected i.e., $\tau = \lim_{N\to\infty} \frac{T_N}{N}$. When $R_0 > 1$, τ is the largest

solution (note that 0 is always a solution) of the non-linear equation:

$$\tau = 1 - \exp(-\lambda E(I)\tau).$$

Here we assume that proportion of initial infectives is negligible. Then for $\rho = 1 - \tau$ we have

$$T_N \sim N\left(\tau N, \frac{N\left[\rho(1-\rho) + \lambda^2\sigma^2\tau\rho^2\right]}{\left(1 - \lambda E(I)\rho\right)^2}\right), \tag{2.5}$$

where $N(a, b)$ denotes the density of a normal random variable with mean a and variance b. The theorem holds for the important practical case when there is (in an infinite population) a finite number of initial infectives. We shall describe the validity of the Gaussian approximation graphically since evaluating distance measures (such as the total variation distance) when comparing exact with approximate results is of limited practical use. We have created three figures that summarise the validity of the normal approximation. The first is figure 2.2 where we plot the distribution of the final size when $R_0 = 1.5$ with the corresponding Gaussian density described in (2.5). Note that in this case we have $\tau \approx 0.583$. This is a very interesting case because a large number of papers concerned with statistical inference for epidemics, including the approach that we adopt in chapter 3, implicitly or explicitly use a Gaussian approximation of the kind described in (2.5) simply by conditioning on the occurrence of a major epidemic and without any rescaling to take into account for this, and this figure is meant to clarify exactly this. However, since there is a large extinction probability the exact probabilities are quite small when compared to the normal approximation and thus, difficult to assess from figure 2.2.

A perhaps more fair and complete comparison is presented in figure 2.3 where we set the probabilities of the "small" final sizes to 0 (since the normal approximation is not valid in this case) and we scale the remaining probabilities accordingly. Hence, for this particular case, we have chosen to work with

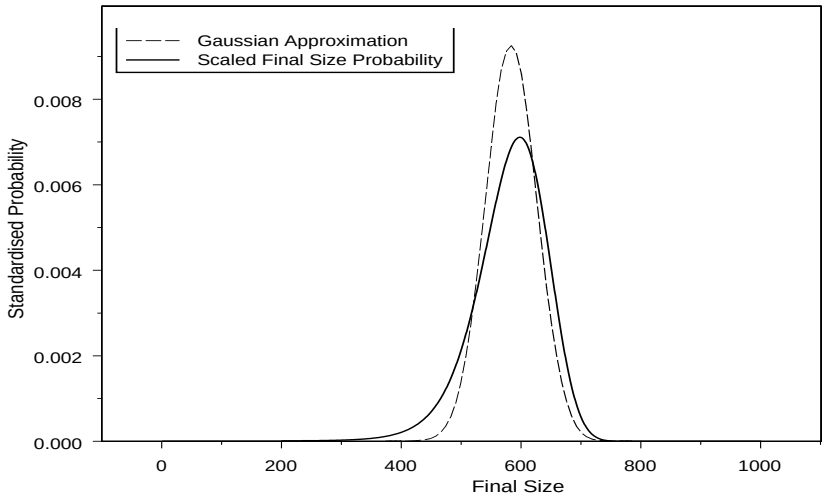

Figure 2.3: The scaled final size distribution of an epidemic among a population of 1000 individuals and the corresponding Gaussian approximation.

final sizes above 200 and we divided the remaining final size probabilities with 0.32672 = 1 - Pr(epidemic extinction). In this comparison the Gaussian approximation performs reasonably well, the main drawback being that a slight left tail (inherently present in many final size distributions) cannot be captured by a normal distribution. This is a feasible approach, since the rescaling factor can be evaluated by the branching process approximation. In the following, we present a comparison that is not feasible in many cases, since we shall rescale with a quantity that is typically unknown, unless one can solve the triangular equations exactly.

Additionally to the previous cases, we have created a third figure where we set the initial probabilities to 0 (again due to the fact that the Gaussian part does not refer to these final sizes) and we standardise the probability distributions so that they have the same likelihood value at the mode. The

outcome can be seen in figure 2.4 where the agreement is very good. These findings remain very similar for population sizes as low as (approximately) 100. Below these population sizes there is a "conflict" between the left tail of the normal approximation and the (not negligible) probabilities that correspond to epidemic extinction. In particular, for $N = 100$ the mean and standard deviation in (2.4) are 58.3 and 18.34 respectively. Hence, the "$\mu - 3\sigma$" 0.005 percentile of the normal distribution is already in the area of final sizes as small as 3 or 4 while the mode is relatively close to the exact probabilities, similarly to figure 2.2. We recall that these findings are for $\lambda = R_0 = 1.5$. For smaller population size the normal approximation breaks down. For $N = 50$ for example we get $\mu = 29.15$ and $\sigma = 12.97$ and the left tail of the Gaussian approximation is obviously insufficient to describe the probabilities of final sizes that correspond to epidemic extinction.

The above assessment is not only interesting from the probabilistic point of view. As we shall see in the next chapter, limit theorems of this kind can be used to perform approximate statistical inference. Thus, the two key observations for the validity of the Gaussian approximation are (i) the Gaussian distribution performs reasonably well with respect to the location of the "likely" final sizes as long as the population is above 100 and (ii) it can be a feasible option with respect to variability measures under appropriate rescaling. If it could be rescaled with the (typically unknown) exact probability of the most likely final size then the approximation, for a large number of realistic examples would be very satisfactory. Thus, care is required when using the original (not rescaled) normal approximation if it is believed that there is a considerable probability for the event $R_0 \leq 1$. In the following section we shall explore the use of the exactly evaluated final size probabilities for conducting statistical inference for the basic reproduction number R_0.

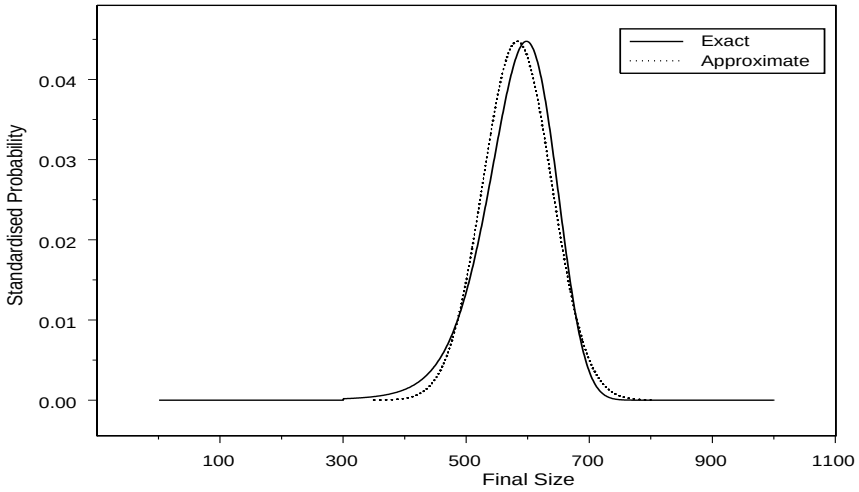

Figure 2.4: The standardised final size probabilities of an epidemic among a population of 1000 individuals and the corresponding Gaussian approximation.

2.5 Statistical Inference for the GSE

2.5.1 Bayesian Inference

Based on the exact solution of the triangular equations, it is straightforward to obtain the posterior distribution of the infection rate λ and the basic reproduction number R_0. We recall from the first chapter Bayes' Theorem which states that for a given final size data point x the posterior distribution is given by

$$\pi(\lambda \mid x) \propto \pi(x \mid \lambda)\pi(\lambda)$$

where $\pi(\lambda)$ is the prior density of λ. We classify this problem as non-standard since we essentially attempt to estimate the rate of a stochastic process from a single data point. The main tool in this effort is the structure we impose in this process, but the problem remains ambitious from the statistical point of

view. We now describe the Markov chain Monte Carlo algorithm we shall use as our inferential tool.

2.5.2 MCMC algorithm

Updating λ: A simple Gaussian Random-Walk type Metropolis algorithm, as described in the first chapter, was found to be sufficient. Hence, we used a normal proposal density $q(\cdot \mid \lambda)$, centred to the current value, where each proposed infection rate parameter $\lambda^* < 0$ is rejected with probability 1 since the assumptions of the model reduce the state-space of the Markov chain to the positive real line. In the case of a positive proposed infection rate, we accept the proposed value with probability

$$\frac{\pi(x \mid \lambda^*)\pi(\lambda^*)}{\pi(x \mid \lambda)\pi(\lambda)} \wedge 1,$$

where $A \wedge B$ is defined as $\min\{A, B\}$, since the ratio of the proposals is cancelled due to symmetry: $q(\lambda^* \mid \lambda) = q(\lambda \mid \lambda^*)$.

Prior specification: The rate of the Poisson contact process λ was assumed *a priori* to follow a Gamma distribution with mean 1 and variance 10000. It is generally recommended in the MCMC literature to use prior densities with large variance and some sensitivity analysis with respect to the prior assumption is presented in the following subsection.

Note that it is straightforward to use the distribution of the final size to perform likelihood inference for the infection rate parameter. A simple, although numerically intensive, method to perform such inference would be the evaluation of the likelihood of the final size over a fine grid of λ values. However, we would expect that the use of the locally flat prior would eliminate the influence of the prior. Further sensitivity analysis with respect to the choice of the prior will be presented at the end of the following section. We will now

present the results from the MCMC algorithm.

2.6 Results

In this section we will explore the effect of different infectious periods by examining the various posterior distributions of the threshold parameter R_0 for three different choices for the infectious period distribution. We have chosen three distributions that we would expect to cover a large spectrum of the behaviour of the R_0 estimates as the distribution of the infectious period varies namely, a constant($\equiv 4.1$) infectious period that has the smallest possible amount of variability, an exponential infectious period with mean 4.1, with a large variability for the infectious period distribution, and a gamma distributed infectious period, being the sum of two exponentials with mean 2.05. With a constant infectious period the generalised stochastic epidemic corresponds to a continuous time version of the Reed-Frost epidemic model (see Bailey 1975) while the exponential infectious period converts the model to its Markovian version, the widely studied general stochastic epidemic. Note that the exponential infectious period is rather extreme in the sense that the standard deviation equals the mean. Hence, the gamma distribution we have chosen appears to be an intermediate choice with respect to the variability of the infectious period and we shall explore the effect of these choices from the output of the MCMC algorithm.

These results can be compared with the estimated R_0 derived by Rida (1991). It should be noted that our method is exact while the results in Rida (1991) are only valid as the population size gets infinite. Moreover, the estimators derived in (3.8) and (5.3), (5.4) of Rida (1991) regarding the MLE of R_0 and the standard error of the estimator are consistent in the case that the

Posterior estimates for R_0 when $x = 30$	Infectious Period		
	Constant	Gamma	Exponential
Mean	1.1773	1.2188	1.2565
Median	1.1649	1.1929	1.2103
S. dev.	0.211	0.273	0.336
Equal-tailed 95% C. I.	(0.799,1.624)	(0.760,1.825)	(0.729,2.047)

Table 2.1: Posterior summary statistics for the three different infectious periods when 30 out of 120 individuals are ultimately infected.

epidemic is above threshold, that is, $R_0 > 1$. This is a common assumption in classical inference for epidemics. In contrast, the methods we describe do not rely on conditioning upon non extinction of the epidemic. Additionally, it would be interesting to compare our results with the martingale estimators described in chapter 7 of Becker (1989).

We should mention that it is not straightforward in the epidemics context to verify the regularity conditions under which the posterior mode (and mean and median for symmetric posteriors) should agree with the maximum likelihood estimator. Typically, these conditions hold true when the relative contribution of the prior is "small" in some sense and since we have no closed form for the likelihood, a theoretical proof appears very difficult. However, for examples where the epidemic has a significant proportion of ultimately infected individuals i.e., $R_0 > 1$, and the population size is large, we would expect good agreement between the two location estimates as well as approximate equality between the posterior standard deviation and the standard error of the MLE, despite the fact that the two quantities describe completely different measures.

For comparison purposes we apply our algorithm to a dataset from a smallpox outbreak in the closed community of Abakaliki in south-eastern Nigeria,

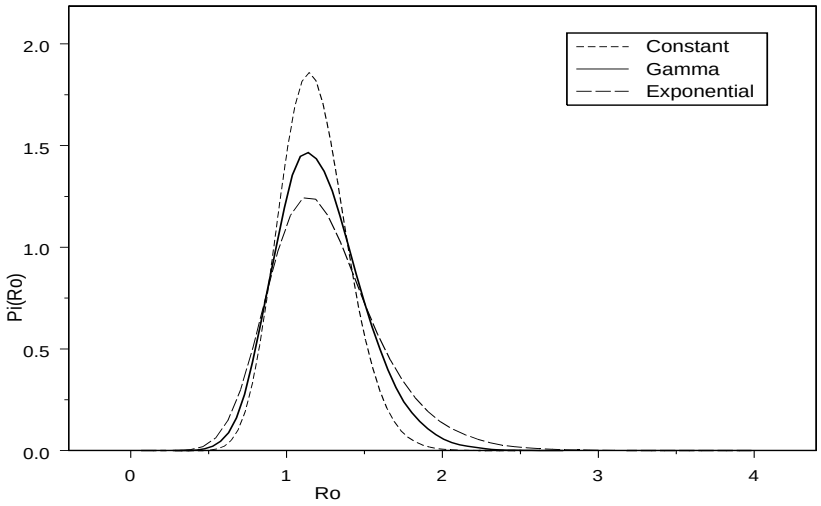

Figure 2.5: Posterior density of R_0 for the three different infectious periods when $x = 30$.

Posterior estimates for R_0 when $x = 60$	Infectious Period		
	Constant	Gamma	Exponential
Mean	1.4236	1.4396	1.4673
Median	1.4128	1.4255	1.4433
S. dev.	0.182	0.228	0.274
Equal-tailed 95% C. I.	(1.092,1.812)	(1.037,1.918)	(0.999,2.077)

Table 2.2: Posterior summary statistics for the three different infectious periods when 60 out of 120 individuals are ultimately infected.

see section 6.4 of Becker (1989). The results are summarised in Table 2.1. The first important observation arises with respect to estimation of the location measures for R_0. In fact it is straightforward to see it when considering figure 2.5. In particular, the mode appears to be very close in all the three cases. This observation is in harmony with the asymptotic results from equation (6) in Becker and Britton (2001) since, in the limit as the population size tends to infinity, the maximum likelihood estimate (that should agree with the posterior mode under weak prior assumptions) does depend on the mean of the infectious period but not the variance.

With respect to comparison with previously derived estimates of R_0, the results in Rida (1991) report an MLE of 1.108 while the posterior mode for the general stochastic epidemic (exponential infectious period) is approximately 1.15. Similar (slight) underestimation occurs when considering martingale methods. In particular, Becker (1989) p.153 estimates the threshold parameter as $R_0 = 1.10$. An interesting remark here though is that the results of both authors are approximately correct despite the fact that there is clearly a significant amount of the posterior density in the area where R_0 is below unity, see figure 2.5.

The second important observation from the output of the MCMC algorithm arises with respect to the estimates of the variability of R_0. In particular, the posterior distribution of R_0 becomes less peaked as the variance of the infectious period reduces. The effect of the variance of the infectious period is more obvious when considering the location measures that are affected by the shape of the posterior distribution, particularly the mean. Hence, the exponentially distributed infectious period displays the most skew distribution. The posterior when using a constant infectious period is almost symmetric while the posterior distribution of R_0 for the Gamma infectious period has

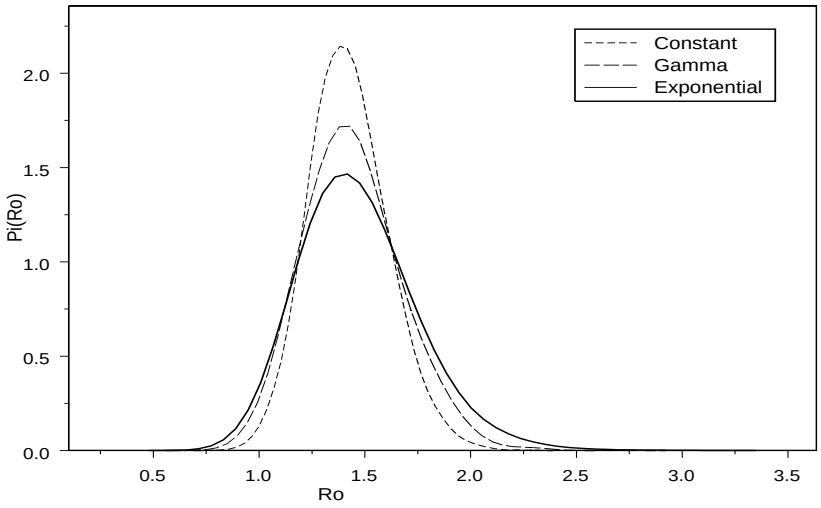

Figure 2.6: Posterior density of R_0 for the three different infectious periods when $x = 60$.

an intermediate behaviour. This is an effect that cannot be reproduced with classical inference procedures.

With respect to the actual estimates of the variance, both standard errors that are reported in Rida (1991) p.278 are slightly smaller than the posterior standard deviation of the general stochastic epidemic. The most interesting observation however, is that both confidence intervals that she evaluates namely, (0.531,1.685) and (0.542,1.674) have a lower bound below unity, despite the $R_0 > 1$ assumption underlying the theory developed in that paper.

We have also applied our algorithm to the case where 60 out of 120 individuals are getting infected when an epidemic is initiated from one initial infective. The results are presented in Table 2.2. The actual estimates of R_0 are (naturally) larger but the findings reported in the case $x = 30$ remain similar. This

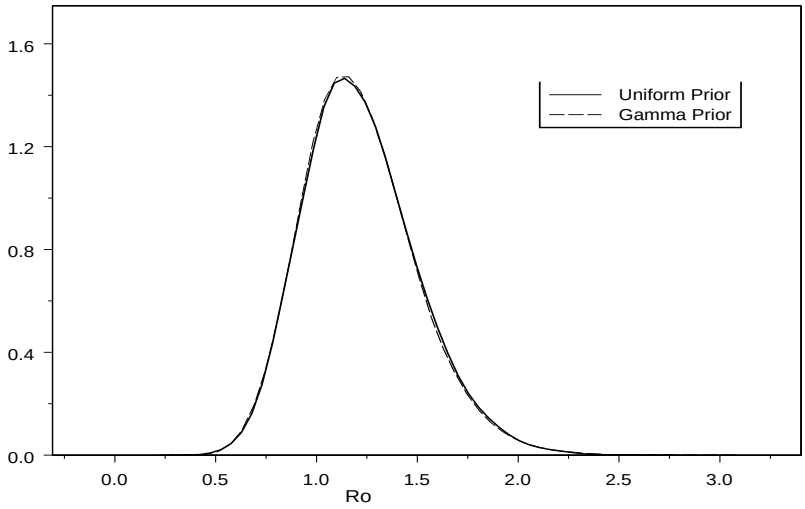

Figure 2.7: Posterior density of R_0 for the two different priors.

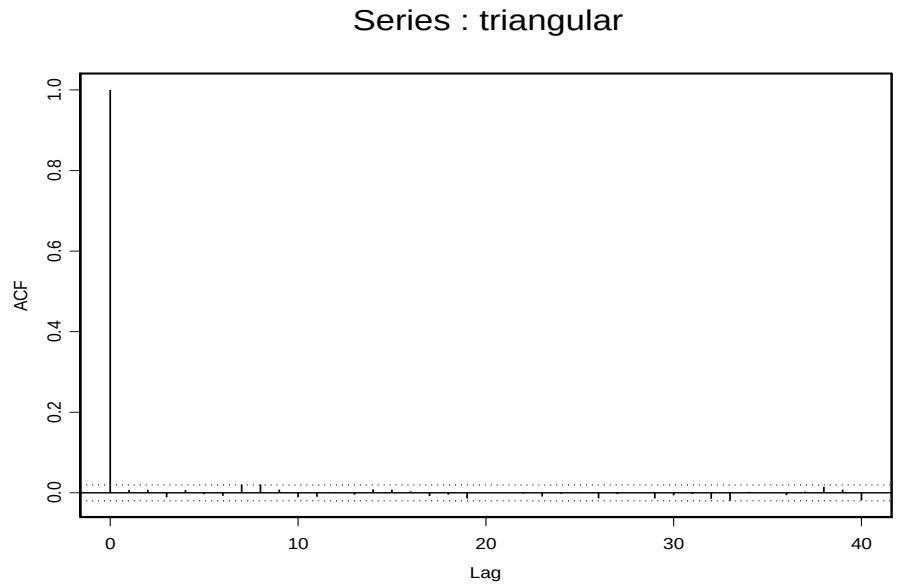

Figure 2.8: Plot of the autocorrelation function based on the posterior output of R_0.

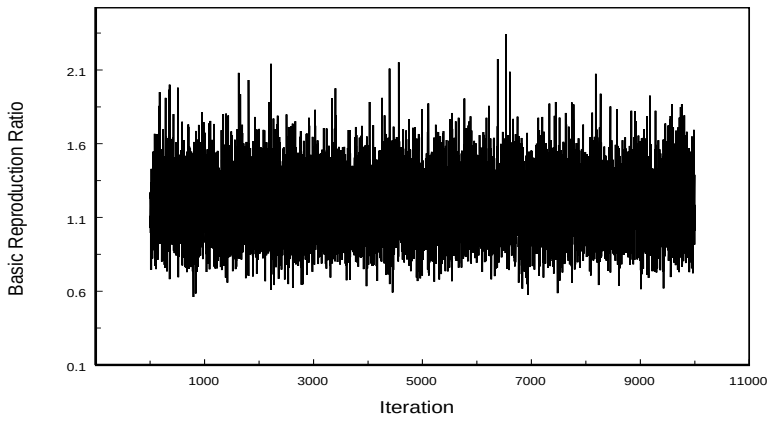

Figure 2.9: Trace plot of the posterior output of R_0.

is particularly obvious from figure 2.6 where again the larger is the (assumed) variance of the infectious period, the larger is the posterior variance of R_0.

The sensitivity of the algorithm to the prior specification was examined using a number of different priors and the results remained virtually unchanged. For illustration we present the output using the Gamma prior that we have used throughout this thesis, and the output based on a Uniform prior over the $(0.0001, 10000)$ interval for the infection rate parameter. Note that this prior restricts the posterior space of R_0 to values below 40000 but this is not an important restriction in practical terms. As can be seen from figure (2.7) the output does not change much and thus, we shall consider the resulting inference as "objective", in the sense that the posterior output is largely determined by the data.

Additionally, we have been exploring the convergence properties of the

Markov chains used through a series of plots of the "trace" and the autocorrelation function of the parameters involved. Two plots of this kind are presented in figures (2.8) and (2.9) for the autocorrelation and the trace respectively, and they are quite satisfactory, particularly the ACF plot that displays negligible autocorrelation for any lag from 1 upwards. Thus, the algorithms we have employed in this chapter appear to have desirable convergence properties.

2.7 Conclusion

We have presented a method for solving the triangular equations that describe the final size distribution of the generalised stochastic epidemic. We therefore have presented methodology that overcomes a well-known problem in the analysis of epidemics. The methods rely on computer intensive algorithms that utilise multiple precision arithmetic. We have presented two applications of the methodology, the evaluation of the precision of limit theorems that appear to be interesting for inference purposes, and statistical inference for the threshold parameter of the epidemic. These methods explore, using exact results, a number of previously obtained probabilistic and statistical results that could only be evaluated in the past using simulation methods. However, the focus of this thesis is towards statistical methodology for realistic (and hence complex) stochastic epidemic models and a method of this kind is considered in the following chapter for the epidemic with two levels of mixing described in the first chapter.

Chapter 3

Approximate Bayesian Inference for Epidemics with two levels of mixing

3.1 Introduction

This chapter describes approximate Bayesian inference for the epidemic model with two levels of mixing presented in chapter two. Given final size data, a set of (vector) triangular equations would have a very large dimension, for all but very small populations. Hence, it is necessary to evaluate probabilities that are often smaller than the machine precision. Hence, the likelihood, the conditional density of the data given a particular set of parameter values, is numerically intractable for most realistic populations. Thus, in order to facilitate inference procedures for the two-level-mixing model, an approximation method is introduced that makes the evaluation of the approximate likelihood possible. The underlying idea of this approximate method of inference is to introduce an appropriate latent random variable for which we have asymptotic

distributional information.

The rest of the chapter is organized as follows. The model and known results that are relevant for inference purposes are presented in section 2 while in section 3 we describe the types of data that we shall consider, the potential difficulties with the associated likelihood and we introduce an approach for approximating the likelihood by augmenting the parameter space. In section 4 we describe the MCMC algorithm used as the inferential tool and in section 5 we apply our methodology to a dataset from an influenza outbreak as well as various illustrative final outcome datasets. In section 6 the inference method is evaluated for the special case where the households have size one, when the model reduces to the generalised stochastic epidemic, and we complete the chapter with some concluding remarks.

3.2 Epidemic models with two levels of mixing

3.2.1 Stochastic Epidemic model

In this section we briefly recall the salient features of the two-level-mixing model and we mention the relevant results that are required for inference purposes from the first chapter where a more detailed description of the model is given. We consider a closed population that is partitioned into groups. Suppose that the population contains m_j groups of size j and let $m = \sum_{j=1}^{\infty} m_j$ be the total number of groups. Then the total number of individuals is $N = \sum_{j=1}^{\infty} j m_j$. Since we have an S-I-R model, a susceptible individual j becomes infectious as soon as he is contacted by an infective and remains so for a time I_j distributed according to the distribution of a specified non-negative random variable I.

The epidemic is initiated at time $t = 0$ by a (typically small) number of

infectives while the rest of the population is initially susceptible. Infective individuals mix at two levels. Thus, while infective, an individual makes infectious population wide contacts at the points of a Poisson process with rate λ_G. Each such contact is with an individual chosen uniformly at random from the N initially susceptibles. Hence, the individual to individual rate is $\frac{\lambda_G}{N}$. Additionally, each infective individual makes person to person contacts with susceptibles in its own household according to a Poisson process with rate λ_L. All the Poisson processes (including the two processes associated with the same individual) and the random variables $I_j, j = 1, \ldots, N$, describing the infectious periods of different individuals, are assumed to be mutually independent.

In what follows we describe some results regarding the final size distribution of a single population stochastic epidemic model where infection from outside is permitted. In this model the global infections are taken into consideration through a fixed probability of avoiding non-local infection, instead of being modelled explicitly. This model is simpler than the two level mixing model from a mathematical point of view. However, it can be used as an approximating model as we shall see in the sequel.

3.2.2 Final outcome of a homogeneous SIR epidemic with outside infection

The Epidemic Model

The model described in Addy *et al.* (1991) is defined for a population that is partitioned into groups. The within group epidemics are modelled using the generalised stochastic epidemic while interactions between groups are not modelled explicitly. However, it is assumed that each individual avoids infection from outside its group independently with probability π. Thus, conditionally

on π, the final outcomes of epidemics in different groups are independent.

The limiting two-level-mixing model

In the two level mixing model the fates of different groups are not independent, the reason being that while infectious, an individual has both local and global contacts. As noted in Ball *et al.* (1997) though, as the number of groups goes to infinity and conditional on the occurence of a major epidemic, a given individual avoids infection from the population at large with probability $\pi = \exp(-\lambda_G z E(I))$, where $z = z(\lambda_L, \lambda_G)$ is the (unique for each (λ_L, λ_G) pair) deterministic proportion of susceptibles who ultimately become infected. This limit can be derived by the solution of a non-linear equation that we describe in 3.7. This limiting behaviour means that we can surmount the complication of the explicit global infections and conduct approximate (in the sense that the number of groups becomes large) inference using the Addy *et al.* (1991) model.

The use of the final severity

We make a refinement by replacing $zE(I)$ with the actual (scaled) final severity $\frac{A}{N}$. The *final severity* is an important final state random variable which we denote by A and equals the total number of time-person units of infection. A is defined as the (random) sum of the infectious periods of the ultimately infected individuals, $A = \sum_{k=1}^{T} I_k$ and is sometimes referred to as *the total area under the trajectory of infectives*. Note that T denotes the (random) final size of the epidemic. The joint distribution of the final size and the final severity was derived for a wide class of epidemic models known as the collective Reed-Frost epidemic, in a series of papers by Lefèvre and Picard (e.g. Lefèvre and Picard (1990)). They have used a non-standard family of polynomials known as Gontcharoff polynomials to assist their algebraic computations, and we now

recall their definition.

Let $U = u_0, u_1, \ldots$ be a given sequence of real numbers. Then the Gontcharoff polynomials attached to U, $G_0(x \mid U), G_1(x \mid U), \ldots$, are defined recursively by the triangular system of equations:

$$\sum_{j=0}^{i} \frac{u_j^{i-j}}{(i-j)!} G_j(x \mid U) = \frac{x^i}{i!}, \qquad i = 0, 1, \ldots. \tag{3.1}$$

A useful property of Gontcharoff polynomials that we shall require (see for example (3.3) in Ball *et al.* (1997)) is

$$G_i^{(j)}(x \mid U) = G_{i-j}(x \mid E^j U), \qquad 0 \leq j \leq i, \tag{3.2}$$

where $E^j U$ is the sequence $U = u_j, u_{j+1}, \ldots$ and $G_i^{(j)}(x \mid U)$ is the jth derivative of $G_i(x \mid U)$. Note that $G_i^{(j)}(x \mid U) = 0$ if $j > i$.

We now focus on the two level mixing model with a large number of groups. Suppose that each of the initial susceptibles in a single group has probability π of avoiding infection from outside the group, independently of the dynamics of the within-group epidemic. Consequently, we cease taking into account the global infection dynamics and the effect of global infections will be instead modelled using π. Then the dynamics of the within-household epidemics can be considered as independent. Assume that a group initially consists of n susceptibles and a infectives. Let $\phi_{n,a}(s, \theta) = E(s^{n-T} \exp(-\theta A)), \theta \geq 0$ be the joint generating function of the group final size and severity, (T, A). Then it follows from Ball *et al.* (1997) that

$$\phi_{n,a}(s, \theta) = \sum_{i=0}^{n} \frac{n!}{(n-i)!} \phi(\theta + \lambda_L i)^{n+a-i} \pi^i G_i(s \mid U), \tag{3.3}$$

where the sequence U is given by $u_i = \phi(\theta + \lambda_L i), i = 0, 1, \ldots$, where $\phi(\cdot)$ denotes the moment generating function of the infectious period. Let $\mu_{n,a} = E(T_n)$ be the mean final size of an epidemic initiated by a infectives in a group

with n susceptibles. Then by differentiating equation (3.3) with respect to s and setting $s = 1$ and $\theta = 0$ it follows that

$$\mu_{n,a}(\pi) = n - \sum_{i=1}^{n} \frac{n!}{(n-i)!} q_i^{n+a-i} \pi^i \alpha_i, \tag{3.4}$$

where $q_i = \phi(\lambda_L i)$ and $\alpha_i = G_{i-1}(1 \mid V)$. Here the sequence V is given by $v_i = \phi(\lambda_L(i+1)) = q_{i+1}$ (for $i = 0, 1, \dots$).

It is also straightforward to obtain the distribution of the final size T_n. Let $p_{kn} = Pr\{T_n = k\}, k = 0, 1, \dots, n$. Then setting $\theta = 0$ in (3.3) and differentiating $n - k$ times with respect to s yields

$$p_{kn} = \frac{1}{(n-k)} \sum_{i=n-k}^{n} \frac{n!}{(n-i)!} q_i^{n+a-i} \pi^i G_{i-n+k}(0 \mid E^{n-k}U), \qquad k = 0, 1, \dots, n, \tag{3.5}$$

where $E^{n-k}U$ is the sequence $q_{n-k}, q_{n-k+1}, \dots$.

In the following we shall use the above within-group exact results to describe asymptotic approximations related to the epidemic over the whole population.

3.2.3 Asymptotic approximations

The threshold parameter

Approximating the initial stages of the epidemic with a branching process yields a threshold parameter that dictates whether or not a major epidemic can occur. Here we summarise the results mentioned in chapter 2. For $n = 1, 2, \dots$, let the proportion $\frac{m_n}{m}$ of groups of size n converge to θ_n as the number of groups $m \to \infty$. Let also $\hat{g} = \sum_{n=1}^{\infty} n\theta_n$ be the asymptotic mean group size and assume that $\hat{g} < \infty$. Then the threshold parameter associated with the two-level-mixing model is defined as

$$R_* = \lambda_G E(I) \nu, \tag{3.6}$$

where $\nu = \nu(\lambda_L) = \frac{1}{\tilde{g}} \sum_{n=1}^{\infty} (1+\mu_{n-1,1}(1)) n\theta_n$ is the mean size of an outbreak in a group, started by a randomly chosen individual, in which only local infections are permitted. The initial infective is also included in ν. Note that $\mu_{n-1,1}(1)$ can be evaluated from (3.4) with $\pi = 1$.

We will now describe results related to a normal approximation for the final state of an outbreak in the case where $R_* > 1$.

Gaussian approximation

The asymptotic distribution of the final size and severity of the two level mixing model was derived by Ball *et al.* (1997) using the embedding representation of Scalia-Tomba (1985). We recall the relevant features of this normal approximation from Ball *et al.* (1997) where additional details can be found.

The representation employed is based on a process describing the infections through "generations". These generations are not necessarily representative of the real time dynamics of the epidemic. However, they provide us with an adequate description of the epidemic that is particularly beneficial for the derivation of the asymptotic distribution of quantities related to the final state of the epidemic.

It is convenient for this construction to think of the infection process by assigning a pair of exponential random variables to each individual, say $Z_L(k)$ and $Z_G(k)$ for individual k, that correspond to the individual's "threshold" to local and global infections. Then, the (exponential with rate λ_L) random variable Z_L represents the total time-units of local infections necessary to *locally* infect a given individual, while the (exponential with rate λ_G/N) random variable Z_G determines the global infections. Consider a group of size n with no initial infectives. For $t \geq 0$, let $R(t)$ and $A(t)$ denote respectively the final size and severity of the epidemic within that group when each individual is exposed

to t units of infectious pressure. Let $(R, A) = \{(R(t), A(t)), t \geq 0\}$. In this construction it is assumed that the infections are instantaneous. Hence, R and A have a simultaneous jump when a susceptible individual gets infected from outside (with probability $1 - \exp(-\lambda_G t/N)$) and they remain constant otherwise. In order to obtain a realisation of (R, A) let us mark the n Z_G values on the t-axis. The first infection (and thus the first jump of (R, A)) occurs at the smallest Z_G. This first infective of the group starts an epidemic among the remaining $n - 1$ susceptibles of the group with final size T, say. Next, we delete the $T - 1$ marks of the new infectives from the t-axis. The next jump of the process occurs at the smallest remaining mark where one of the individuals that was not infected by the epidemic started in the first mark was globally infected. This individual initiates a new epidemic among the other remaining susceptibles, that results in the second jump of the (R, A) process and so forth.

Let us now define with $(R_i(t), A_i(t)), i = 1, \ldots, m$, the (R, A) process of each group, and let $R_\bullet(t) = \sum R_i(t)$ and $A_\bullet(t) = \sum A_i(t)$. Assume that we apply to the initially susceptible population an amount T_0 of infectious time. Then the epidemic is described in terms of generations as follows: The first generation is completed at the (stochastic) time T_1 after the occurrence of the within-group epidemics initiated by the T_0 infectious time-units. The T_1 amount of infection might generated new global infections that may create additional within-group infections. At the end of these new infections there will be in the population $T_2 = T_0 + A_\bullet(T_1)$ infectious time-units. The process stops when the additional infectious time cannot give rise to further global infections, i.e. at the time $T_\infty \equiv \min\{t \geq 0 : t = T_0 + A_\bullet(t)\}$. Thus $R_\bullet(T_\infty)$ and $T_\infty = A_\bullet(T_\infty) + T_0$ represent the final size and severity of the epidemic, respectively.

Then the asymptotic, as the population size goes to infinity, final size of the

epidemic is either small, as obtained from the branching process approximation or, in the case of non-extinction, is normally distributed around an appropriate deterministic limit. The stopping time of the epidemic serves as a means for the formulation of a law of large numbers. Subsequently, the authors derive a Gaussian law around this deterministic limit. We now present these results in detail. Let as define

$$a(t) = \frac{E(I)\sum_{n=1}^{\infty}\theta_n\mu_{n,0}(\exp(-\lambda_G t/N))}{\hat{g}}, \qquad t \geq 0.$$

Then, as the population size goes to infinity, the limit of the mean final severity $\frac{E(A)}{N}$ is defined as the stopping time

$$\tau := min\{t : t = a(t)\},$$

see Ball *et al.* (1997) section 4.2. This limit can also be derived by solving the non-linear equation 3.35 in Ball *et al.* (1997), namely

$$z = 1 - \sum_{n=1}^{\infty}\hat{g}^{-1}\theta_n\sum_{i=1}^{n}\frac{n!}{(n-i)!}q_i^{n-i}\pi^i\alpha_i, \tag{3.7}$$

where the α_i's are the sequence given in (3.4) and $\pi = \exp(-zE(I)\lambda_G)$. Hence, (3.7) is an implicit equation for z. Note that zero is always a solution of (3.7). Additionally, there is a unique second solution of the above equation in which case $\tau = zE(I)$, the latter being a Wald's identity for epidemics, see Ball (1986). Finally define

$$\begin{aligned} \mu &= \mu(\lambda_L, \lambda_G) = a(\tau), \\ \sigma^2 &= \sigma^2(\lambda_L, \lambda_G) = \frac{\sum_{n=1}^{\infty}\theta_n Var(A_n(\tau))}{[\hat{g}(1-a'(\tau))]^2}, \end{aligned} \tag{3.8}$$

where $A_n(\tau)$ is the final severity of an epidemic in a group of n initial susceptibles which started with no initial infectives and the probability of infection from outside is $\pi = \exp(-\lambda_G\tau)$. Note that $Var(A_n(\tau))$ can easily be obtained

by differentiating $\phi_{n,a}(s,\theta)$ given in (3.3) with respect to θ and setting $s=1$ and $\theta=0$. In particular, $Var(A_n(\tau)) = E(A_n^2(\tau)) - (E(A_n(\tau)))^2$ which using (3.2) can be evaluated as

$$E(A_n(\tau)) = \sum_{i=1}^{n} \frac{n!}{(n-i-1)!} q_i^{n-i-1} \pi^i G_i(1 \mid U) + \sum_{i=1}^{n} \frac{n!}{(n-i)!} q_i^{n-i} \pi^i G_{i-1}(1 \mid V),$$

and

$$\begin{aligned} E(A_n^2(\tau)) = &\sum_{i=2}^{n} \frac{n!}{(n-i-2)!} q_i^{n-i-2} \pi^i G_i(1 \mid U) + \sum_{i=2}^{n} \frac{n!}{(n-i)!} q_i^{n-i} \pi^i G_{i-2}(1 \mid W) \\ &+ 2 \sum_{i=1}^{n} \frac{n!}{(n-i-1)!} q_i^{n-i-1} \pi^i G_{i-1}(1 \mid V). \end{aligned}$$

The sequence U is given by $u_i = q_i = \phi(\lambda_L i)$, while the sequences V and W are given by $v_i = \phi(\lambda_L(i+1)) = q_{i+1}$ and $w_i = \phi(\lambda_L(i+2)) = q_{i+1}$ respectively. Also $a'(\tau) = \frac{E(I)\sum_{n=1}^{\infty} \theta_n \mu'_{n,0}(\exp(-\lambda_G \tau))}{\hat{g}}$ with $\mu'_{n,0}(\exp(-\lambda_G \tau)) = \left(n - \sum_{i=1}^{n} \frac{n!}{(n-i)!} q_i^{n-i} \pi^i a_i\right)' = \lambda_G \sum_{i=1}^{n} \frac{n!}{(n-i)!} i q_i^{n-i} \pi^i a_i$.

Then, as stated in section 4.2.2 of Ball *et al.* (1997) it follows that the quantity $\sqrt{m}\left(\frac{A}{N} - \mu(\lambda_L, \lambda_G)\right)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma^2(\lambda_L, \lambda_G)$ as the number of groups $m \to \infty$. Thus, we can approximate the distribution of the final severity with

$$A \sim N\left(N\mu(\lambda_L, \lambda_G), \frac{N^2}{m}\sigma^2(\lambda_L, \lambda_G)\right). \tag{3.9}$$

In the following section we will employ the central limit theorem in (3.9) to approximate the likelihood of the two level mixing model given final size data.

3.3 Data and Augmented Likelihood

3.3.1 Final outcome data

We consider data of the form $\tilde{n} = \{n_{ij}\}$ where n_{ij} is the number of households in which i out of j susceptibles ultimately become infected. We initially assume that the whole population is observed. However, our analysis also applies to the important practical case where the data are recorded only on a fraction of the population. Specifically, a "representative" random sample would have the same proportions for the different group sizes.

In the sequel, we consider the problem of statistical inference for the two infection rates λ_L and λ_G given the final size data $\tilde{n}$. Hence, in a Bayesian framework, we wish to explore the posterior distribution $\pi(\lambda_L, \lambda_G \mid \tilde{n})$ of the two infection rates that are the basic model parameters given the observed data $\tilde{n}$. The Bayesian paradigm is particularly suitable for inference in epidemics, notably because the quantities of interest, e.g. R_*, are functions of the basic model parameters. Hence, in a sampling based approach it is straightforward to sample from these, possibly non Gaussian, important quantities. A classical inference approach typically relies on normality assumptions that are not required in the Bayesian framework. Hence, when sampling functions of the basic model parameters, the procedure adopted here has a natural advantage.

In Bayesian statistics inference concentrates on the posterior distribution of the model parameters. In order to study the posterior distribution of interest we need to be able to evaluate the likelihood $\pi(\tilde{n} \mid \lambda_L, \lambda_G)$ since, by Bayes' theorem,

$$\pi(\lambda_L, \lambda_G \mid \tilde{n}) \propto \pi(\tilde{n} \mid \lambda_L, \lambda_G)\pi(\lambda_L, \lambda_G).$$

For realistic data sets, the distribution of the final size is numerically intractable as we have seen in chapter two. Moreover, in the two level mixing model

taking into account the between household infections increases enormously the dimension of the problem compared to the models that assume homogeneous mixing. For this reason we need to resort to an alternative method of inference.

3.3.2 Augmented Likelihood

The difficulties described above suggest considering some form of imputation that would make the evaluation of the (augmented) likelihood feasible. Here we augment the parameter space using the final severity A, as defined in section two. Since A is the sum of the infectious periods of all the ultimately infected individuals, it turns out that, asymptotically, an individual avoids infection from outside his group during the course of the epidemic with approximate probability $\exp(-\lambda_G A/N)$. Note that in this section the final severity it taken over the whole population, which does not necessarily correspond to the final severity of the individuals that the data refer to.

When the number of households is large we can assume that, conditional on the final severity A, the households can be thought of as independent since they are exposed to (approximately) the same force of infection from outside. Thus, we do not need to model the global contacts explicitly. Instead, we replace the infection dynamics outside the group with a single probability. Hence, we approximate the two-level-mixing model with the model described in Addy *et al.* (1991) where the probability of avoiding infection from outside is $\pi = \exp(-A\lambda_G/N)$. Then p_{kl}, the probability that k out of the l individuals of a group ultimately become infected, can be calculated from the following system of triangular equations described by Addy *et al.* (1991):

$$\sum_{k=0}^{i} \frac{\binom{l-k}{i-k} p_{kl}}{q_{l-i}^{k} \pi^{l-i}} = \binom{l}{i}, \qquad i = 0, 1, \ldots, l. \tag{3.10}$$

where n_{kl} is the number of households of size l with k infected individuals. A

little algebra can show that this definition of p_{kl} agrees with the one given in (3.5). Then the likelihood can be approximated by

$$\tilde{\pi}(\tilde{n} \mid \lambda_L, \lambda_G, A) = L(\lambda_G, \lambda_L, A) = \prod_{l=1}^{\infty} \prod_{k=1}^{l} (p_{kl})^{n_{kl}}. \tag{3.11}$$

It is worth reemphasizing that the formula in (3.11) is approximate. However, the underlying assumptions appear to be reasonable in the supercritical case that is, when a significant proportion of the population is ultimately infected. Then the final size and severity are approximately normally distributed according to the central limit theorem in section 4 of Ball *et al.* (1997). Unfortunately, the authors do not provide a more detailed central limit theorem regarding for instance the within household outbreaks, although a related result is proved in Ball and Lyne (2001). Nevertheless, the formula in (3.9) is attractive because it provides us with an accurate result concerning the exposure of individuals to infection from outside their household. Hence, the *pseudolikelihood* approximation in (3.11) should be a reasonable inferential tool whenever the assumptions of the CLT hold. In practice, we would expect this approach to be a good approximation in the case where the probability that the epidemic dies out quickly is small. In that case, provided that we have a large number of households, the probabilities of within-household epidemics should be approximated reasonably well by the p_{kl}'s in (3.10).

Inference procedures proceed by utilising the pseudolikelihood defined in (3.11) to approximate the true, numerically intractable, likelihood. Let $\pi(\lambda_L)$ and $\pi(\lambda_G)$ be the priors for λ_L and λ_G respectively. Then we can approximate the augmented posterior by considering

$$\tilde{\pi}(\lambda_L, \lambda_G, A \mid \tilde{n}) \propto \tilde{\pi}(\tilde{n} \mid \lambda_L, \lambda_G, A)\pi(A \mid \lambda_L, \lambda_G)\pi(\lambda_L)\pi(\lambda_G), \tag{3.12}$$

where the density $\pi(A \mid \lambda_L, \lambda_G)$ for this model is naturally approximated using the Gaussian approximation of (3.9), since there is not a closed form available

for the exact conditional distribution of $\pi(A \mid \lambda_L, \lambda_G)$. With respect to the case where the dataset consists of a random $\alpha \times 100\%$ sample of the population let us denote by A_s the total severity of the ultimately infected individuals in the sample and by N_s the total number of individuals in the dataset. Then it is easy to see from (3.9) that $E\left(\frac{A}{N}\right) = E\left(\frac{A_s}{N_s}\right)$ and $Var\left(\frac{A}{N}\right) = \alpha Var\left(\frac{A_s}{N_s}\right)$. Hence, in the case where the data are a sample from a larger population, we would expect a less diffuse posterior for A/N, with the estimate becoming "exact" as $\alpha \to 1$. Note however that as $\alpha \to 0$ the assumption that the data are a representative sample of the population becomes stronger, particularly because we assume that the form of the disease spread in the population is the same as in our dataset.

Since we only consider final outcome data, no temporal information is available regarding the disease propagation throughout the population. Hence, we need to make specific assumptions about the distribution of the infectious period. This is not a serious restriction in practice, since for most of the commonly occurring diseases we do have some information about the infectious period and we can use it prior to any data analysis. However, the length of the infectious period implicitly sets a time scale and every interpretation of the results should be presented with respect to this scale, an exception being R_* where the length of the infectious period is taken into account.

3.4 Markov chain Monte Carlo algorithm

We used a single-component Metropolis - Hastings algorithm as described in the first chapter. Thus, the model parameters A, λ_L and λ_G were updated in one block. Since all the three parameters are positively defined the state-space of the Markov chain is constrained in the positive quadrant of $\mathbb{R}^3$. Hence, the

proposal distributions used for the Markov chain updates need to take this constraint into account. Note that we cannot use the Gibbs sampler for the parameter updates because the full conditional distributions $\tilde{\pi}(\lambda_L \mid \tilde{n}, \lambda_G, A)$ and $\tilde{\pi}(\lambda_G \mid \tilde{n}, \lambda_L, A)$ appear to be analytically intractable.

Updating $(\lambda_L, \lambda_G, A)$: A simple Gaussian random walk proposal for the two infection rates was found to be sufficient. Hence, the proposed sample $(\lambda_L^*, \lambda_G^*)$ was generated from a bivariate, negatively correlated normal density centered around the current value (λ_L, λ_G). This is intuitively reasonable since we would expect the two infection rates to be negatively correlated, the rationale being that a within-group infection could always arise as a result of a "global" close contact. Thus, it is possible that two different infection rates $(\lambda_L^1, \lambda_G^1)$ and $(\lambda_L^2, \lambda_G^2)$ can produce identical final outcome data. So we use a negatively correlated proposal to improve algorithmic efficiency (in terms of acceptance rates) since we spend less time proposing low-likelihood values. Consequently, it is natural to consider block-updating of (λ_L, λ_G) since it well known in the MCMC literature that blocking can improve mixing of the resulting Markov chain. For each proposed sample $(\lambda_L^*, \lambda_G^*)$ we calculate the proposed threshold parameter R_*^* using (3.6) and if $R_*^* \leq 1$, the corresponding $(\lambda_G^*, \lambda_G^*)$ is rejected because our methodology is only valid in the event of a major epidemic. Else, for $R_*^* > 1$ we sample the proposed A, A^*, using the normal proposal density that is naturally implied by the model (see 3.9) with $\mu = \mu(\lambda_L^*, \lambda_G^*)$ and $\sigma^2 = \sigma^2(\lambda_L^*, \lambda_G^*)$. The proposed sample $(\lambda_G^*, \lambda_G^*, A^*)$ is then accepted with probability

$$\frac{\tilde{\pi}(\tilde{n} \mid \lambda_L^*, \lambda_G^*, A^*)\pi(\lambda_L^*)\pi(\lambda_G^*)}{\tilde{\pi}(\tilde{n} \mid \lambda_L, \lambda_G, A)\pi(\lambda_L)\pi(\lambda_G)} \wedge 1,$$

where $A \wedge B$ is defined as $\min\{A, B\}$, $\tilde{\pi}(\tilde{n} \mid \lambda_L, \lambda_G, A)$ is the pseudolikelihood defined in (3.11) and $\pi(\cdot)$ denotes the prior density.

Prior specification: The two parameters (λ_L, λ_G) were assumed *a priori* to follow independent Gamma distributions with mean 1 and variance 10000. A

large prior variance appears to be a reasonable choice and sensitivity analysis with respect to the prior distribution of λ_L and λ_G will be presented in the following subsection.

3.5 Application to data

3.5.1 Influenza outbreak data

The data The above methodology was applied to the influenza dataset described below. This is an important dataset because the diagnoses were verified by laboratory tests. We consider the observed distribution of influenza A(H3N2) infections in 1977-1978 and 1980-1981 combined epidemics in Tecumseh, Michigan, see Addy *et al.* (1991). The actual numbers of the ultimately infected individuals are presented in a form that is convenient for analysis in Table 3.1. The actual dataset was, approximately, a 10% sample of the entire population under study. Thus, in the notation used above we have $\alpha = 0.1$. The results presented in this section are mainly for the purpose of illustrating our methodology. Hence, we do not analyse the two datasets from the separate epidemics, nor do we attempt to analyse the full dataset, that includes different types of individuals classified by age. We do however compare our results with those from previous analyses that used simpler models and we comment on the effect of using such models.

Implementation details The algorithm was implemented using Fortran 90 on a mainframe computer. We used a burn-in of 10^4 cycles (sweeps), a sampling gap of 50, and all the results presented are from a sample size of 10^6. The actual run time was of the order of 5000 cycles/sec. All the algorithms were also tested with three more Gamma priors with identical variance and means equal to 10,

	Susceptibles per household				
No. infected	1	2	3	4	5
0	110	149	72	60	13
1	23	27	23	20	9
2		13	6	16	5
3			7	8	2
4				2	1
5					1
Total	133	189	108	106	31

Table 3.1: The Tecumseh influenza data

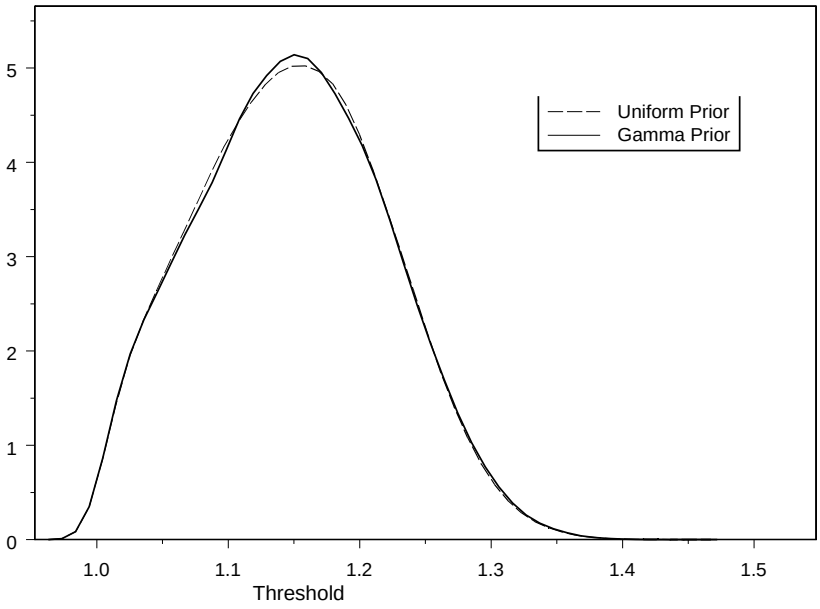

Figure 3.1: Posterior density of R_* for two different priors.

100 and 1000 respectively. Additionally, we used a Uniform prior over the $(0.001, 10000)$ interval for both λ_L and λ_G. The results were virtually identical to the numbers presented here in all cases. For illustration we present the output of the posterior density of the threshold parameter R_* for the Gamma with mean unity and the uniform priors in figure (3.1). This particular output stems from the case where the data available are assumed to represent the whole population. Note that despite the minimum value of R_* lies just above unity, an unconstrained (to be above 1) figure in S-Plus can create the artifact that some of the posterior density appears below 1, due to the way S-Plus draws the kernel density estimators. In conclusion, the results were largely unaffected by the choice of the prior distribution, at least when one uses a large prior variance. Note that our Uniform prior constrains the posterior density of λ_L and λ_G below 10000 but this is not a particularly restrictive assumption in practice.

The convergence of the Markov chains was tested informally with plots of the "trace" of the chain and all the results reported are from chains that appear to have converge to stationarity, like the trace plot of λ_L presented in figure (3.2). Moreover, the plots of the autocorrelation functions from the "thinned" chains show a negligible autocorrelation for lags larger than 1, see for example figure (3.3) for an example of this kind. Thus, the Markov chains for the collection of the algorithms presented in this chapter appear to mix well. It should be noted that it is rather straightforward to (informally) check the convergence of these algorithms since we only have a small number of parameters. Generally MCMC diagnostics are not straightforward and the problem increases when there is a large number of parameters involved.

As was mentioned earlier, since we have final outcome data, we have to make specific distributional assumptions about the infectious period. It is

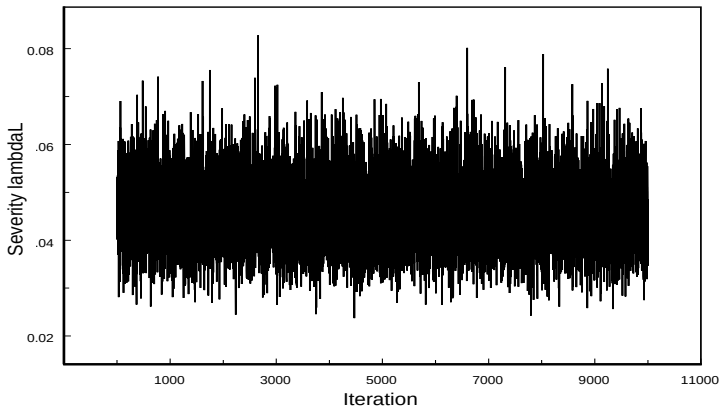

Figure 3.2: Posterior trace plot of λ_L.

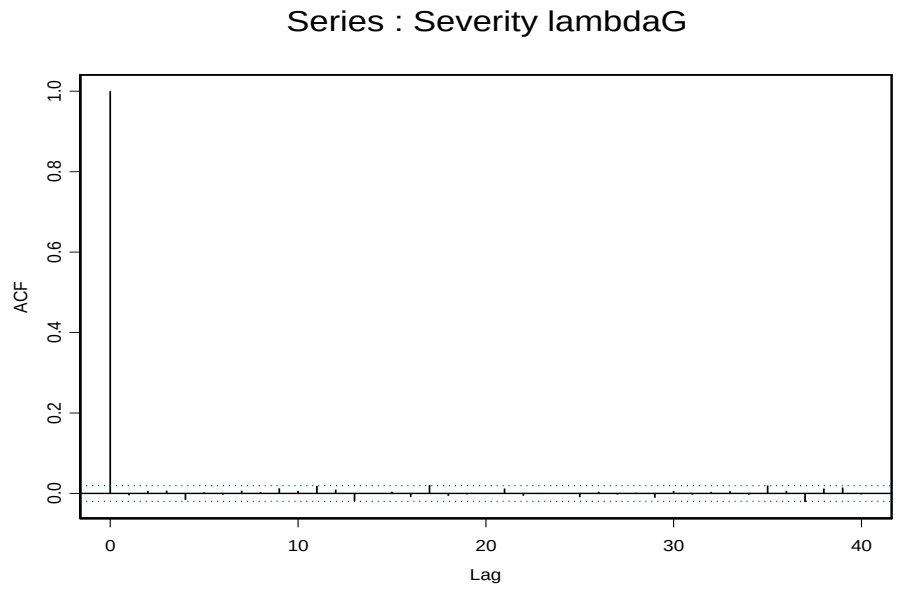

Figure 3.3: Plot of the autocorrelation function for λ_L.

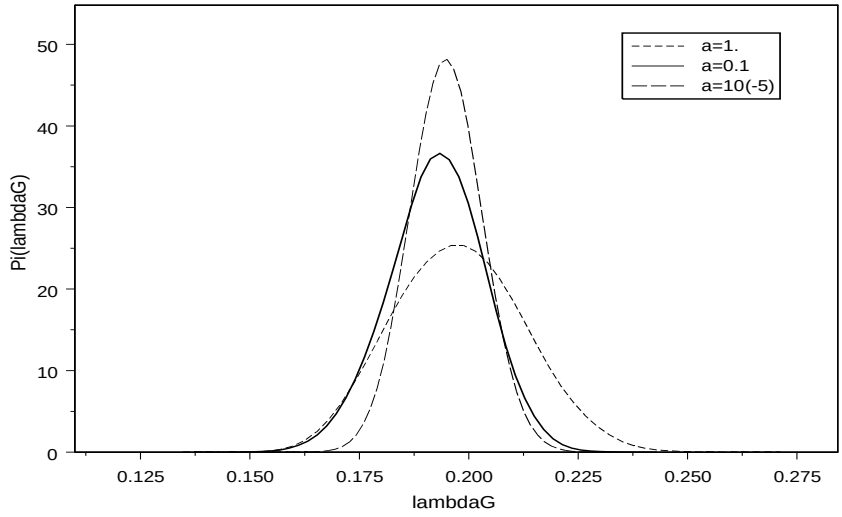

Figure 3.4: Posterior density of λ_G as α varies.

well-known (e.g. Addy *et al.* (1991)) that final size data are typically not sufficient to easily distinguish between different distributions with respect to the infectious period since if the length of I, $E(I)$, is the same, the resulting inferences are similar. Thus, we used the same distribution as in Addy *et al.* (1991) for the infectious period, namely Gamma with mean 4.1 resulting as the sum of two exponential random variables. Moreover, we assumed that we have a "representative" dataset i.e., that the proportions of the different household sizes in our sample are the same with the corresponding proportions in the population. This is not a necessary constraint in our methodology, in practice any choice can be used for θ_n, such as data from the census bureau or data from previous studies, possibly for a larger fraction of the population.

Results We consider the cases $\alpha = 1$, $\alpha = 0.1$ and $\alpha = 10^{-5}$ that correspond to observing the whole population, observing a 10% sample and observing a

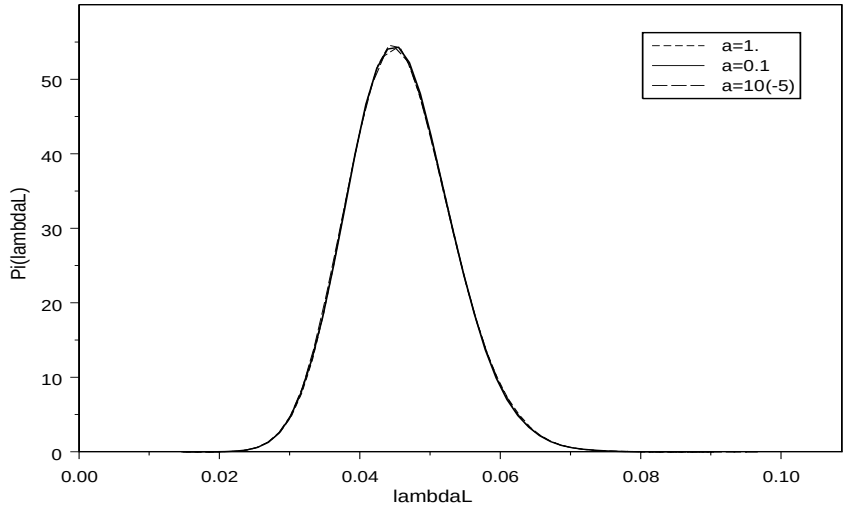

Figure 3.5: Posterior density of λ_L as α varies.

tiny fraction of the population, respectively. The latter case is not very useful in practice but it does serve as a means of comparison with previous analyses where the models used are a special case of the present model for $\alpha \to 0$. Moreover, the results do not change for values of α smaller than 10^{-5}. We will therefore consider this case as $\alpha \approx 0$. Note that in some of the figure annotation a and α are used interchangeably.

The posterior density plot for the global infection rate λ_G of the epidemic for the three different values of α is shown in Figure 3.4 while the corresponding plots for λ_L and R_* are presented in Figures 3.5 and 3.6, respectively.

The two infection rates are clearly negatively correlated and this effect is confirmed by the scatterplot in Figure 3.7. The posterior density plot for R_* appears to be truncated at unity. This occurs because the method we have adopted for inference assumes that a major outbreak has taken place.

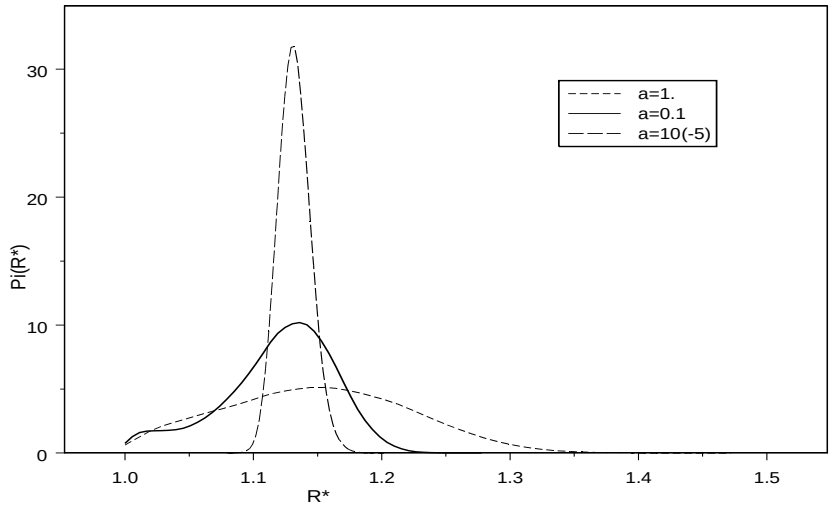

Figure 3.6: Posterior density of R_* as α varies.

Specifically, the algorithm will reject proposed (λ_L, λ_G) values that result in a threshold parameter below unity. It does however look like the case $\alpha = 1$ should have some mass below threshold. Thus, the resulting shape of the posterior density of R_* is in accordance with the methodology we use. In contrast, the (pseudo)likelihood methods used in Ball and Lyne (2003) can result in 95% confidence intervals with the lower limit being below unity, despite the same assumption of $R_* > 1$ being used. This is a well-known advantage of a sampling-based approach to inference for problems of this kind as described in Gelfand *et al.* (1992). We find the posterior correlation of the two basic parameters to be $\rho(\lambda_L, \lambda_G) = -0.5544$. This is intuitively reasonable since for a given dataset one would expect that a decrease in e.g. the local rate would give rise to an increase in the global rate in order to ensure the same number of infections.

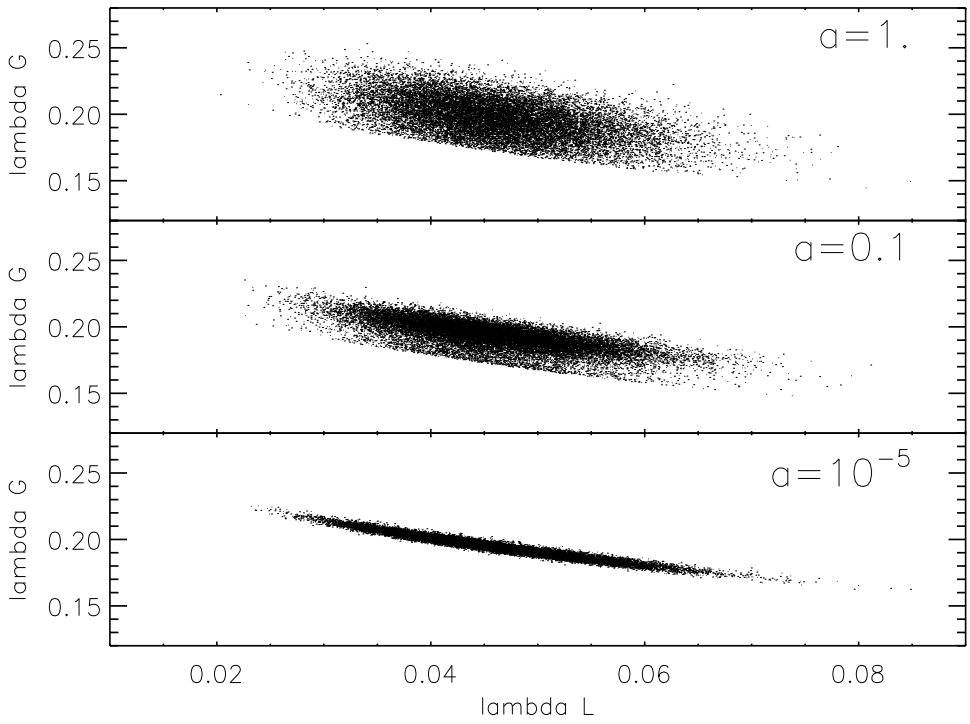

Figure 3.7: Scatterplot of λ_L and λ_G as α varies.

	Parameter			
	λ_L	λ_G	A	R_*
Mean	0.045	0.197	1032.33	1.148
Median	0.045	0.197	1026.36	1.148
S. dev.	0.007	0.014	102.64	0.071
Equal-tailed 95% C. I.	(0.03,0.06)	(0.17,0.26)	(848,1247.9)	(1.02,1.28)

Table 3.2: Posterior parameter summaries from MCMC algorithm using the Tecumseh dataset in the case that $\alpha = 1$.

	Parameter			
	λ_L	λ_G	A	R_*
Mean	0.045	0.193	1053.03	1.114
Median	0.045	0.193	1049.24	1.125
S. dev.	0.007	0.010	87.55	0.044
Equal-tailed 95% C. I.	(0.03,0.06)	(0.17,0.21)	(892,1236.2)	(1.02,1.19)

Table 3.3: Posterior parameter summaries from MCMC algorithm using the Tecumseh dataset in the case that $\alpha = 0.1$.

Posterior density summaries for the parameters of the model associated with the influenza data of Tecumseh for the case $\alpha = 1$ are presented in Table 3.2. It would be rather inappropriate to expect that our results with match previous (classical) analyses, namely Addy *et al.* (1991) and Britton and Becker (2000), since the models they analyse correspond in our framework to the case $\alpha = 0$. However, inference for λ_L is almost identical since our approach also utilises the same pseudolikelihood which can be regarded as a function of π and λ_L alone, independently of α. The results with respect to λ_G and R_* are not similar. In particular, point estimation is in agreement with the previous methods used but the variability associated with these parameters is smaller in the classical approaches. This is essentially the result of using a model that corresponds to a two level mixing model with $\alpha = 0$. We shall comment further on this in the analysis of the two cases with $\alpha < 1$. Furthermore, using an infectious period with mean 4.1 and analysing a dataset with 250 ultimately infected individuals, we would expect the mean of the total severity to result in a value close to 1025. Indeed, the posterior location estimates, particularly the median and the mode, are reassuringly close.

For the case $\alpha = 0.1$ the posterior summaries are given in Table 3.3. The

extent to which the resulting inferences are affected by the assumption $R_* > 1$ reduces as α decreases. The main observation here is that as α reduces, the assumption that we have a "representative" sample becomes stronger. In particular, since the variance of the mean severity $\frac{A}{N}$ in the central limit theorem described in (3.9) reduces with α, so does the posterior variance of $\frac{A}{N}$ and, consequently the variance of λ_G and R_*. Of course the dataset that we use does not change with α. However, it is easy to see from the formula of the pseudo-likelihood (3.11) that the posterior estimation for λ_L and $\pi = \exp(-\lambda_G A/N)$ should not change with α. Indeed, the posterior mean and standard deviation of π for all three cases of α was 0.8677 and 0.0097 respectively. Also note that in all the above tables the posterior summaries for λ_L remain unchanged. Hence, the uncertainty about λ_G (and R_*) as α changes is accommodated through the central limit theorem for the severity. The posterior correlation in this case has naturally increased to $\rho(\lambda_L, \lambda_G) = -0.722$. This was expected since the marginal posterior distribution of λ_G is more peaked. Hence, when we reduce α we effectively assume that we have stronger information about the infectious pressure exerted from outside the group. Consequently, the number of ways that a final outcome data could have arisen decreases, resulting in larger posterior correlation of λ_L and λ_G.

Finally, the case $\alpha = 10^{-5}$ is summarised in Table 3.4. The results in this case will agree with the analyses in Addy *et al.* (1991) and Britton and Becker (2000) since their model, which assumes independent groups, arises in our framework as the special case $\alpha = 0$. In particular, both papers reported in their analysis $\lambda_L = 0.446$ (with a standard error of 0.007) and $\pi = 0.867$ (with a standard error of 0.097) and our results agree with these numbers up to the third decimal point. Hence, the present analysis could potentially be used to "correct" for the variability in the estimation of the global rate and

	Parameter			
	λ_L	λ_G	A	R_*
Mean	0.045	0.194	1039.21	1.160
Median	0.045	0.194	1037.43	1.158
S. dev.	0.007	0.008	76.02	0.024
Equal-tailed 95% C. I.	(0.03,0.06)	(0.18,0.21)	(895,1192)	(1.11,1.16)

Table 3.4: Posterior parameter summaries from MCMC algorithm using the Tecumseh dataset in the case that $\alpha = 10^{-5}$.

consequently the variability of the vaccination coverage required to prevent epidemics in a population partitioned into groups which, as described in Britton and Becker (2000), is a function of the basic model parameters. The posterior correlation for this rather extreme case was $\rho(\lambda_L, \lambda_G) = -0.972$.

3.5.2 Artificial final outcome data

Artificial data can be potentially useful since they allow us to explore the behaviour of the MCMC algorithm in different settings. Here we shall explore the robustness of the proposed methodology by considering two rather extreme artificial datasets.

Example 1

The first artificial dataset can be found in Table 3.5. It has the same number of groups as the influenza dataset from Tecumseh although in this case, despite a rather large number of groups acquiring infection, there are only a few groups where the disease propagates locally. Hence, the dataset consists of a significant number of groups with a single ultimately infected individual. Therefore, we

	Susceptibles per household				
No. infected	1	2	3	4	5
0	40	49	22	20	3
1	93	137	83	80	23
2		3	2	4	2
3			1	1	1
4				1	1
5					1
Total	133	189	108	106	31

Table 3.5: The dataset of the first example

	Parameter			
	λ_L	λ_G	A	R_*
Mean	0.0004	0.40	1346.94	1.66
Median	0.0003	0.39	1358.23	1.64
S. dev.	0.0004	0.04	145.70	0.17
Equal-tailed 95% C. I.	(0.0001,0.0018)	(0.34,0.50)	(1041,1593)	(1.39,2.06)

Table 3.6: Posterior parameter summaries for the data in Example 1.

	Susceptibles per household				
No. infected	1	2	3	4	5
0	1	0	0	0	100
1	249	0	0	0	0
2		0	0	0	0
3			0	0	1
4				0	1
5					48
Total	250	0	0	0	150

Table 3.7: The dataset of the second example

would expect a rather small local infection rate λ_L and a relatively large global rate λ_G.

The output of the MCMC algorithm verifies this, since the global infection rate mainly drives the epidemic. In fact the posterior summaries, presented in Table 3.6, show clearly that the local infection rate is rather negligible. Consequently, the threshold parameter essentially corresponds to the standard basic reproduction ratio. Hence, in household outbreak data where the vast majority of households have only one ultimately infected individual, the generalised stochastic epidemic can be used as a crude approximation of the two level mixing model for inference purposes.

Example 2

In contrast to the previous example, we next created a rather extreme dataset representing a highly infectious disease. Thus, almost no member of the "infected" groups escapes infection. This is a more realistic senario than the first

	Parameter			
	λ_L	λ_G	A	R_*
Mean	0.449	0.298	1344.52	4.65
Median	0.428	0.292	1342.17	4.55
S. dev.	0.117	0.054	230.30	0.85
Equal-tailed 95% C. I.	(0.24,0.81)	(0.21,0.43)	(906.1,1798.9)	(3.11,7.19)

Table 3.8: Posterior parameter summaries for the data in Example 2 using the a Normal random walk proposal.

artificial dataset and it can be thought of as a highly infectious disease such as the 2001 foot and mouth epidemic in the UK. It is also a common situation from many well known diseases of different severity, ranging from the common cold to measles. The actual dataset can be found in Table 3.7. For final outcome data of this kind it is likely that the posterior density of λ_L can have a heavy right tail since the local basic reproduction number is large and there is no temporal information to prevent λ_L from getting very large. Hence, we tuned the component of the (normal) bivariate proposal corresponding to λ_L in order to accommodate for this effect.

The posterior summaries for this analysis are presented in Table 3.8. The main observation here is that the large local infection rate affects the estimation of the threshold parameter drastically. In particular, reporting in this case threshold parameter, $R_0 = \lambda_G E(I) \approx 1.21$ should be regarded as a misleading indicator of the propagation of the epidemic since the actual parameter is amplified by a factor of almost 3. Hence, in highly infectious diseases where the within-group spread of infection is greatly facilitated, it is vital that the epidemic process used to model disease propagation can take into account the additional local spread.

Apart from artificial datasets, simulated data can be particularly helpful due to the fact that they provide us with the ability to test the accuracy of the inference procedures that we use. Three simulated datasets were used for these purposes and this analysis is presented in the following section.

3.5.3 Simulated final outcome data

The simulated data sets of this section were kindly provided by Owen Lyne. These data sets, apart from being a useful tool for evaluating the robustness of our algorithm, enabled us to compare our results with the analysis in Ball and Lyne (2003) wherever it was appropriate. In the cases that we would expect to be above threshold resulting in a symmetric posterior distribution, there is typically approximate agreement in interval estimation, while point estimation remains as described in the analysis of the Tecumseh influenza data, namely that the posterior mode is very close to the maximum pseudolikelihood estimator. All the following examples are based on the same household structure as with the Tecumseh data but with a population that is 10 times larger. The large number of households was utilised to enable greater accuracy of the asymptotics used. We also set $\alpha = 1$ and we use the same, Gamma-distributed, infectious period as before.

We first apply our methodology to a "perfect" dataset that consist of the expected values in each cell. However, since the expected values will not be integers in general, there is a rounding error since the maximum is not exactly at the values of parameters for which the data were calculated. This particular dataset was generated with $\lambda_L = 0.06$ and $\lambda_G = 0.23$, the corresponding threshold parameter being $R_* = 1.46$. The actual data are presented in Table 3.9. As can be seen from the table, the cell values are non-integers which, of course, is never the case in practice. It does however serve as a means of accu-

	Susceptibles per household				
No. infected	1	2	3	4	5
0	865.0089	799.4654	297.1191	189.6626	36.0750
1	464.9911	681.5445	308.6312	217.5999	43.5575
2		408.9901	275.0391	223.5430	47.1339
3			199.2106	231.6536	54.5101
4				197.5409	65.0051
5					63.7184
Total	1330	1890	1080	1060	310

Table 3.9: The "perfect" dataset

rately exploring the precision of our inferential tool and as such it is a useful hypothetical example.

The posterior distributions of interest are summarised in table 3.10. Indeed, the location estimates of λ_L and λ_G are very close 0.06 and 0.23. This is quite reassuring with respect to point estimation. The estimation of the threshold parameter was precise as well, as expected from the posterior distributions of the two infection rates. This dataset was clearly simulated with a threshold parameter above unity. In fact, we would expect that the MCMC algorithm will produce approximately symmetric posteriors since the high posterior probability of the model parameters appears to be sufficiently far from the two regions that result in extreme epidemic outcomes, the second being a severe epidemic where almost all the individuals are ultimately getting infected. In an epidemic of this kind the number of the individuals avoiding infection is Poisson distributed (e.g. Ball and Neal (2004)). Indeed, all the resulting posteriors are fairly symmetric. Also our results agree with the results obtained by

	Parameter			
	λ_L	λ_G	A	R_*
Mean	0.0600	0.2299	26474.8	1.459
Median	0.0600	0.2300	26452.5	1.460
S. dev.	0.0022	0.0050	658.3	0.029
Equal-tailed 95% C. I.	(0.055,0.064)	(0.22,0.24)	(25493,27494)	(1.41,1.51)

Table 3.10: Posterior parameter summaries for the "perfect" data.

	Susceptibles per household				
No. infected	1	2	3	4	5
0	463	196	48	14	2
1	867	871	238	90	16
2		823	468	321	41
3			326	419	109
4				216	102
5					40
Total	1330	1890	1080	1060	310

Table 3.11: The second simulated dataset

Ball and Lyne (2003) for the same dataset. The posterior correlation of the two infection rates was $\rho(\lambda_L, \lambda_G) = -0.439781$, not too close to -1, in accordance with the results for the influenza data when $\alpha = 1$.

The second dataset we considered was simulated with $\lambda_L = 0.001$ and $\lambda_G = 0.4$ with a corresponding threshold $R_* = 1.653$. The rationale here is similar to the first artificial dataset, namely testing the MCMC algorithm in a globally driven epidemic. The actual final outcome data are presented in Table 3.11.

	Parameter			
	λ_L	λ_G	A	R_*
Mean	0.004	0.397	38629.6	1.661
Median	0.003	0.397	38618	1.661
S. dev.	0.003	0.006	647.9	0.024
Equal-tailed 95% C. I.	(0.0001,0.0021)	(0.38,0.41)	(37780,39485)	(1.62,1.70)

Table 3.12: Posterior parameter summaries for the second simulated dataset.

The posterior summaries are presented in Table (3.12). The posterior modes for λ_L, λ_G and R_* were 0.002, 0.4 and 1.67 respectively. This bias with respect to the estimation of λ_L and, to a lesser extent, R_* is expected since our model implicitly assumes that the infection rates are positive. Hence, when we attempt to estimate a positively defined parameter which in reality can actually be zero, we would expect some bias to arise. In this particular case, where we use a full two-level-mixing model, λ_L could be positive even if all the infections were actually global. The explanation for this is similar to the one regarding the posterior correlation of the two infection rates described in the analysis of the Tecumseh data. The results show that the algorithm captures the globally driven nature of this epidemic. Thus, the proposed methodology performs well except in the case where we try to estimate a quantity that is very close to zero with a positively defined parameter.

In contrast, the third dataset we considered was simulated with a more locally driven outbreak representing a highly infectious disease. The actual infection rates used where $\lambda_L = 0.4$ and $\lambda_G = 0.15$ with a corresponding threshold parameter $R_* = 1.687$. We would hope that in this case the posterior distribution of both parameters will be symmetric since the threshold parameter appears to be sufficiently far away from the boundary of unity. The dataset is

	Susceptibles per household				
No. infected	1	2	3	4	5
0	943	929	404	269	59
1	387	235	62	38	10
2		726	68	16	2
3			546	40	2
4				697	2
5					235
Total	1330	1890	1080	1060	310

Table 3.13: The third simulated dataset

	Parameter			
	λ_L	λ_G	A	R_*
Mean	0.394	0.1497	33282.2	1.644
Median	0.394	0.1498	33260.9	1.645
S. dev.	0.012	0.007	838.6	0.032
Equal-tailed 95% C. I.	(0.37,0.42)	(0.135,0.164)	(32122,34470)	(1.59,1.70)

Table 3.14: Posterior parameter summaries for the third simulated dataset.

presented in Table 3.13.

The results, presented in Table 3.14, are indeed very close to the values used for the simulation. Note that the last two datasets are the outcome of a single realisation of a simulated epidemic and, in contrast with the "perfect" data we would not necessarily expect complete agreement of the output with the values used for simulation. However, these values always fall within the 95% credible intervals of the corresponding parameters.

The conclusion from this simulation study is that the methodology appears

to be working reasonably well, at least for point estimation purposes, provided that the real (positively defined) parameters are not too close to the boundary of zero where some overestimation might occur. It would clearly be desirable to test the accuracy of estimating the dispersion of the posterior distribution of the infection rates and this is the subject of the next section.

3.6 Evaluation for the homogeneous case

The likelihood of the two level mixing model is numerically intractable for realistic population sizes as we described earlier in this chapter. This is not necessarily the case for the one level mixing model, as was demonstrated in chapter two. Hence, considering a homogeneously mixing population can allow us to evaluate to some extent the accuracy of the underlying approximations used in the inference method for the two level mixing model.

The two level mixing model reduces to the so-called generalised stochastic epidemic when the households are of size one. This can also be thought of as the general model when $\lambda_L = 0$ and the person to person contact rate between individuals is the population wide rate $\frac{\lambda_G}{N}$. In this case (3.10) becomes

$$\sum_{k=0}^{l} \frac{\binom{N-k}{l-k} p_k}{\left[\phi\left(\frac{\lambda_G(N-l)}{N}\right)\right]^{k+a}} = \binom{N}{l}, \qquad l = 0, 1, \ldots, N, \tag{3.13}$$

where a denotes the number of initial infectives, p_k is the probability that k out of the N initial susceptibles ultimately become infected and ϕ is the moment generating function of the infectious period i.e., $\phi(\theta) = E(e^{-I\theta})$.

We wish to compare the final size probabilities $p_k(\lambda_G)$, as derived by (3.13), to the approximate final size probabilities, say $\tilde{p}_k(\lambda_G)$, evaluated using the final severity. We will attempt to explore the potential differences in a variety of ways.

3.6.1 Exact formula

The approximate final size probabilities can be calculated if we can integrate out the final severity that underlies the approximation. Specifically, we would like to evaluate the probability that x out of the N individuals of the population become infected, as a function of λ_G:

$$I_x^N(\lambda_G) = \frac{1}{\Phi(\frac{\mu}{\sigma})} \int_0^\infty \pi(x \mid \lambda_G, A)\pi(A \mid \lambda_G)dA$$

where Φ denotes the cumulative distribution function of the standard normal,

$$\pi(x \mid \lambda_G, A) = \binom{N}{x} \left(e^{-\frac{\lambda_G}{N}A}\right)^{N-x} \left(1 - e^{-\frac{\lambda_G}{N}A}\right)^x$$

is the likelihood and

$$\pi(A \mid \lambda_G) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(A-\mu)^2}{2\sigma^2}}$$

is the Gaussian distribution of the final severity with $\mu = \mu(\lambda_G)$ and $\sigma^2 = \sigma^2(\lambda_G)$ given by (3.10) for $\lambda_L = 0$. The $\frac{1}{\Phi(\frac{\mu}{\sigma})}$ term comes from constraining the integral of a normal random variable to the positive real numbers since the final severity is positively defined. Then it follows using the binomial theorem that

$$I_x^N = \frac{\binom{N}{x}}{\Phi(\frac{\mu}{\sigma})} \int_0^\infty \exp\left(\frac{-\lambda_G A(N-x)}{N}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(A-\mu)^2}{2\sigma^2}\right) \left(\sum_{k=0}^{x} \binom{x}{k} (-1)^{x-k} \exp\left(\frac{-\lambda_G A(x-k)}{N}\right)\right) dA$$

$$= \frac{\binom{N}{x}}{\Phi(\frac{\mu}{\sigma})} \sum_{k=0}^{x} \binom{x}{k} (-1)^{x-k} \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\right) \left[A^2 - 2A\left(\mu - \frac{\lambda_G\sigma^2(N-k)}{N}\right) + \mu^2\right] a$$

$$= \frac{\binom{N}{x}}{\Phi(\frac{\mu}{\sigma})} \sum_{k=0}^{x} \binom{x}{k} (-1)^{x-k} \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{A^2 - 2A\tilde{\mu} + \tilde{\mu}^2 - \tilde{\mu}^2 + \mu^2}{2\sigma^2}\right) dA,$$

where $\tilde{\mu}(k) = \mu - \frac{\lambda_G \sigma^2 (N-k)}{N}$. Thus, the quantity of interest is:

$$I_x^N = \frac{\binom{N}{x}}{\Phi(\frac{\mu}{\sigma})} \sum_{k=0}^{x} \binom{x}{k} (-1)^{x-k} \exp\left(\frac{\mu^2 - \tilde{\mu}^2(k)}{2\sigma^2}\right) \Phi\left(\frac{\tilde{\mu}(k)}{\sigma}\right).$$

Unfortunately, the above expression is numerically unstable because the terms $\exp\left(\frac{\mu^2 - \tilde{\mu}^2(k)}{2\sigma^2}\right)$ can be very large. Hence, in order to evaluate the validity of the underlying approximation, we will have to utilise simulation methods.

3.6.2 Likelihood comparison for the GSE

Since we cannot evaluate numerically the probability of observing a final size x for a given value of λ_G, we will use three methods to test the simulated likelihood. The simulated (under the approximation method) values will be tested against the exact values, as obtained by solving the triangular equations (3.13), by calculating the total variation distance between the two sequences, performing a statistical distance test and by comparing the two sequences graphically. In the following subsections we describe these three methods.

Simulation

The simulation of the final size probabilities for the generalized stochastic epidemic, when the Gaussian approximation is used for the severity, proceeds in two steps as described below.

For a given $\lambda_G > E(I)^{-1}$ (so that $R_* > 1$)

(i) Sample a final severity A_j from $\phi\left(N\mu(\lambda_G), N\sigma^2(\lambda_G)\right) \mathbb{1}_{\{A>0\}}$ where N is the number of individuals, $\phi(a, b)$ is the density of a normal distribution with mean a and variance b, and $\mathbb{1}_E$ is the indicator function of the event E.

(ii) Use A_j to evaluate a final size T_j from $Bin(N, 1 - e^{-\frac{\lambda_G}{N} A_j})$ where $Bin(N, p)$ is the binomial distribution and p is the probability of "success".

Let $\{T_1, T_2, \ldots, T_M\}$ be the output of the simulation, where M is a large integer. Then, for a given λ_G, the probability q_x that x out of the N initial susceptibles become infected can be approximated by $q_x(\lambda_G) = P[T = x] = \frac{\sum_{j=1}^{M} \mathbb{1}_{\{T_j = x\}}}{M}$.

For comparison purposes we will also consider the approximation considered in Britton and Becker (2000) where essentially $\frac{A}{N}$ is replaced with its deterministic limit $\mu(\lambda_G)$, namely $v(x) = \binom{N}{x} (1 - \exp(\lambda_G \mu(\lambda_G)))^x \exp((N-x)\lambda_G \mu(\lambda_G))$. We will firstly evaluate the accuracy of our approximation using a distance measure.

The Total Variation Distance

The two distributions can be compared using a distance measure. Let X and Y be integer-valued random variables. Then the *Total Variation Distance* (TVD) is defined as

$$d_{TV}(X, Y) = \sum_k |\mathbb{P}(X = k) - \mathbb{P}(Y = k)| = 2 \sup_{A \subseteq \mathbb{Z}} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|.$$

Since the approximation is only valid for the "gaussian part" of the final size distribution it is probably sensible to use a *truncated* TVD for comparison of the two distributions of the final size. Let p_x be the "exact" probability that x out of the N initial susceptibles become infected, calculated using multiple precision arithmetic. Since the initial probabilities correspond to *minor* epidemics we have rescaled the p_x's by dividing by $1-s$, where $s = \sum_{x=0}^{100} p_x$. We have chosen 100 because it appears to be in the middle of the final sizes with extremely low probability. Hence, we will consider the sequence $\tilde{p}_x = \frac{p_x}{1-s}$. Let also q_x be the probability that x out of the N initial susceptibles become infected, calculated using the simulation method described above. The truncated TVD between the "exact" sequence $P = p_1, p_2, \ldots$ and the simulated probability sequence

$Q = q_1, q_2, \ldots$ in this particular case was

$$d_{TV}(P, Q) = \sum_{x=101}^{500} |\tilde{p}_x - q_x| = 0.445273$$

However, for this comparison it is probably more useful to make use of an appropriate test statistic that would give us some quantitative idea about the validity of the approximation.

Graphical comparison of the likelihood values

The results that follow are based on a well known smallpox dataset (see e.g. Bailey, 1975, p.125) where $N = 120$ and $x = 30$. We employed the same Gamma (being the sum of two exponentials) random variable for the infectious period I and we set $M = 10^6$. Although this datapoint (the observed final size) is observed in a relatively small population, the results were qualitatively similar for different infectious periods. Furthermore, as the population size increases the approximation becomes better in the sense that the underestimation of the variance reduces, but the general pattern remains similar to the results presented here.

Figure 3.8 contains plots of the three likelihoods under consideration namely the exact likelihood, evaluated using multiple precision arithmetic and the two approximations based on the central limit theorem for the final severity. The two main observations are (i) the approximations perform quite well with respect to the location estimation and (ii) the variances are underestimated by both approximations, particularly by the more crude one where we impute the deterministic limit of the severity. This is likely to arise due to the fact that both approximations assume that the epidemic is above threshold while the exact inference gives some mass below 1. Hence, ignoring this probability mass results in a more peaked likelihood. Thus, the approximate likelihoods, espe-

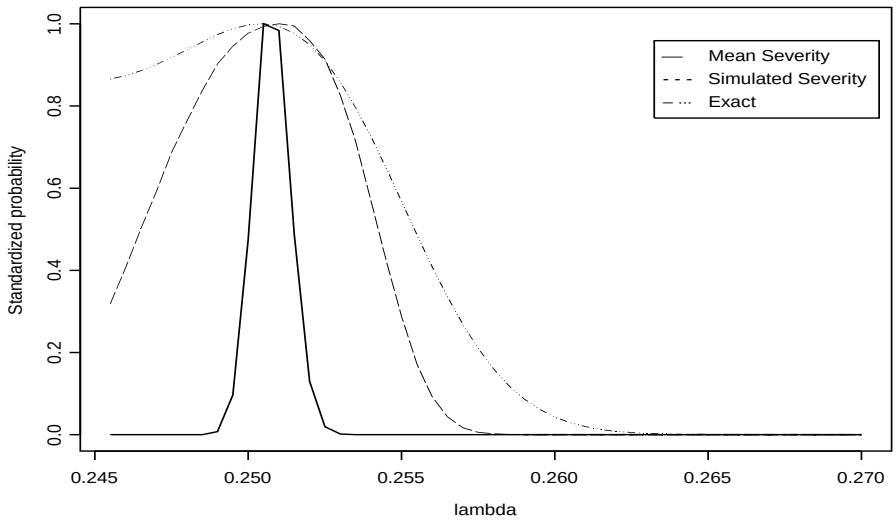

Figure 3.8: Graphical comparison of the likelihood for the three different methods.

cially the mean-based one, are more concentrated around the mode. Naturally, we would expect these findings to be maintained in the inference comparison. Thus, it is likely that the inference resulting from the approximations could underestimate the variance. This is the reason we did not consider inference based on the more crude approximation since ignoring the variance of the severity is expected to give less accurate results.

3.6.3 Inference Comparison

The findings from the likelihood comparison are strongly indicative of the accuracy of the resulting inferences. In particular, for the smallpox dataset used in the previous subsection the mean of λ_G based on the exact and the approximate likelihood was found to be 0.296 and 0.303 respectively. Hence, the results

are in accordance with the likelihood comparison where the estimation of the mean is rather accurate. The corresponding posterior variances of the resulting estimates were 0.0043 and 0.00094. Thus, the pattern of the results remains the same, namely, the variance is underestimated when inference is based on the approximation method.

3.7 Discussion

This chapter presented Bayesian methodology for the approximate analysis of a stochastic epidemic model with two levels of mixing. Real populations are inherently complex and this statistical analysis aims towards capturing an important source of population heterogeneity such as the "local" mixing that facilitates faster spread of a disease. Needless to say, the picture is far from complete. However, our methodology can, in principle, be extended to more complex models such as the overlapping groups (household-workplace or school-household) model described in Ball and Neal (2002), models with additional spatial spread, or multitype household epidemic models (Ball and Lyne (2001)). A crude description of the latter extension is based on the multitype version of the triangular equations described in Addy *et al.* (1991). Specifically, the central limit theorem proved in Ball and Lyne (2001) can be used to derive the (vector) probability of avoiding infection from outside. Then one can evaluate the corresponding pseudolikelihood in a similar manner to the method described here. In practice some identifiability problems may occur for particular datasets. These could be potentially dealt with either by explicitly employing some prior assumptions about the range of a number of infection rates, or by implicitly imposing structural relationships between some infection rates. Both ways appear as possible solutions and this is a worthwhile

extension of our methodology since individual heterogeneities combined with group structure can describe a more realistic population.

The methods described in this chapter assume that we have a large population of households and that the final outcome data come from a major epidemic. The latter is not a serious restriction in practice since outbreak data are typically from a major epidemic. However, it would clearly be beneficial to take into account the bimodal nature of the distribution of the final severity. A natural way is to consider a mixture of two distributions, the first component, in the case that $R_* > 1$, being the normal density used in this chapter, and the second component coming from the branching process described in section three of Ball *et al.* (1997). This refinement should aid towards the correct estimation of the posterior variance for λ_G and R_*.

Moreover, we make use of known asymptotic theory regarding the final outcome of the epidemic model by imputing the final severity. Hence, any generalisation of the above methodology would require the development of appropriate asymptotic results. It would therefore be desirable to develop "exact" methodology that does not rely on large sample results. This appears to be possible if the imputation consists of more information related to the epidemic spread. This is the subject of the following chapter.

Chapter 4

Bayesian Inference for Stochastic Epidemics using Random Graphs

4.1 Introduction

This chapter describes methodology for the statistical analysis of stochastic epidemic models defined on a population with known structure that remains fixed during the course of the epidemic. We shall restrict our attention to the two models described in the first chapter of this thesis, the generalised stochastic epidemic and the epidemic with two levels of mixing. However, our methods can be extended to more general contact structures.

When final outcome data are available the likelihood for these models is hard to compute for all but very small population sizes. Two distinct ways to surmount this problem are presented in chapters two, for the homogeneous case only, and three respectively. The idea that this chapter uses is to impute

appropriate (pseudo)temporal information about the underlying disease progression. In particular, we shall augment the parameter space to introduce a *random graph*, a latent process which essentially describes whom each infective would infect in the absence of other infectives. For every realisation of the random graph we obtain the *out degree* of each individual, i.e. the number of (potentially) "infectious" links emanating from the individual. This information is sufficient to enable the calculation of the likelihood. Subsequently we develop Markov chain Monte Carlo methods to facilitate Bayesian inference for the infection rates and the random graph.

The rest of the chapter is organized as follows. The two models and known results that are relevant for inference purposes are presented in section two as well as the types of data that we shall consider. In section three we describe two different ways for constructing random graphs of the required kind and we derive the likelihood given a simulated random graph. Section four contains the Markov chain Monte Carlo algorithm used to update the graph and the infection rates and in section five we illustrate our methodology using two datasets from real life outbreaks, the first on a homogeneously mixing population and the second among a community partitioned into households. Additionally we examine our methods using various artificial final outcome datasets. In section six the inference method is further evaluated for the generalised stochastic epidemic using exact results obtained with multiple precision arithmetic and we complete the chapter with some concluding remarks.

4.2 Epidemic models and Data

4.2.1 Stochastic Epidemic Models

Epidemic model with two levels of mixing

In this section we recall the notation for the two-level-mixing model which is described in detail in the first chapter. We consider a closed population that is partitioned into m_j groups of size j and $m = \sum_{j=1}^{\infty} m_j$ is the total number of groups. The total number of individuals is $N = \sum_{j=1}^{\infty} jm_j$. The random variables $I_j, j = 1, \ldots, N$ are distributed according to the distribution of a specified non-negative random variable I. Each infectious individual makes population-wide close contacts at the points of a Poisson process with rate λ_G. Each such contact is with an individual chosen uniformly at random from the N initially susceptibles. Hence, the individual-to-individual contact rate is $\frac{\lambda_G}{N}$. Additionally, each infective individual makes person to person contacts with individuals in its own household according to a Poisson process with rate λ_L. All the Poisson processes (including the two processes associated with the same individual) and the random variables $I_j, j = 1, \ldots, N$, describing the infectious periods of different individuals, are assumed to be mutually independent. In the special case where all the households are of size 1 the model reduces to the generalised stochastic epidemic process and we shall now recall the main features of this latter model.

Generalised Stochastic Epidemic

In this model the local contact rate becomes irrelevant. Thus, while infectious, an individual makes contacts at the points of a time homogeneous Poisson process with rate $\frac{\lambda_G}{N}$. We recall results for the final size distribution from

chapter one. Let $\phi(\theta) = E(\exp(-\theta I))$ be the moment generating function of the infectious period I and p_k^N the probability that the final size of the epidemic is equal to k, $0 \leq k \leq N$. Then

$$\sum_{k=0}^{l} \frac{\binom{N-k}{l-k} p_k^N}{\left[\phi\left(\frac{\lambda(N-l)}{N}\right)\right]^{k+m}} = \binom{N}{l}, \qquad 0 \leq l \leq N, \tag{4.1}$$

where m is the number of initial infectives. As we have seen in the second chapter, the system of equations in (4.1) is numerically unstable even for moderate population sizes. One way to overcome this problem was described in chapter two where multiple precision arithmetic was employed to solve (4.1). However, for realistic population sizes the evaluation of the final size probability becomes infeasible. Additionally, when the model is not homogeneous the problem gets worse since the dimension of the system increases enormously. The random graph we shall employ presents a unified approach to inference from final outcome data even for complex population structures. We shall now recall the threshold parameter for the two epidemic models of interest.

The threshold parameter

We will summarise the threshold results mentioned in the first chapter. For $n = 1, 2, \ldots$, let the proportion $\frac{m_n}{m}$ of groups of size n converge to θ_n as the population size $N \to \infty$. Let also $g = \sum_{n=1}^{\infty} n\theta_n$ be the asymptotic mean group size and assume that $g < \infty$. Then the threshold parameter, probably the quantity of the highest epidemiological interest, associated with the two-level-mixing model is defined as

$$R_* = \lambda_G E(I)\nu, \tag{4.2}$$

where $\nu = \nu(\lambda_L) = \frac{1}{g}\sum_{n=1}^{\infty}(1+\mu_{n-1,1}(1))n\theta_n$ is the mean size of an outbreak in a group, started by a randomly chosen individual, in which only local infections are permitted. The initial infective is also included in ν.

For the generalised stochastic epidemic we have $\nu = 1$. Hence, the threshold parameter reduces to the well known basic reproduction number: $R_0 = \lambda_G E(I)$.

4.2.2 Final Outcome Data

We consider data of the form $\tilde{n} = \{n_{ij}\}$ where n_{ij} is the number of households in which i out of j susceptibles ultimately become infected. Hence, in a Bayesian framework, we wish to explore the posterior distribution $\pi(\lambda_L, \lambda_G \mid \tilde{n})$ of the two infection rates that are the basic model parameters given the observed data $\tilde{n}$. For the homogeneous case the data consist of a single point, the final size of the epidemic. From this observation it becomes apparent that we shall attempt to impute a large amount of unobserved information given relatively uninformative data. In order to pursue this purpose, which is a challenging statistical problem, we will use the detailed and complex nature of the stochastic epidemic models under consideration. We shall now describe a random graph that provides us with appropriate information regarding the infection mechanism that could result to the observed final size. This graph is not necessarily representative of the exact temporal dynamics of the epidemic. However, the resulting posterior information is sufficient since we are only concerned with the final outcome in the event of an outbreak.

4.3 Random Graphs and the Likelihood

It has been observed that the final outcome of stochastic epidemic models can be considered in terms of directed random graphs, see e.g. Barbour and Mollisson (1990). We describe in this section how random graphs can be simulated and used for inference purposes. See chapter 7 of Andersson and Britton (2000) for a nice introduction to the characterisation of epidemic models in terms of

random graphs.

4.3.1 The Random Graph

In what follows we describe how to represent the infectious contacts during an outbreak with a random digraph (directed graph). This graph is defined on N labelled vertices that may be partitioned into clusters according to the group structure in the epidemic of interest. In this representation the individuals of the population under study correspond to the vertices of the random graph. Each vertex i has an associated random variable I_i, which is the infectious period of the corresponding individual. Global (population-wide) directed links from i appear with probability $1 - \exp(-\lambda_G I_i/N)$ while local (within-cluster) links emanate from i with probability $1 - \exp(-\lambda_L I_i)$. Conditional upon I_i, the appearance of such links is independent of the appearance, or not, of any other link. Thus, a close contact from individual i, while i is infectious, to individual j corresponds in our random graph representation to a directed link from the vertex that represents i to the vertex that corresponds to individual j. Hence, we shall be using the terms "individual" and "vertex" interchangeably. The same applies for contacts and directed links. Thus an individual j in the epidemic is ultimately infected if and only if there exists a directed path (i.e. a sequence of connected edges) from the initial infective vertex (or vertices) to the vertex j in the graph.

In the sequel we shall be using random graphs of the kind just described as imputed latent variables. It follows that a key challenge is to efficiently construct and update graphs such that the number of vertices that are connected to the vertices that correspond to the initial infectives agrees with the data, i.e is equal to the number of ultimately infected individuals. Note that the random graph essentially describes who each individual would attempt to infect

in the event that they become infectious. If an individual so contacted has not previously been infected, then the attempt is successful; otherwise it has no effect. Furthermore the random graph does not contain real-time temporal information, although it does implicitly contain a description of the outbreak in terms of generations. Since we are only concerned with final size data, this "pseudotemporal" information can be imputed with no loss of generality.

The method we shall employ is concerned with simulating random graphs over the set that contains the "ultimately infected" vertices. For example, if we observe ℓ out of N initial susceptibles infected, we need only construct the graph on ℓ vertices. This approach facilitates the crucial requirement that the graph should agree with the data. Subsequently, it is easy to evaluate (conditional on the current λ's and I's) the probability that the vertices included in G fail to infect the vertices outside this set. This probability is necessary for the calculation of the likelihood.

We shall now introduce some notation. We assume that there is one initial infective labelled κ. This is for illustrative purposes but the methodology we shall describe is straightforward to apply to any (finite) number of initial infectives. Let us denote by ℓ the final size. For the two level mixing case we have:

$$\ell = \sum_{j=1}^{h} \sum_{k=1}^{j} k n_{kj},$$

where h is the maximum group size. Let also $i, i = 1, \ldots, \ell$ be the label of the vertices (including κ) linked to κ via a path of directed links. These vertices correspond to the ℓ ultimately infected individuals. Graphs of the required kind can be constructed in a number of ways. We shall focus on two types of construction mechanisms. However, these two methods, including combinations of them, are far from exhaustive and any construction that results in valid configurations can be potentially useful.

The first of those contains proposals that attempt to construct a completely new random graph in each iteration. We shall refer to mechanisms of this kind as *independent proposals* and we would expect this set of algorithms to work well when the proposal distribution is well calibrated with respect to the likelihood.

The second group of construction mechanisms contains proposals that concentrate on perturbing the existing graph. A variety of ways to achieve this can be used. We shall focus on the case where the proposal is based on either adding a new edge, or deleting one of the current directed links, as long as the remaining infection pathway is still valid. Additionally, particularly for heavily structured populations, moving an edge between two vertices could potentially improve mixing. We shall refer to this class of proposals as *birth-death proposals*. In the following section we describe these two groups of proposal distributions.

4.3.2 The Graph Construction

The Independent Construction

In the first part we shall describe the construction algorithm for the generalised stochastic epidemic model. It is then relatively straightforward to extend this approach to the two level mixing model and this method will be described in the second part of this subsection.

The Independence Sampler in the Homogeneous Case The construction of random graphs of the required kind can be achieved using a stepwise method. Essentially we simulate a realisation of the potential contacts that the ultimately infected individuals would have while infective. This process includes the pathway of infection and a number of contacts from infectives to

already infected or removed individuals that do not result in new infection. The construction is completed in a random number of *generations* γ. Since the probability of a close contact for individual i with infectious period I_i, $p_i = 1 - q_i = 1 - \exp(-\lambda I_i/N)$ is typically small, at each generation, where the total number infected remains less than the observed final size, we use a "special" link in order to preserve the continuity of the algorithm. This special link is drawn before the other links of each generation from a current (within generation) infective to a current susceptible. Hence, each new generation contains at least one link to the remaining susceptibles. Recall that for this construction we are only concerned with vertices that correspond to ultimately infected individuals. However, we shall use the terminology "infected" and "susceptible" in the context of the construction to refer to vertices infected so far, or not yet infected, respectively.

The use of the "special" link is motivated by the need for efficient simulation of the random graph G and it is expected that conditioning on at least one infective per generation should aid towards this direction. Alternatively, a naive approach to this construction would consist of repeatedly sampling potential links until the resulting graph consists of a set of linked vertices that agrees with the final size data set. Such an approach would clearly be highly inefficient.

Let us assume that the epidemic initiates at generation 0, with c_0 infectives and s_0 susceptibles. Without loss of generality we initiate the epidemic with one individual labelled κ. Suppose that the construction is at generation i, $i = 0, 1, 2, \ldots$, and that the epidemic is not complete, i.e. that not all the ℓ vertices have yet had their links assigned to them. If the epidemic is complete then the algorithm terminates. At generation i we pick (among the c_i currently "infective" vertices) the vertex that corresponds to the infective from which the special link emanates, uniformly at random with probability $1/c_i$. In the case

where the infectious periods $I_1, \ldots, I_\ell$ are explicitly included as latent variables, this procedure can be easily modified. In particular, given a realisation of the infectious periods $I_j, j = 1, \ldots, \ell$ we can pick the "special infector" j with probability $I_j/(\sum_{k=1}^{c_i} I_k)$.

Next, we choose the individual infected by the special link uniformly at random among the s_i current "susceptible" vertices. The special link ensures that the following generation will contain $c_{i+1} \geq 1$ infectives.

After selecting the special link, we draw additional links from the c_i currently "infective" vertices. This can be achieved by first determining the number of such links and then deciding which vertices they are linked to. The number of additional links can follow any, appropriately truncated, discrete distribution. We have explored two distinct senarios. In the first case, the number of links from i, additionally to a potential special link, follows a binomial distribution $Bin(\ell - 1, p_i)$ where $p_i = 1 - \exp\left(\frac{-\lambda I_i}{N}\right)$. In the case that we have not imputed the infectious periods the above probability may be replaced by $p = 1 - \exp\left(\frac{-\lambda E(I)}{N}\right)$. In fact, it is not necessary for the random graph construction to employ the actual infection probabilities of the epidemic model p_i. The corresponding link probabilities can be arbitrary. An approach that appears to be worthy of exploration is the use of a modified set of probabilities $p_i^\alpha = 1 - \exp\left(\frac{-\alpha\lambda E(I)}{N}\right)$ that is "adapted" to the infection rate parameter λ. Note that α does not have to be fixed either. In particular, it is straightforward to update α as an extra parameter in the MCMC algorithm that will be used and we shall illustrate this approach in the following section. The intuition behind the use of α can be described as follows. It is well known that independence samplers can perform badly when the proposal distribution is badly calibrated with respect to the posterior. Hence, the inclusion of α essentially introduces a whole family of proposal distributions, with each α value result-

ing in a different probability of proposing the same graph. Thus, we would hope that α will converge towards values that offer reasonable calibration of the proposal distribution. Note that if the original (without α) proposal is well calibrated then it is always possible that α will oscillate around $\alpha = 1$ values.

The binomial distribution seems like a natural choice implied by the model but is not necessary. We also used a constrained Poisson distribution where the constraint is determined by the number of potential edges. The rate of the Poisson distribution ξ can be either fixed or "adapted" to λ. The latter case can, in principle, aid towards faster mixing of the resulting MCMC algorithm.

Once the number of links has been obtained the actual links are assigned uniformly at random among the possible choices. In practice an efficient way to draw d links to e target vertices is to randomly permute the e vertices and consequently to draw a link to the first d of them.

For each of the c_j infectives in generation j let us denote with f_{ij} the number of *forward* links from infective i, i.e. the number of links from i to the vertices that are not yet "infected" at the start of the generation. Let also b_{ij} be the number of *backward* links from i to the current "infective" vertices of generation j, or to the "removed" vertices i.e., the vertices of the generations $c_0, c_1, \ldots, c_{j-1}$. This distinction is essential for the calculation of the proposal probability. Then $f_j = \sum_{i=1}^{c_j} f_{ij} \geq c_{j+1} \geq 1$ with the convention that $f_j = 0$ in the last generation. Similarly $b_j = \sum_i b_{ij}$. Note that $f_j + b_j = \sum_{i=1}^{c_j} \delta_i$ is the sum of the out-degrees, δ_j, of all the vertices that comprose generation c_j. In order to assist the exposition we supress α from p_i^α and we denote both the standard (implied by the model) and the adapted probability by p. In general, the proposal probability is evaluated by summing over all possible ways that could result in a particular configuration. Hence, the probability of proposing

the event of generation j, conditional on λ and α is

$$q(\mathcal{G}_j \mid \lambda, \alpha) = \sum_{k=1}^{f_j} \frac{1}{s_j}(1-p)^{f_j-1}(1-p)^{b_j}p^{mc_j-b_j-f_j} = \frac{f_j}{s_j}(1-p)^{f_j+b_j-1}p^{mc_j-b_j-f_j}$$

For the last generation when $f_\gamma = 0$ (and $\delta_\gamma = b_\gamma$) we get $q(\mathcal{G}_\gamma) = (1-p)^{b_\gamma}p^{mc_\gamma-b_\gamma}$.

Thus, the probability of proposing the random graph $\mathcal{G}$ is

$$q(\mathcal{G} \mid \lambda, \alpha) = \prod_{j=1}^{\gamma} q(\mathcal{G}_j) = \left(\prod_{j=1}^{\gamma-1} \frac{f_j}{s_j}(1-p)^{f_j+b_j-1}p^{mc_j-b_j-f_j}\right)(1-p)^{b_\gamma}p^{mc_\gamma-b_\gamma} \tag{4.3}$$

The procedure described here implies that $\gamma \leq \ell$. The random variable γ can be thought of as a measure of the *connectivity* of G. There is a number of measures of the connectivity of random graphs (e.g. Bollobás (2001)) and here we use γ as a convenient measure of this kind.

The Independence Sampler for the two level mixing model The construction in this case is tailored towards the two-level-mixing behaviour. In particular, the probability distribution of the directed links emanating from a given vertex is a mixture of two components. The first component of the mixture refers to the probability of a potential link to the *local* vertices, i.e. the vertices that belong to the group of the vertex of interest. The second component corresponds to the (typically smaller) probability of a *global* directed link, i.e. an edge to any of the ℓ vertices that correspond to the final size. This construction could aid towards faster convergence of the algorithm since it attempts to mimic the actual infection process induced by the model.

Without loss of generality, we assume throughout the section that individuals 1 up to n_{11} correspond to the individuals of the n_{11} households with the only individual of the household being finally infected. The individuals $n_{11}+1$

up to $n_{11} + n_{12}$ correspond to the infected individuals in the n_{12} households with only one of the two members of the household being ultimately infected while the individuals $n_{11} + n_{12} + 1$ up to $n_{11} + n_{12} + n_{13}$ correspond to the infected individuals in the n_{13} households with only one of the three members of the household being ultimately infected and so forth. We recall that h is the size of the largest household in the data. The labelling ends with individuals $\ell - hn_{hh} + 1, \ldots, \ell$ that reside in the n_{hh} households with all their members being infected at the end of the epidemic. Furthermore, in addition to the s_G current susceptibles in the graph population that each of the c_j infectives in generation j can infect, assume that there are s_{Li} susceptibles in i's group.

We can pick a special link in a similar manner to the homogeneous case using a three-stage procedure. The first step is to pick the vertex that corresponds to the infective from which the special link emanates. We sample this particular individual according to the length of his infectious period, if known. Hence, each currently infective individual i has probability $\frac{I_i}{\sum_{j=1}^{c_i} I_j}$ of being selected. In the case where there is no information about the length of the infectious period, this step can be replaced with picking one of the c_i current infectives uniformly at random.

The second step is to choose between local or global infection and we do so with corresponding probabilities $\frac{s_{Li}p_{iL}}{s_{Li}p_{iL}+s_G p_{iG}}$ and $\frac{s_G p_{iG}}{s_{Li}p_{iL}+s_G p_{iG}}$, where $p_{iL} = 1 - q_{iL} = 1 - \exp(-\lambda_L I_i)$ is the probability of a local infectious contact and $p_{iG} = 1 - q_{iG} = 1 - \exp(-\lambda_G I_i/N)$ is the probability of a global infectious contact. In the case where a realisation of the infectious periods is not available, we can replace I_i in p_{iL} and p_{iG} with $E(I)$. Consequently, we choose the special link uniformly at random among the s_{Li} (s_G) local (global) susceptibles.

After selecting the special link, we draw additional local and global links from the c_i currently "infective" vertices according to their corresponding prob-

abilities. This method corresponds to a binomial distribution $Bin(n_{Li}, p_{Li})$ where $p_{Li} = 1 - \exp(-\lambda_L I_i)$ and n_{Li} is the final size in i's group. In the case that i denotes a vertex from which a local special link emanates, we draw additional local links from a binomial $Bin(n_{Li} - 1, p_{Li})$ distribution. Likewise, the number of global links from i follows a binomial distribution $Bin(\ell, p_{Gi})$ ($Bin(\ell - 1, p_{Gi})$ if i had a global special link) where $p_{Gi} = 1 - \exp\left(\frac{-\lambda_G I_i}{N}\right)$. We pick the d out of e vertices that denote the destinations of the directed links using the procedure described above i.e., by permuting the e potential link "receivers" and choosing the first d of them.

Arguing as in the homogeneous case it is straightforward to extend this approach to more general link probabilities. Hence, the probability of a local and global link can be replaced by $p^{\alpha}_{Li} = 1 - \exp(-\alpha\lambda_L I_i)$ and $p^{\alpha}_{Gi} = 1 - \exp\left(\frac{-\alpha\lambda_G I_i}{N}\right)$ respectively, where α is a parameter that can be updated in the MCMC algorithm.

Similarly to the homogeneous case we have used an alternative possibility for the distribution of the additional number of links. Thus, the extra number of links is determined by a pair of constrained Poisson distributions with rates ξ_L and ξ_G for the local and global contacts respectively. As before, the Poisson rates ξ_L and ξ_G can be either fixed or "adapted" to λ_L and λ_G. The procedure is repeated until all the ℓ vertices are linked to the initial infective κ.

We now describe the procedure for obtaining probabilities of proposing random graphs when an "independence"construction is used. The key idea is that we should assign a unique probability to each graph configuration and the sum of these probabilities over the large (but finite) set of graphs should equal unity. We shall illustrate the approach in the case of the constrained Poisson distribution.

We will now introduce some notation. Let X_i^{Lf} denote the random variable

the realisations of which provide us with the number of forward within-group links from vertex i, i.e., the local links to the within generation susceptibles. Then we denote by x_i^{Lf} the realisations of X_i^{Lf}. In a similar manner X_i^{Gf} denotes the random variable that is concerned with the global forward links from i, X_i^{Lb} is the random variable that describes the local backward links and X_i^{Gb} is the random variable whose realisations give the global backward links. Note that for the graph construction forward links refer to links to within-graph susceptible vertices i.e., vertices that will ultimately get connected to the initial infective. Let us also denote by s_i^L and s_i^G the within-generation local and global susceptibles. Additionally, we denote the final size in i's group with n_i since it is only the final size that matters in the graph construction, as opposed to the actual group size that is important for the evaluation of the likelihood. Finally, let $\mathcal{F}_i$ be the event that i is the special infector i.e., the vertex from which the special link emanates. Then the probability of proposing $\mathcal{G}$ is given by

$$q(\mathcal{G} \mid \xi_L, \xi_G) = \prod_{j=1}^{\gamma} q(\mathcal{G}_j \mid \xi_L, \xi_G),$$

where

$$\begin{aligned} q(\mathcal{G}_j \mid \xi_L, \xi_G)) \quad &= \textstyle\sum_{i:x_i^{Lf}\geq 1} Pr(\mathcal{F}_i) Pr(i \text{ local} \mid \mathcal{F}_i) Pr(\mathcal{G}_j \mid i \text{ local}, \mathcal{F}_i) \\ &+ \textstyle\sum_{i:x_i^{Gf}\geq 1} Pr(\mathcal{F}_i) Pr(i \text{ global} \mid \mathcal{F}_i) Pr(\mathcal{G}_j \mid i \text{ global}, \mathcal{F}_i), \end{aligned} \tag{4.4}$$

where "i local" (global) means that the local (global) special link emanates from i. After a little algebra the probability of proposing generation j, $q(\mathcal{G}_j \mid \xi_L, \xi_G))$, can be rewritten as

$$q(\mathcal{G}_j \mid \xi_L, \xi_G) = \mathcal{H}_1\mathcal{H}_2\mathcal{H}_3\mathcal{H}_4\mathcal{H}_5 \frac{1}{\sum_{k=1}^{c_j} I_k} \left(\sum_{i:x_i^{Lf}\geq 1} I_i\mathcal{H}_6 + \sum_{i:x_i^{Gf}\geq 1} I_i\mathcal{H}_7 \right)$$

where the $\mathcal{H}_i$'s, $i = 1, \ldots, 7$ denote the following probabilities:

$$\mathcal{H}_1 = \prod_{k=1}^{c_j} Pr_{\xi_L}(X = x_k^{Lf})$$

$$\mathcal{H}_2 = \prod_{k=1}^{c_j} Pr_{\xi_L}(X = x_k^{Lb})$$

$$\mathcal{H}_3 = \prod_{k=1}^{c_j} Pr_{\xi_G}(X = x_k^{Gf})$$

$$\mathcal{H}_4 = \prod_{k=1}^{c_j} Pr_{\xi_G}(X = x_k^{Gb})$$

$$\mathcal{H}_5 = \binom{s_k^L}{x_k^{Lf}}^{-1} \binom{n_k - s_k^L}{x_k^{Lb}}^{-1} \binom{s_k^G}{x_k^{Gf}}^{-1} \binom{\ell - s_k^G}{x_k^{Gb}}^{-1}$$

$$\mathcal{H}_6 = \frac{p^L s_i^L}{p^L s_i^L + p^G s_i^G} \left(PS_{\xi_L, s_i^L} - 1\right) \left(PS_{\xi_L, s_i^L}\right)^{c_j - 1} \left(PS_{\xi_L, n_i - s_i^L}\right)^{c_j}$$

$$\mathcal{H}_7 = \frac{p^G s_i^G}{p^L s_i^L + p^G s_i^G} \left(PS_{\xi_G, s_i^G} - 1\right) \left(PS_{\xi_G, s_i^G}\right)^{c_j - 1} \left(PS_{\xi_G, \ell - s_i^G}\right)^{c_j}.$$

Hence, $\mathcal{H}_1$ denotes the probability that the Poisson(ξ_L) random variable X takes the value x_k^{Lf}. The partial sums $PS_{\xi,s}$ arise because of the constrained Poisson samples. Hence, the conditional (Poisson with rate ξ) probability $Pr_\xi(X \mid X \leq s)$ produces the partial sums

$$PS_{\xi,s} = Pr_\xi(X \leq s) = Pr_\xi(X = 0) + Pr_\xi(X = 1) + \cdots + Pr_\xi(X = s) = e^{-\xi} \sum_{k=0}^{s} \frac{\xi^k}{k!}.$$

In the following section we describe a different approach to derive the required graph.

The Birth-Death Construction

We shall focus in this section to an approach that appears to be more promising as far as mixing of the corresponding MCMC algorithm is concerned. The rationale behind this is that MCMC methods based on independence samplers

can suffer from poor convergence properties. The key idea is to start with a given configuration of the random graph and to attempt a perturbation of the current graph at each iteration. Naturally, the way we initialise the graph is not essential because when the Markov chain reaches equilibrium the algorithm will not "remember" the initial configuration. Hence, we can choose any initial graph, G_0 say, as long as all the ℓ vertices that correspond to the final size are linked to the initial infective κ either directly or through a chain of directed links. For example, for data sets where there is a significant number of individuals that ultimately escape infection, a suitable initialisation is to construct a graph where each individual infects one other, until all ℓ vertices are infected. Additionally, for datasets where a large proportion of the individuals is ultimately infected, we can initialise the graph using a "complete" graph, i.e., a graph where each individual has contacts with every "ultimately infected" vertex both within the group and in the population.

Given a current configuration of the random graph G, we attempt to update the graph by either adding a directed link, or deleting one of the existing edges of the graph. In particular, we pick addition or deletion uniformly at random. Let us denote by δ_j, the out-degree of individual j. Additionally, for the two level mixing model, we denote by δ_{Gj} the number of global links i.e., the directed links to any of the ℓ vertices and δ_{Lj} the number of edges to the vertices that correspond to the individuals that reside in j's group. Obviously the out degree of each individual admits the decomposition $\delta_j = \delta_{Gj} + \delta_{Lj}$, and posterior information about δ_{Gj} and δ_{Lj} can provide us with a better understanding of the infection process.

Adding a link We now describe a proposal mechanism for transition from a graph G, with the total number of directed links being $\sum_{j=1}^{\ell} \delta_j$, to a graph G' with total number of edges equal to $\sum_{j=1}^{\ell} \delta_j + 1$. The addition of an edge

is always possible as long as $\sum_{j=1}^{\ell} \delta_j$ does not exceed the maximum number of potential links.

Adding a link in the Homogeneous Case In the homogeneous case the total number of potential directed links is $\ell(\ell-1)$. The actual number implied by the model is ℓ^2 since each individual has infectious contacts with every member of the population, including the individual itself. However, in the random graph representation we can, equivalently, restrict our attention to the case where each individual has potential contacts with the remaining $\ell-1$ individuals. Hence, when $\sum_{j=1}^{\ell} \delta_j$ is strictly smaller than $\ell(\ell-1)$ we pick the edge to be added uniformly at random among the $\ell(\ell-1) - \sum_{j=1}^{\ell} \delta_j$ potential links. In practice a simple device to achieve this is to start from 0 and sequentially add the $\ell - 1 - \delta_j$ number of non-links for each individual j until we reach the uniform random number that indicates the link to be added. The probability of adding a specific link is simply $\left(\ell(\ell-1) - \sum_{j=1}^{\ell} \delta_j\right)^{-1}$ since we add the link uniformly at random and there are $\ell(\ell-1) - \sum_{j=1}^{\ell} \delta_j$ possible choices. This procedure can easily generalised to more complex senarios and we shall describe the two level mixing case in the next paragraph.

Adding a link in the Epidemic with Two Levels of Mixing The stochastic epidemic with two levels of mixing naturally offers two distinct options for adding a link. Hence, each individual that belongs to a group with at least two infectives can have both local and global infectious contacts. Thus, in addition to the $\ell(\ell-1)$ potential global links, there are $\sum_{j:n_{Lj}\geq 2} n_{Lj}(n_{Lj}-1)$ potential local links, where n_{Lj} is the final size in j's group. Similarly to the homogeneous case, an addition is always possible as long as $\sum_{j=1}^{\ell} \delta_{Gj} < \ell(\ell-1)$ or $\sum_{j=1}^{\ell} \delta_{Lj} < \sum_{j:n_{Lj}\geq 2} n_{Lj}(n_{Lj}-1)$, for global and local additions respectively.

If the strict inequality that refers to the global (local) links is satisfied then

we choose to add a global (local) link. When both of these inequalities are satisfied we need to choose the addition of a local or global link. The probability of choosing local, p_L say, can be determined by the user. We have used a number that depends on some preliminary estimation of the local and global reproduction numbers, the rationale being that the reproduction number can be thought of as the average number of infectious contacts per individual. However, the probability of choosing local or global can depend on other measures such as the total number of potential local and global links. In practice, since we use the same probability for choosing to delete a local or global links, this choice is not critical for the behaviour of the algorithm. Finally, we choose the edge to be added uniformly at random among the $\sum_{j:n_{Lj}\geq 2} n_{Lj}(n_{Lj} - 1) - \sum_{j=1}^{\ell} \delta_{Lj}$ and $\ell(\ell-1) - \sum_{j=1}^{\ell} \delta_{Gj}$ local and global potential links respectively. Hence, the probability of adding a given local (global) link is $p_L / \left(\sum_{j:n_{Lj}\geq 2} n_{Lj}(n_{Lj}-1) - \sum_{j=1}^{\ell} \delta_{Lj}\right) \left((1-p_L)/\left(\ell(\ell-1) - \sum_{j=1}^{\ell} \delta_{Gj}\right)\right)$. In the next paragraph we shall describe a mechanism for deleting an edge.

Deleting a link The object of interest is a proposal mechanism for transition from a graph G, with the total number of edges being $\sum_{j=1}^{\ell} \delta_j$, to a graph G' with total number of directed links equal to $\sum_{j=1}^{\ell} \delta_j - 1$. The deletion of an edge is always possible as long as $\sum_{j=1}^{\ell} \delta_j \geq \ell - 1$. Additionally, in the two level mixing senario, the number of global directed links should exceed the number of groups with ultimately infected individuals minus one, since at least one member of such groups must be infected globally. These observations are only useful for efficiency since in practice each potential update will be controlled by a procedure that confirms the validity of the proposed infection pathway.

Deleting a link in the Generalised Stochastic Epidemic This step is executed in a similar fashion to the addition step with the proposed link

to be deleted chosen uniformly at random among the $\sum_{j=1}^{\ell} \delta_j$ current directed edges of the graph. Note that again we pick the link to be deleted, as opposed to choosing an individual, introducing the correct bias in the sense that the greater number of directed links an individual has, the more likely it is that he is elected to "lose" a link. The probability of selecting a particular link is simply $\left(\sum_{j=1}^{\ell} \delta_j\right)^{-1}$. One of the main advantages of this construction method is that it is straightforward to generalise to more complex population structures and in the following paragraph we shall describe the two level mixing case.

Deleting a link in the Epidemic with Two Levels of Mixing Once we have chosen to delete a link, we delete a local directed link with the same probability that we choose to add a local link. Consequently we pick the actual directed link to be deleted uniformly at random among the $\sum_{j=1}^{\ell} \delta_{Lj}$ ($\sum_{j=1}^{\ell} \delta_{Gj}$) current local (global) edges. Hence, the corresponding probability of deleting an edge is $p_L / \left(\sum_{j=1}^{\ell} \delta_{Lj}\right)$ and $(1-p_L) / \left(\sum_{j=1}^{\ell} \delta_{Gj}\right)$ for local and global links respectively.

Regardless of the way we obtain the random graph required, the information that this imputed stochastic process contains is sufficient to evaluate the likelihood, the function of the data for a given value of the parameters governing the epidemic. We shall derive the necessary formulas in the next section.

4.3.3 The Likelihood

Once a particular configuration of the random graph has been obtained, we can derive the likelihood of the data conditional upon the current infection rates. We shall present two different versions that correspond to different levels of information being available. The first approach does not require the actual infectious periods since the likelihood is derived by taking expectations i.e., by

integrating over the infectious periods space. However, in a sampling based approach this is not necessary. Specifically, a sample of the infectious periods can be obtained by sampling from the prior, a method that is commonly used in multiple imputation methods. Subsequently, the space of the infectious periods will naturally be explored by the MCMC sampler.

The methods are actually equivalent since in both cases we assume that we know the distribution of the infectious period. In particular, in the absence of temporal information, the mean of the infectious period sets a scale with respect to which all the results, including the infection rates, should be interpreted. An exception is the threshold parameter R_* where the length of the infectious period $E(I)$ is already taken into account.

The likelihood is an integral part of Bayesian inference. However, the methods we describe could also be possibly useful in a simulated likelihood framework. We shall now derive the formula for the likelihood in the homogeneous case.

The Likelihood for the Generalised Stochastic Epidemic

In a Bayesian approach, the target density can be written as

$$\pi(\lambda \mid \ell) \propto \pi(\ell \mid \lambda)\pi(\lambda).$$

Since the likelihood $\pi(\ell \mid \lambda)$ can be extremely difficult to calculate, we augment the parameter space using a random graph as described above. Thus, the augmented posterior is

$$\pi(\lambda, G \mid \ell) \propto \pi(\ell \mid G, \lambda)\pi(G \mid \lambda)\pi(\lambda) \tag{4.5}$$

Note that $\pi(\ell \mid G, \lambda) = \mathbb{1}_{\{G \in \Gamma\}} Pr(\omega)$ where $\mathbb{1}_{\{A\}}$ is the indicator function of the event A, $\Gamma = \{G : \forall j \in \{1, \ldots, \ell\}, \kappa \rightarrow j\}$, $i \rightarrow j$ means that there

is a path from i to j and $Pr(\omega)$ denotes the probability that no links exist from the ℓ ultimately infected individuals to the remaining $N-\ell$ members of the population that ultimately escape infection. Given a realization of G, $\mathcal{G}$ say, where $\mathcal{G} \in \Gamma$, the essential likelihood $L(\mathcal{G} \mid \lambda) = \pi(\mathcal{G} \mid \lambda)Pr(\omega)$ can be evaluated as follows.

$$L(\mathcal{G} \mid \lambda) = \int\limits_{[0,\infty)^{\ell}} \pi(\mathcal{G} \mid \lambda, \{I_j\})Pr(\omega)d\Pi(\{I_j\}) = \mathop{\mathbb{E}}_{\{I_j\}}\{\pi(\mathcal{G} \mid \lambda, \{I_j\})Pr(\omega)\}.$$

However, recalling that the out-degree of individual j is denoted by δ_j we have

$$\begin{aligned} L(\mathcal{G} \mid \lambda, \{I_j\}) &= \prod_{j=1}^{\ell} \left(1-e^{-\lambda I_j}\right)^{\delta_j} \left(e^{-\lambda I_j}\right)^{(\ell-\delta_j-1)} e^{-\lambda I_j(N-\ell)} \\ &= \prod_{j=0}^{\ell} \left(1-e^{-\lambda I_j}\right)^{\delta_j} e^{-\lambda I_j(N-\delta_j-1)} \end{aligned} \tag{4.6}$$

The formula in 4.6 enables us to evaluate the likelihood when the infectious periods of the individuals and their out-degrees are available. Additionally, when only the graph information is available further progress can be made using the binomial theorem. Specifically we can rewrite $\left(1-e^{-\lambda I_j}\right)^{\delta_j}$ as follows:

$$\left(1-e^{-\lambda I_j}\right)^{\delta_j} = \sum_{k=0}^{\delta_j} \binom{\delta_j}{k} (-1)^{(\delta_j-k)} e^{-\lambda I_j(\delta_j-k)}.$$

Hence,

$$\left(1-e^{-\lambda I_j}\right)^{\delta_j} e^{-\lambda I_j(N-\delta_j-1)} = \sum_{k=0}^{\delta_j} \binom{\delta_j}{k} (-1)^{(\delta_j-k)} e^{-\lambda I_j(N-k-1)}.$$

Since the infectious periods $\{I_j\}$ are mutually independent we have that:

$$\mathbb{E}\left(f(I_k)g(I_j)\right) = \mathbb{E}\left(f(I_k)\right)\mathbb{E}\left(g(I_j)\right), k \neq j,$$

for any functions f, g such that the expectations exist. Thus, we get

$$\mathop{\mathbb{E}}_{\{I_j\}}\{L(\mathcal{G} \mid \lambda, \{I_j\})\} = \prod_{j=0}^{l} \mathbb{E}\left(\sum_{k=0}^{\delta_j} \binom{\delta_j}{k} (-1)^{(\delta_j-k)} e^{-\lambda I_j(N-k-1)}\right).$$

Finally the likelihood can be evaluated from

$$L(\mathcal{G} \mid \lambda) = \prod_{j=0}^{\ell} \left(\sum_{k=0}^{\delta_j} \binom{\delta_j}{k} (-1)^{(\delta_j - k)} \phi(\lambda(n-k)) \right), \tag{4.7}$$

with $\phi(\theta) = \mathbb{E}\left(e^{-\theta I}\right)$ where the expectation is taken with respect to the probability distribution of I. The two formulas given in 4.6 and 4.7 are valid for every appropriate random graph that contains the required information, independently of the method used to construct the graph. Additionally, (4.7) illustrates the value of considering all possible links in the augmented space. The above calculations are straightforward to extend to more complex population structures and the corresponding results for the two level mixing model are presented in the following section.

The Likelihood for the Two-level-mixing Model

For a given realisation of G, say $\mathcal{G}$, the likelihood becomes a conditional density that can be written as the product of three components

$$L = L_1 L_2 L_3,$$

where L_1 can be evaluated from $\mathcal{G}$ as in the homogeneous case and $Pr(\omega) = L_2 L_3$ is the contribution to the likelihood of the non-links between the ℓ ultimately infectives and the remaining $N - \ell$ individuals of the population, L_2 corresponding to global and L_3 to local infections.

L_1 can be evaluated as the product of the appropriate binomial probabilities. We recall that we denote by δ_{Gj} the number of *global* directed links emanating from individual j and by δ_{Lj} the number of *local* contacts of j while he has been infectious. Also the probability of this individual having a local infectious contact with a specified individual in his household is $p_{jL} = 1 - q_{jL}$

where $q_{jL} = \exp(-\lambda_L I_j)$ and the probability of a global contact with a specified individual in the population is $p_{jG} = 1 - q_{jG}$, $q_{jG} = \exp\left(-\frac{\lambda_G I_j}{N}\right)$. Note that for the evaluation of the likelihood we can ignore the pseudotemporal, per generation, information obtained by the random graph and focus on each individual. Let us recall that s_{Lj} is the number of not yet infected vertices in j's group and s_G is the number of global within-graph susceptible vertices. All these individuals will however be infected at the end of the graph construction. Then it is easy to see that

$$L_1 = \prod_{j=1}^{\ell} p_{jG}^{\delta_{Gj}} q_{jG}^{s_G - \delta_{Gj} - 1} p_{jL}^{\delta_{Lj}} q_{jL}^{s_{Lj} - \delta_{Lj} - 1}. \tag{4.8}$$

The second component is the contribution of the ℓ ultimately infected individuals which fail to *globally* infect the $N - \ell$ remaining susceptibles and is given by

$$L_2 = \prod_{j=1}^{\ell} \exp\left(-(\lambda_G/N) I_j (N - \ell)\right) = \exp\left(\lambda_G \left(\frac{\ell}{N} - 1\right) \sum_{j=1}^{\ell} I_j\right). \tag{4.9}$$

The quantity $\sum_{j=1}^{\ell} I_j$ is a realisation of the final severity of the epidemic.

The third component is due to the individuals who avoid infection *locally* from the set of the ℓ infectives and is calculated as

$$\begin{aligned} L_3 &= \prod_{i=1}^{h-1} \prod_{j=i}^{h-1} \prod_{k=n_{ii}+1}^{n_{ii}+n_{ij+1}} \exp\left(-\lambda_L (j - i + 1) I_k\right) \\ &= \exp\left(-\lambda_L \sum_{i=1}^{h-1} \sum_{j=i}^{h-1} (j - i + 1) \sum_{k=n_{ii}+1}^{n_{ii}+n_{ij+1}} I_k\right), \end{aligned} \tag{4.10}$$

where h is the maximum household size and n_{ij} is the number of households of size j of whom i individuals become infected. Combining equations (4.8), (4.9) and (4.10) and after some elementary calculations we get

$$L(\mathcal{G} \mid \lambda_L, \lambda_G, \{I_j\}) = \prod_{j=1}^{\ell} p_{jG}^{\delta_{Gj}} q_{jG}^{N - \delta_{Gj} - 1} p_{jL}^{\delta_{Lj}} q_{jL}^{n_{Lj} - \delta_{Lj} - 1}. \tag{4.11}$$

The formula in (4.11) is the likelihood for the two level mixing model given the vector of the infectious periods and a realisation of the random graph that contains the local and global out-degrees of each individual. This formula could be derived directly simply by taking the product over each individual of the corresponding (binomial) probabilities of a particular realisation, conditional upon his local and global out-degree. However, we do include the likelihood decomposition to the L_1, L_2 and L_3 components since they represent a natural split into the three important aspects of the infection process. As with the generalised stochastic epidemic we can attempt to integrate out the actual realisations of the infectious periods and present the likelihood as a function of the out-degrees of the vertices that correspond to the ultimately infected individuals.

Using the binomial theorem we can rewrite the probability that corresponds to the local infections as follows:

$$\left(1-e^{-\lambda_L I_j}\right)^{\delta_{Lj}} = \sum_{k=0}^{\delta_{Lj}} \binom{\delta_{Lj}}{k} (-1)^{(\delta_{Lj}-k)} e^{-\lambda_L I_j(\delta_{Lj}-k)}.$$

Similarly for the global infections $p_{jG}\left(1-e^{-\lambda_G I_j}\right)^{\delta_{Gj}}$ we obtain:

$$\left(1-e^{-\lambda_G I_j/N}\right)^{\delta_{Gj}} = \sum_{k=0}^{\delta_{Gj}} \binom{\delta_{Gj}}{k} (-1)^{(\delta_{Gj}-k)} e^{-\lambda_G I_j(\delta_{Gj}-k)/N}.$$

Then,

$$\left(1-e^{-\lambda_L I_j}\right)^{\delta_{Lj}} e^{-\lambda_L I_j(n_{Lj}-\delta_{Lj}-1)} = \sum_{k=0}^{\delta_{Lj}} \binom{\delta_{Lj}}{k} (-1)^{(\delta_{Lj}-k)} e^{-\lambda_L I_j(n_{Lj}-k-1)}$$

and

$$\left(1-e^{-\lambda_G I_j/N}\right)^{\delta_{Gj}} e^{-\lambda_G I_j(N-\delta_{Gj}-1)/N} = \sum_{k=0}^{\delta_{Gj}} \binom{\delta_{Gj}}{k} (-1)^{(\delta_{Gj}-k)} e^{-\lambda_G I_j(N-k-1)/N}.$$

Since the infectious periods $\{I_j\}$ are mutually independent we obtain

$$L(\mathcal{G} \mid \lambda_L, \lambda_G) = \mathop{\mathbb{E}}_{\{I_j\}} \{L(\mathcal{G} \mid \lambda_L, \lambda_G, \{I_j\})\} = \prod_{j=0}^{\ell} \mathbb{E}\left(\left\{\sum_{k=0}^{\delta_{Lj}} \mathcal{A}_k\right\}\left\{\sum_{k=0}^{\delta_{Gj}} \mathcal{B}_k\right\}\right),$$

where $\mathcal{A}_k = \binom{\delta_{Lj}}{k}(-1)^{(\delta_{Lj}-k)}e^{-\lambda_L I_j(n_{Lj}-k)}$ and $\mathcal{B}_k = \binom{\delta_{Gj}}{k}(-1)^{(\delta_{Gj}-k)}e^{-\lambda_G I_j(N-k)/N}$.

Using the fact that

$$\left(\sum_{k=0}^{\delta_{Lj}}\mathcal{A}_k\right)\left(\sum_{k=0}^{\delta_{Gj}}\mathcal{B}_k\right) = \sum_{k=0}^{\min\{\delta_{Lj},\delta_{Gj}\}}\sum_{i=0}^{k}\mathcal{C}_{ik} + \sum_{k=\min\{\delta_{Lj},\delta_{Gj}\}+1}^{\delta_{Lj}+\delta_{Gj}}\sum_{i=k-\min\{\delta_{Lj},\delta_{Gj}\}}^{\max\{\delta_{Lj},\delta_{Gj}\}}\mathcal{C}_{ik},$$

where $\mathcal{C}_{ik} = \mathcal{A}_i\mathcal{B}_{k-i}$, we obtain the likelihood:

$$L(\mathcal{G} \mid \lambda_L, \lambda_G) = \prod_{j=0}^{\ell}\left(\sum_{k=0}^{\min\{\delta_{Lj},\delta_{Gj}\}}\sum_{i=0}^{k}\mathcal{D}_{ik} + \sum_{k=\min\{\delta_{Lj},\delta_{Gj}\}+1}^{\delta_{Lj}+\delta_{Gj}}\sum_{i=k-\min\{\delta_{Lj},\delta_{Gj}\}}^{\max\{\delta_{Lj},\delta_{Gj}\}}\mathcal{D}_{ik}\right), \tag{4.12}$$

where $\mathcal{D}_{ik} = \binom{\delta_{Lj}}{i}\binom{\delta_{Gj}}{k-i}(-1)^{(\delta_{Gj}+\delta_{Lj}-k)}\phi\left(\lambda_L(n_{Lj}-i) + \frac{\lambda_G(N-k+i)}{N}\right)$ and $\phi(\theta) = \mathbb{E}\left(e^{-\theta I}\right)$ is the moment generating function for I. Like in the homogeneous case, the two formulas given in 4.11 and 4.12 are valid for every appropriate random graph that contains the required local and global contact information, independently of the method the graph is constructed. Once a graph of this kind has been obtained we can proceed with an appropriate MCMC algorithm in order to explore the posterior density of interest. In the following section we describe the algorithm in detail.

4.4 Markov chain Monte Carlo algorithm

The posterior density of interest can be obtained using Bayes' Theorem since it is proportional to the product of the likelihood and the prior. We assume independent prior distributions for each parameter. Let us denote by $\tilde{I}$ the ℓ-dimensional vector with the infectious periods of the ℓ ultimately infected individuals. Then the posterior is given by

$$\begin{aligned}\pi(\lambda_L, \lambda_G, \mathcal{G}, \kappa, \tilde{I} \mid \{n_{ij}\}) \quad &\propto \quad \pi(\{n_{ij}\} \mid \lambda_L, \lambda_G, \mathcal{G}, \kappa, \tilde{I})\pi(\mathcal{G} \mid \lambda_L, \lambda_G, \kappa, \tilde{I}) \\ &\quad \pi(\lambda_L)\pi(\lambda_G)\pi(\kappa)\pi(\tilde{I}),\end{aligned} \tag{4.13}$$

where

$$\pi(\{n_{ij}\} \mid \lambda_L, \lambda_G, \mathcal{G}, \kappa, \tilde{I}) = \mathbb{1}_{\{\mathcal{E}\}} Pr(\omega),$$

with ω being the event that there are no links from the ℓ ultimately infectived individuals to the remaining $N - \ell$ individuals of the population and $\mathcal{E}$ denoting the event that the digraph $\mathcal{G}$ agrees with the data $\{n_{ij}\}$. Hence, $\pi(\mathcal{G} \mid \lambda_L, \lambda_G, \kappa, \tilde{I})Pr(\omega)$ essentially represents the likelihood, since we sample the random graph $\mathcal{G}$ conditional on the data. In the homogeneous case (4.13) reduces to

$$\pi(\lambda, \mathcal{G}, \tilde{I} \mid \ell) \propto \pi(\ell \mid \lambda, \mathcal{G}, \tilde{I})\pi(\mathcal{G} \mid \lambda, \tilde{I})\pi(\lambda)\pi(\tilde{I}). \tag{4.14}$$

Here $\pi(\ell \mid \lambda, \mathcal{G}, \tilde{I}) = \mathbb{1}_{\{\mathcal{E}\}} Pr(\omega)$ and the likelihood can be evaluated by $L(\mathcal{G} \mid \lambda, \tilde{I}) = \pi(\mathcal{G} \mid \lambda, \tilde{I})Pr(\omega)$. We now proceed to the corresponding MCMC algorithms that provide us with approximate samples from the posterior densities of interest.

4.4.1 The Independence Sampler

Generalised Stochastic Epidemic

We use a single component Metropolis-Hastings algorithm in which the parameters λ, $\mathcal{G}$ and α are updated in one block as follows. We first sample the proposed values for (λ, α), say (λ^*, α^*) from a Gaussian random walk proposal. If λ^* (or α^*) is negative then the sample is rejected since it has likelihood 0. Otherwise, we proceed to sample the proposed graph $\mathcal{G}^*$ according to the method described in section 4.3.2 where the probability mass function $q(\mathcal{G}^* \mid \lambda^*, \alpha^*)$ was derived. The proposed new parameter vector is then accepted with probability

$$\frac{L(\mathcal{G}^* \mid \lambda^*)\pi(\lambda^*)q(\mathcal{G} \mid \lambda, \alpha)}{L(\mathcal{G} \mid \lambda)\pi(\lambda)q(\mathcal{G}^* \mid \lambda^*, \alpha^*)} \wedge 1.$$

Two level mixing model

In this algorithm we update the parameters in two blocks. One block consists of the two infection rates λ_L and λ_G, while at the second block of each sweep of the algorithm we update the random graph $\mathcal{G}$, the initial infective κ and the vector with the infectious periods $\tilde{I}$. We use a discrete uniform over $\{1, \ldots, \ell\}$ prior for κ, we assign the distribution of the infectious periods $\tilde{I}$ and we denote the prior density|mass function with $\pi()$.

Sampling $\mathcal{G}$, κ and $\tilde{I}$: We propose new values for κ and $\tilde{I}$ sampling from their prior distribution. Let us denote with κ^* the proposed initial infective and with $\tilde{I}^*$ the proposed infectious periods. Based on κ^* and $\tilde{I}^*$ we sample the proposed graph $\mathcal{G}^*$. The new samples are then accepted with probability

$$\frac{L(\mathcal{G}^* \mid \lambda_L, \lambda_G, \kappa^*, \tilde{I}^*)\pi(\lambda_L)\pi(\lambda_G)\pi(\kappa^*)\pi(\tilde{I}^*)q(\mathcal{G})q(\kappa)q(\tilde{I})}{L(\mathcal{G} \mid \lambda_L, \lambda_G, \kappa, \tilde{I})\pi(\lambda_L)\pi(\lambda_G)\pi(\kappa)\pi(\tilde{I})q(\mathcal{G}^*)q(\kappa^*)q(\tilde{I}^*)} \wedge 1.$$

Note that $\pi(\kappa)$, $\pi(\kappa^*)$, $\pi(\tilde{I})$ and $\pi(\tilde{I}^*)$ vanish from the acceptance probability since they only enter as the ratio $\frac{\pi(\kappa^*)}{\pi(\kappa)}$ for κ and $\frac{\pi(\tilde{I}^*)}{\pi(\tilde{I})}$ for the infectious periods. The acceptance probability is then:

$$\frac{L(\mathcal{G}^* \mid \lambda_L, \lambda_G, \kappa^*, \tilde{I}^*)q(\mathcal{G} \mid \lambda_L, \lambda_G, \kappa, \tilde{I})}{L(\mathcal{G} \mid \lambda_L, \lambda_G, \kappa, \tilde{I})q(\mathcal{G}^* \mid \lambda_L, \lambda_G, \kappa^*, \tilde{I}^*)} \wedge 1. \tag{4.15}$$

Sampling λ_L and λ_G: We use a bivariate normal proposal based around the current value (λ_L, λ_G). If one of the proposed values λ_L^* or λ_G^* is negative the sample is rejected with probability one. Otherwise, we calculate the new likelihood $\pi(\mathcal{G} \mid \lambda_L^*, \lambda_G^*, \kappa, \tilde{I})$ based on the current graph $\mathcal{G}$ and the current κ and $\tilde{I}$. The proposed sample $(\lambda_G^*, \lambda_G^*)$ is then accepted with probability

$$\frac{L(\mathcal{G} \mid \lambda_L^*, \lambda_G^*, \kappa, \tilde{I})\pi(\lambda_L^*)\pi(\lambda_G^*)}{L(\mathcal{G} \mid \lambda_L, \lambda_G, \kappa, \tilde{I})\pi(\lambda_L)\pi(\lambda_G)} \wedge 1.$$

Note that $q(\lambda_L, \lambda_G) = q(\lambda_L^*, \lambda_G^*)$ as the proposal is symmetric and hence it does not appear in the ratio either. Note also that the covariance matrix of the

bivariate normal proposal for λ_L and λ_G is the only "tuning" parameter for the algorithm.

4.4.2 The Birth Death Sampler

The approach that appears to perform most efficiently in this algorithm is to update the model parameters in three blocks as follows.

Updating the Infection rates

The updates of the infection rates were all based on Gaussian random walk proposals constrained on the positive real line like in chapters 2 and 3. Hence, a negative proposed value λ^* is rejected with probability 1. A positive λ^* is accepted with probability

$$\frac{L(\mathcal{G} \mid \lambda^*, \tilde{I})\pi(\lambda^*)}{L(\mathcal{G} \mid \lambda, \tilde{I})\pi(\lambda^*)} \wedge 1, \tag{4.16}$$

where $\tilde{I}$ denotes the vector of the infectious periods. We update the infection rates in the two level mixing model similarly, and the acceptance probability is again the likelihood ratio multiplied by the ratio of the priors over the parameters to be updated:

$$\frac{L(\mathcal{G} \mid \lambda_L^*, \lambda_G^*, \tilde{I})\pi(\lambda_L^*)\pi(\lambda_G^*)}{L(\mathcal{G} \mid \lambda_L, \lambda_G, \tilde{I})\pi(\lambda_L)\pi(\lambda_G)} \wedge 1. \tag{4.17}$$

We shall now describe the updates of the infectious periods when these are introduced as extra model parameters.

Updating the Infectious Periods

Since we have final size data we need to make specific distributional assumptions about the infectious periods. In a Bayesian framework where the infectious periods appear as model parameters this is equivalent to strong prior

assumptions. In this particular case, sampling from the priors is computationally convenient. This is not necessary but it appears to be a plausible choice when only final outcome data are available. Then the probability of accepting a proposed infectious period vector $\tilde{I}^*$ reduces to the likelihood ratio:

$$\frac{L(\mathcal{G} \mid \lambda_L, \lambda_G, \tilde{I}^*)}{L(\mathcal{G} \mid \lambda_L, \lambda_G, \tilde{I})} \wedge 1. \tag{4.18}$$

Note that for complex stochastic systems, like the epidemic with two levels of mixing, the likelihood is complicated and proposing the update of the whole vector can be too ambitious in that it frequently has a very small acceptance probability. Updating the individual infectious periods might be more efficient in this case and this approach corresponds to a single-site update in a Gibbs sampler. Alternatively the infectious periods could be updated in small blocks, e.g. ten at a time.

Updating the Random Graph

Here we use the acceptance probabilities that were described in the construction of the random graph to obtain the Metropolis-Hastings acceptance probabilities. We shall describe each proposed move separately.

Generalised Stochastic Epidemic

Adding an edge It is straightforward to obtain the probability of moving from a graph $\mathcal{G}$ with $\sum_{j=1}^{\ell} \delta_j$ directed links to a graph $\mathcal{G}'$ with $\sum_{j=1}^{\ell} \delta_j + 1$ edges. The probability of accepting this move is $\alpha \wedge 1$ where:

$$\alpha = \frac{L(\mathcal{G}' \mid \lambda, \tilde{I}) q(\mathcal{G} \mid \mathcal{G}')}{L(\mathcal{G} \mid \lambda, \tilde{I}) q(\mathcal{G}' \mid \mathcal{G})} = \frac{(1 - \exp(-\lambda I_j)) \left(\ell(\ell - 1) - \sum_{j=1}^{\ell} \delta_j \right)}{\exp(-\lambda I_j) \left(\sum_{j=1}^{\ell} \delta_j + 1 \right)}. \tag{4.19}$$

The likelihood ratio $\frac{1-\exp(-\lambda I_j)}{\exp(-\lambda I_j)}$ can then be evaluated by $\exp(\lambda I_j) - 1$ when a realisation of the infectious period is available. When the infectious periods are not available we have to evaluate the likelihood ratio using (4.7).

Deleting an edge The calculation for moving from a graph $\mathcal{G}$ with $\sum_{j=1}^{\ell} \delta_j$ directed links to a graph $\mathcal{G}'$ with $\sum_{j=1}^{\ell} \delta_j - 1$ edges is equivalent to the deletion case and the proposed graph is accepted with probability:

$$\frac{\sum_{j=1}^{\ell} \delta_j}{\left(\ell(\ell-1) - \sum_{j=1}^{\ell} \delta_j + 1\right)(\exp(\lambda I_j) - 1)} \wedge 1. \tag{4.20}$$

Again, in the case that the infectious periods are not available we have to evaluate the likelihood ratio using (4.7). We shall now describe the graph updates in the two level mixing model.

Two level mixing model

Adding a local edge In this case we wish to obtain the probability of moving from a graph $\mathcal{G}$ with $\sum_{j=1}^{\ell} \delta_{Lj}$ directed links to a graph $\mathcal{G}'$ with $\sum_{j=1}^{\ell} \delta_{Lj} + 1$ edges. Using the same arguments as in the generalised stochastic epidemic it is straightforward to see that the probability of accepting this move is $\alpha \wedge 1$ where:

$$\alpha = \frac{\pi(\mathcal{G}' \mid \lambda_L, \lambda_G, \tilde{I}) q(\mathcal{G} \mid \mathcal{G}')}{\pi(\mathcal{G} \mid \lambda_L, \lambda_G, \tilde{I}) q(\mathcal{G}' \mid \mathcal{G})} = \frac{(1 - \exp(-\lambda_L I_j)) \left(\sum_{j:n_{Lj} \geq 2} n_{Lj}(n_{Lj} - 1) - \sum_{j=1}^{\ell} \delta_{Lj}\right)}{(\exp(-\lambda_L I_j)) \left(\left(\sum_{j:n_{Lj} \geq 2} n_{Lj}(n_{Lj} - 1) - \sum_{j=1}^{\ell} \delta_{Lj}\right) + 1\right)}. \tag{4.21}$$

The evaluation of the likelihood ratio $\frac{1-\exp(-\lambda_L I_j)}{\exp(-\lambda_L I_j)}$ proceeds as above.

Adding a global edge Here we wish to obtain the probability of moving from a graph $\mathcal{G}$ with $\sum_{j=1}^{\ell} \delta_{Gj}$ directed links to a graph $\mathcal{G}'$ with $\sum_{j=1}^{\ell} \delta_{Gj} + 1$

edges. Similarly to the local case the probability of accepting this move is $\alpha \wedge 1$ where:

$$\alpha = \frac{(1-\exp(-\lambda_G I_j/N))\left(\ell(\ell-1) - \sum_{j=1}^{\ell}\delta_{Gj}\right)}{(\exp(-\lambda_G I_j/N))\left(\left(\sum_{j=1}^{\ell}\delta_{Gj}\right)+1\right)}. \tag{4.22}$$

Deleting an edge

Deleting a local edge In this case we wish to obtain the probability of moving from a graph $\mathcal{G}$ with $\sum_{j=1}^{\ell}\delta_{Lj}$ directed links to a graph $\mathcal{G}'$ with $\sum_{j=1}^{\ell}\delta_{Lj}-1$ edges. Then that the probability of accepting this move is $\alpha \wedge 1$ where:

$$\alpha = \frac{\left((\exp(-\lambda_L I_j))\sum_{j:n_{Lj}\geq 2} n_{Lj}(n_{Lj}-1) - \sum_{j=1}^{\ell}\delta_{Lj}\right)}{(1-\exp(-\lambda_L I_j))\left(\left(\sum_{j:n_{Lj}\geq 2} n_{Lj}(n_{Lj}-1) - \sum_{j=1}^{\ell}\delta_{Lj}\right)+1\right)}. \tag{4.23}$$

Deleting a global edge Here we wish to obtain the probability of moving from a graph $\mathcal{G}$ with $\sum_{j=1}^{\ell}\delta_{Gj}$ directed links to a graph $\mathcal{G}'$ with $\sum_{j=1}^{\ell}\delta_{Gj}-1$ edges. The probability of accepting this move is $\alpha \wedge 1$ where:

$$\alpha = \frac{\left((\exp(-\lambda_G I_j/N))\sum_{j=1}^{\ell}\delta_{Gj}\right)}{(1-\exp(-\lambda_G I_j/N))\left(\left(\ell(\ell-1) - \sum_{j=1}^{\ell}\delta_{Gj}\right)+1\right)}. \tag{4.24}$$

Again, when the infectious periods are not available we have to evaluate the likelihood ratio using (4.12).

4.5 Application to Data

The methodology described thus far was applied to a number of different datasets in order to illustrate the methods and to evaluate the accuracy of the algorithms where possible. We shall present the results from the birth-death sampler. The mixing of the independence sampler appears to vary and

it can be particularly slow. In a number of runs it has given very similar results to the birth-death algorithm, especially for the estimation of the infection rates. However, looking at the infection rates only can be misleading. More specifically, the convergence of the graph chain can be problematic and indeed extremely slow. The graph appears to remain in a reasonable (and valid) configuration for an exceedingly large number of iterations, see for example figure (4.1) for the posterior of the total number of links when the algorithm was applied to the smallpox dataset used in the second chapter. Since 30 out of 120 individuals are getting ultimately infected we have a minimum of 29 links that can result in valid configurations. This particular algorithm had a sampling gap of 1000 and despite the extended thinning it is clear that the output is unsatisfactory, the main reason being that the graph does not move sufficiently around its posterior space, despite being in the "high posterior probability region". Hence, it is easy to obtain good results for the infection rates under the false impression that the Markov chain mixes well and integrated out the posterior space. The convergence problems seem to occur because with both the binomial and the truncated Poisson proposals the actual proposal probabilities appear to be badly calibrated with respect to the high posterior density region. It is well known that a badly calibrated independence sampler can perform very poorly and indeed not being geometrically ergodic, e.g. Roberts (1996) p.55.

We illustrate in this section the use of the birth-death algorithm by applying it to datasets from influenza epidemics as well as an artificial dataset, and we describe the novel posterior information that can be obtained with our methods. In the following section we shall assess the precision of our approach in two distinct ways.

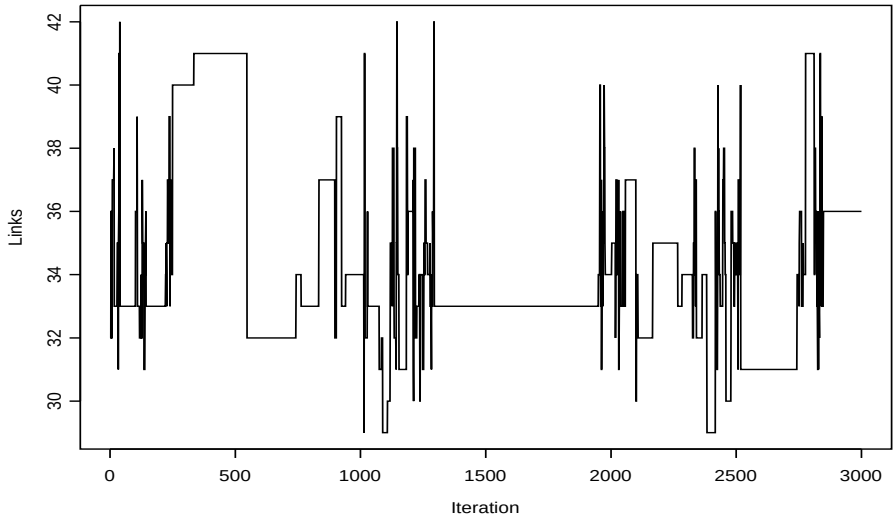

Figure 4.1: The output for the total number of links for the smallpox dataset.

4.5.1 Influenza Outbreak Data

The Data

We apply the above methodology to the influenza outbreak data described in Table 3.1. This dataset consists of the observed distribution of influenza A(H3N2) infections in the 1977-1978 and 1980-1981 combined epidemics in Tecumseh, Michigan as described in chapter 3, see Addy *et al.* (1991) and the references therein for details. The actual household sizes go up to 7 but we use the dataset in Table 3.1 for comparison with the analysis of chapter 3 presented also in Demiris and O'Neill (2003). We shall analyse the two separate datasets as well as the complete (up to size 7) dataset in the following subsection.

	Parameter					
	λ_L	λ_G	$\hat{\delta}_L$	$\hat{\delta}_G$	$\hat{\delta}$	R_*
Mean	0.045	0.199	0.352	0.806	1.156	1.156
Median	0.045	0.198	0.352	0.804	1.156	1.150
S. dev.	0.007	0.018	0.033	0.029	0.028	0.118
95% C. I.	(0.03,0.06)	(0.16,0.24)	(0.28,0.41)	(0.75,0.87)	(1.10,1.22)	(0.94,1.40)

Table 4.1: Posterior parameter summaries for the Influenza dataset with households sizes truncated to 5 and with a Gamma distributed infectious period.

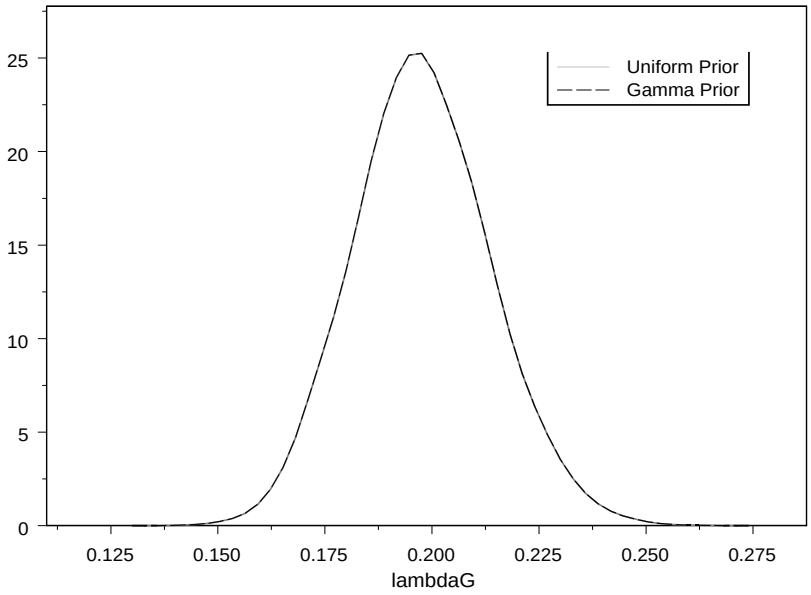

Figure 4.2: The posterior density of λ_G for the two different priors.

Implementation

In accordance with all the algorithms in this thesis, this MCMC algorithm was implemented using Fortran 90 in a mainframe computer. Each iteration (or cycle) of the algorithm consists of sampling each one of the parameters once, no repeated sampling was utilised in the algorithms presented here. The burn-in we used for the results presented here was 2×10^5 for datasets with final size up to 300 and 5×10^5 when the number of ultimately infected individuals in the dataset under study was larger than 300. All the results presented here are from a sample size of 10^4 with a sampling gap of 100. The actual run time was approximately one hour for a dataset with a final size of 250. We used the same prior as before, namely a Gamma with mean 1 and variance 10000. All the algorithms were also tested with three more Gamma priors with identical variance and means equal to 10, 100 and 1000 respectively, as well as a Uniform(0.001,10000) distribution. Again, this prior restricts the posterior state-space but for realistic datasets this is never a problem. The results were virtually identical to the numbers presented here and for illustration we show the output of the posterior distribution of λ_G for the Gamma with mean 1 and the Uniform prior in figure (4.2). To summarise, the output appears to be relatively unaffected by the choice of the prior distribution, at least when we used a large prior variance.

The convergence of the Markov chains was tested informally with plots of the "trace" of the chain and the burn-in we used seems to be sufficient. Additionally, we plotted the autocorrelation functions from the all the "thinned" chains and the autocorrelation reduces drastically for lags larger than 3 or 4. In fact, we found that the ACF plots are more informative for the "exact" behaviour of the algorithm, although a combination of both plots was always used. An example of this kind is presented in figure 4.3 where we plotted

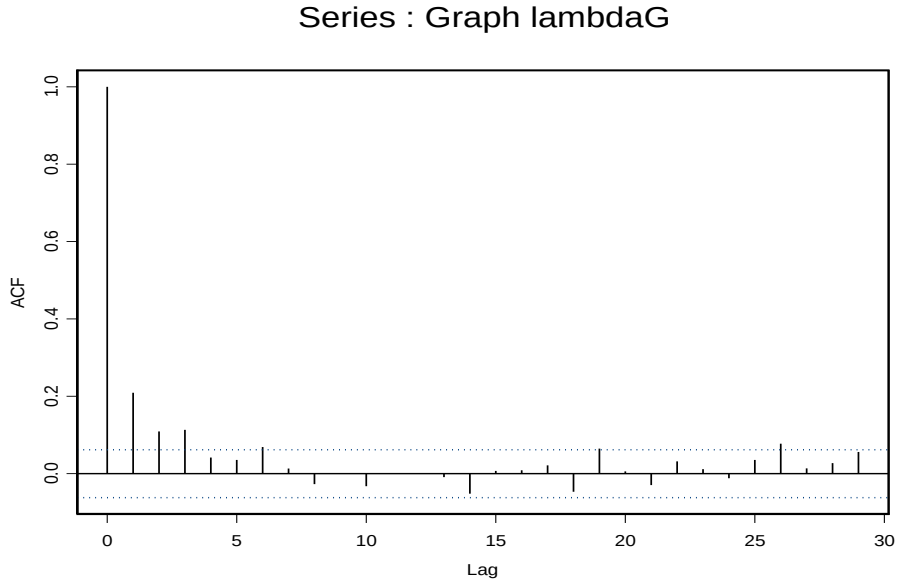

Figure 4.3: The posterior autocorrelation function of λ_G.

the autocorrelation function of λ_G. Additionally, we provide trace plots from the posterior densities of λ_L, λ_G and the total number of local and global out-degrees in figures 4.4, 4.5, 4.6 and 4.7 respectively. Thus, the Markov chains for the birth-death algorithms appear to mix reasonably well. In general convergence diagnostics are an unsolved problem, unless one can find a way to perform perfect simulation (Propp and Wilson (1996)). However, it is relatively easy to (informally) check the convergence of these algorithms since we only have a small number of parameters. Finally, we used the same distributional assumptions, with respect to the infectious period, with Addy *et al.* (1991) and Demiris and O'Neill (2004) namely, a Gamma distribution with mean 4.1 that is the sum of two independent exponential random variables, with mean 2.05.

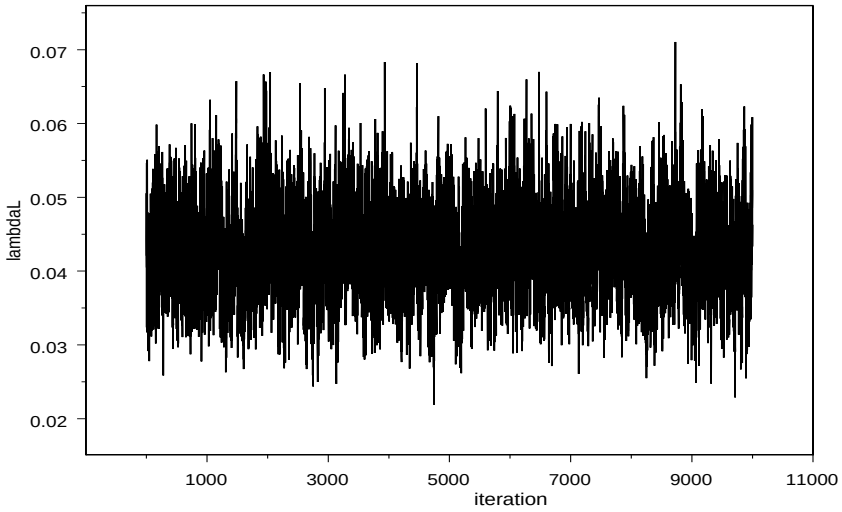

Figure 4.4: Trace of the posterior density of λ_L from the Random Graph algorithm.

Results

The results are summarised in Table 4.1. It is appropriate to compare the outcome with the results in Table 3.2 since we also assume here that the data we observe account for the whole population or, in the terminology of chapter 3, $\alpha = 1$. We shall explore the case that the data we observe are a random sample of the population in the sequel. The key observation here is that the two algorithms give very similar results with respect to point estimation but the approximate approach that utilises the final severity underestimates the variance of λ_G. A possible explanation lies in the fact that the approximate method rejects the proposed (λ_L,λ_G) samples that result in $R_* > 1$. It is actually trivial to obtain from the posterior distribution of R_* that $Pr(R_* < 1) \approx 0.08$. However, it is reassuring that both methods agree with respect to

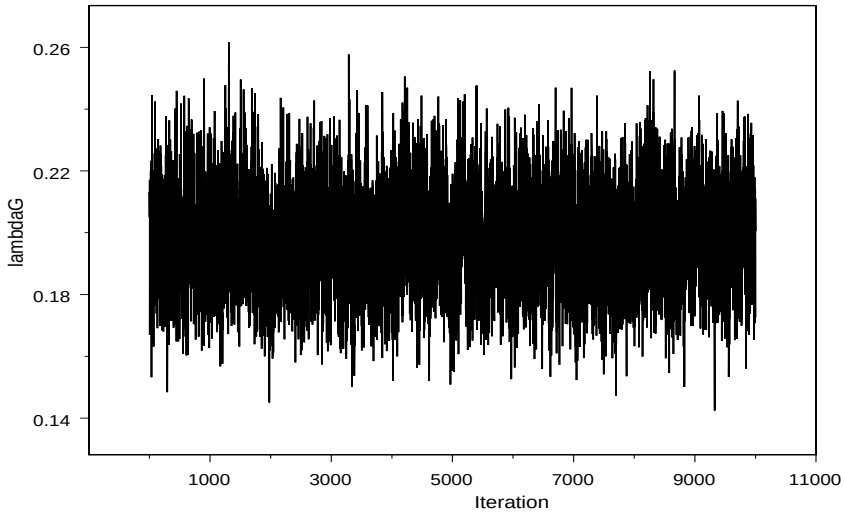

Figure 4.5: Trace of the posterior density of λ_G from the Random Graph algorithm.

the location estimates particularly because the variance underestimation is not large.

The comparison of the two approaches can probably best be summarised in the figures 4.8 and 4.9. In particular, it is clear from figure 4.9 that the approach based on approximating the likelihood using the final severity performs reasonably well with respect to point estimation but it is not as satisfactory in the estimation of dispersion measures for R_*. Note that the output from the severity algorithm in figure (4.9) appears to give non-zero posterior support below unity but this is an S-Plus artifact since the actual minimum of the posterior sample of R_* was slightly above 1. Note also that in this particular case the point estimate of R_* is affected be the extent to which the $R_* > 1$ assumption holds but in general the two approaches result in similar outcomes

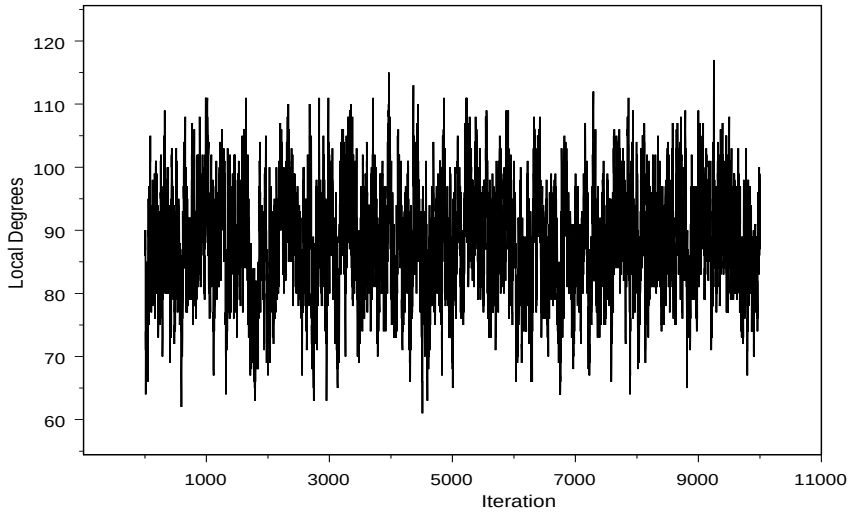

Figure 4.6: Trace of the posterior density of the total number of local out-degrees.

with respect to the location estimates. It is also worth mentioning that the underestimation of the variability is "transferred" to λ_G only since, as can be seen from figure 4.8, the two posterior densities are very close with respect to λ_L. Also the plot of the posterior density of λ_G is identical to the figure 4.9 but located around 0.2 instead of 1.15. The same findings were maintained for a number of different datasets, including the "perfect" data that will be presented later in the evaluation of our method.

Another interesting remark is that the larger posterior variance of λ_G (compared to the final severity approach) reduces the posterior correlation of λ_L and λ_G. This appears to be natural since as the variability of λ_G increases, so does the number of (λ_L,λ_G) combinations that can result in a particular dataset. For the Tecumseh Influenza data of households up to size 5, the posterior cor-

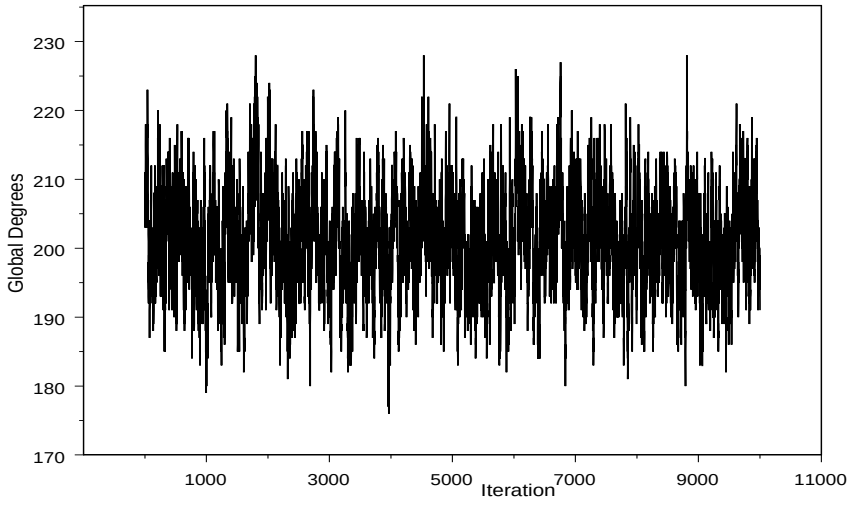

Figure 4.7: Trace of the posterior density of the total number of global out-degrees.

relation was $\rho(\lambda_L, \lambda_G) = -0.0124$. The joint distribution of λ_L,λ_G can be seen from the scatterplot in figure 4.10.

A second interesting outcome is the "decomposition" of the local and global amount of infection as obtained from the mean local and global out-degrees. These are obtained by dividing each sample of the total number of local and global links by the final size i.e., $\hat{\delta}_L = \left(\sum_{j=1}^{\ell} \delta_L(j)\right)/\ell$ and $\hat{\delta}_G = \left(\sum_{j=1}^{\ell} \delta_G(j)\right)/\ell$. The posterior distribution of the mean local and global degree for the Tecumseh influenza data is shown in figure 4.11.

We can obtain the total mean number of links emanating from each ultimately infected individual, by simply adding the separate levels: $\hat{\delta} = \hat{\delta}_L + \hat{\delta}_G$.

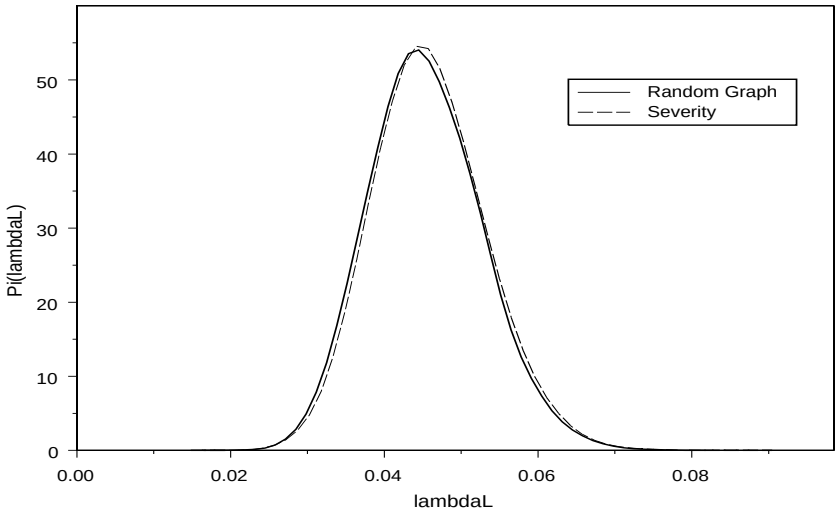

Figure 4.8: The posterior density of λ_L from the Random Graph and the Severity algorithms.

	Susceptibles per household						
No. infected	1	2	3	4	5	6	7
0	66	87	25	22	4	0	0
1	13	14	15	9	4	0	0
2		4	4	9	2	1	0
3			4	3	1	1	1
4				1	1	0	0
5					0	0	0
6						0	0
7							0
Total	79	105	48	44	12		1

Table 4.2: The 1977-1978 Tecumseh influenza data

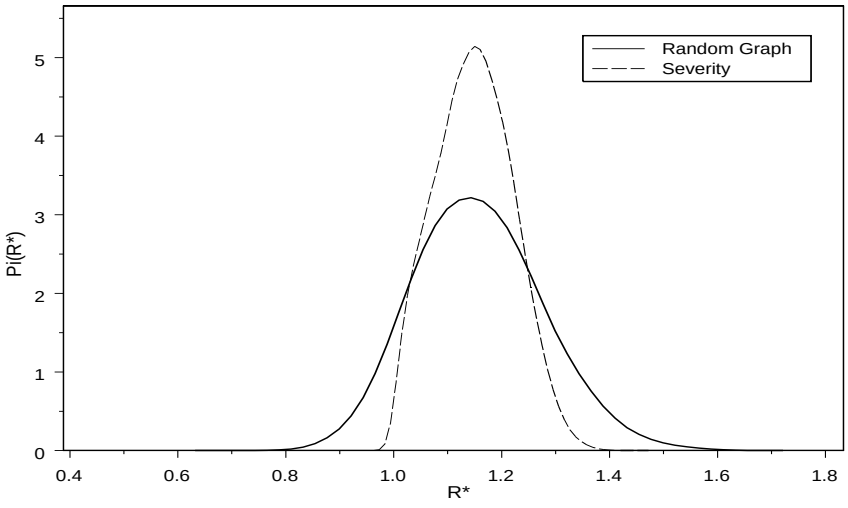

Figure 4.9: The posterior density of R_* from the Random Graph and the Severity algorithms.

4.5.2 Separate and Combined Influenza Data

In this section we shall analyse the individual influenza data. These datasets were kindly provided by Owen Lyne. Addy *et al.* (1991) truncated the combined (both the 1978 and 1981 datasets added together) dataset up to households of size 5 due to numerical problems with larger household sizes. Our approach can cope easily with large household sizes. The dataset from the 1977-1978 outbreak is presented in Table 4.2 while the 1980-1981 outbreak is summarised in Table 4.3.

The results presented are all with the same infectious period, with $E(I) = 4.1$. The posterior summaries for the 1978, 1981 and the combined 1978 and 1981 datasets are presented in Tables 4.4, 4.5 and 4.6 respectively. Note that in the analysis of the combined data set we allow global mixing between individu-

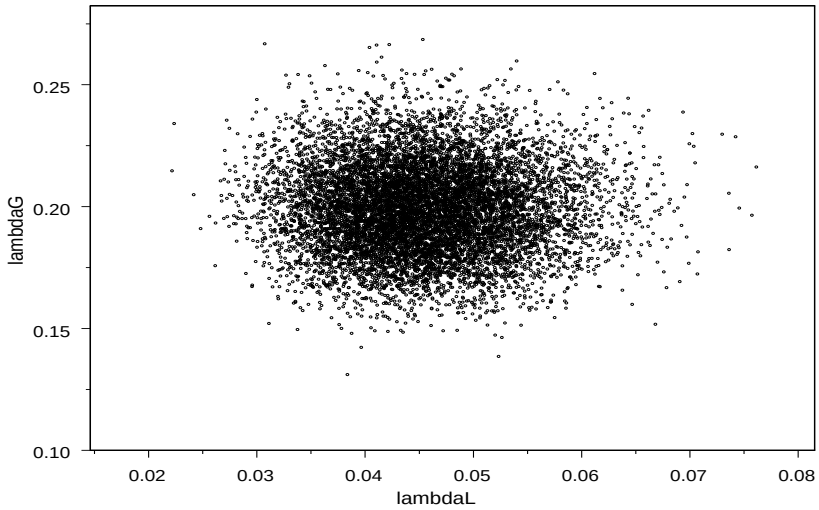

Figure 4.10: Scatterplot of λ_L and λ_G for the Tecumseh data when the infectious period follows a Gamma distribution.

	Susceptibles per household						
No. infected	1	2	3	4	5	6	7
0	44	62	47	38	9	3	2
1	10	13	8	11	5	3	0
2		9	2	7	3	0	0
3			3	5	1	0	0
4				1	0	0	0
5					1	0	0
6						0	0
7							0
Total	79	105	48	44	12		1

Table 4.3: The 1980-1981 Tecumseh influenza data

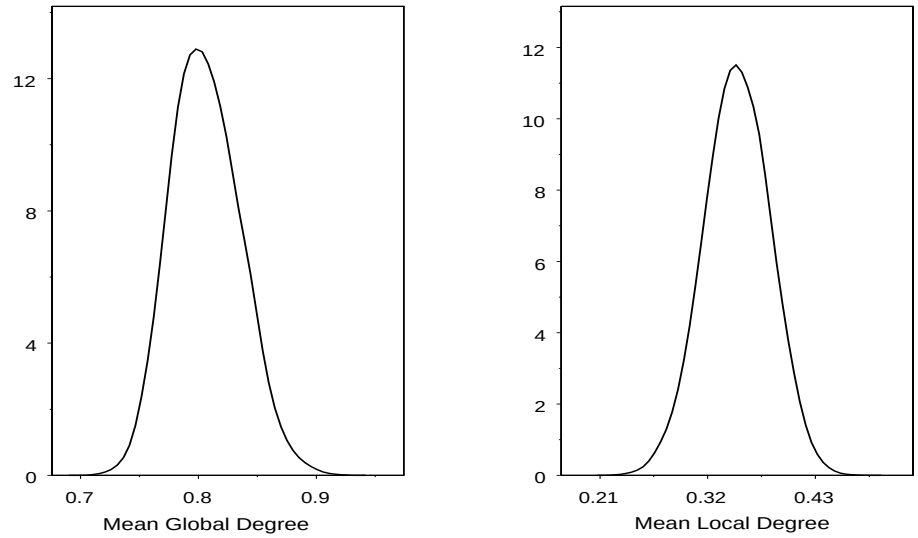

Figure 4.11: The Posterior distribution of the mean local and global degree for the Tecumseh data.

	Parameter					
	λ_L	λ_G	$\hat{\delta}_L$	$\hat{\delta}_G$	$\hat{\delta}$	R_*
Mean	0.039	0.197	0.344	0.797	1.140	1.088
Median	0.039	0.196	0.346	0.797	1.135	1.085
S. dev.	0.007	0.020	0.043	0.038	0.036	0.121
95% C. I.	(0.03,0.05)	(0.14,0.24)	(0.25,0.43)	(0.73,0.88)	(1.07,1.22)	(0.86,1.34)

Table 4.4: Posterior parameter summaries for the full 1977-1978 Influenza dataset.

	Parameter					
	λ_L	λ_G	$\hat{\delta}_L$	$\hat{\delta}_G$	$\hat{\delta}$	R_*
Mean	0.047	0.185	0.399	0.748	1.167	1.161
Median	0.046	0.184	0.398	0.742	1.168	1.156
S. dev.	0.008	0.020	0.042	0.033	0.037	0.138
95% C. I.	(0.03,0.06)	(0.14,0.23)	(0.32,0.49)	(0.68,0.82)	(1.09,1.24)	(0.90,1.45)

Table 4.5: Posterior parameter summaries for the full 1980-1981 Influenza dataset.

	Parameter					
	λ_L	λ_G	$\hat{\delta}_L$	$\hat{\delta}_G$	$\hat{\delta}$	R_*
Mean	0.043	0.189	0.370	0.772	1.143	1.114
Median	0.042	0.189	0.371	0.770	1.142	1.111
S. dev.	0.006	0.015	0.030	0.025	0.025	0.092
95% C. I.	(0.03,0.06)	(0.16,0.22)	(0.31,0.43)	(0.72,0.83)	(1.09,1.20)	(0.94,1.31)

Table 4.6: Posterior parameter summaries for the combined 1977-1978 and 1980-1981 Influenza datasets with household sizes up to 7.

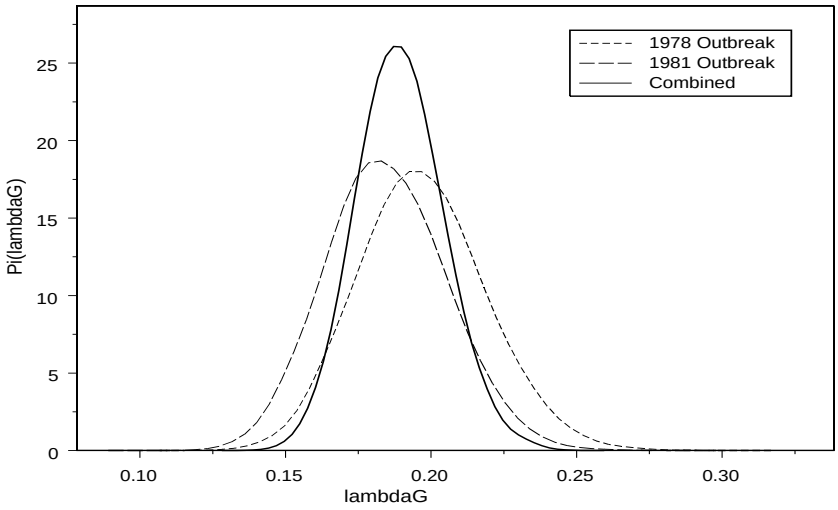

Figure 4.12: Posterior Density of λ_G for the two separate and the combined Tecumseh outbreaks.

als in different epidemics. This is only for illustrative purposes and answers the question posed by Addy *et al.* (1991). A more careful analysis would not allow global links between individuals in different epidemics. Additional information with respect to the infection spread and mixing patterns in the individual data sets can also be incorporated into the algorithm.

The two key observations can be summarised with respect to the posterior location and the posterior dispersion of the parameters. In particular, the 1978 epidemic resulted in relatively (to the 1981 outbreak) large global infection rate and smaller local rate. However, the threshold parameter R_* and the mean out-degree $\hat{\delta}$ are relatively close as we would expect from the fact that the actual proportions infected are close, despite the different pattern of the disease spread. Hence, it appears reasonable to consider the combination of the

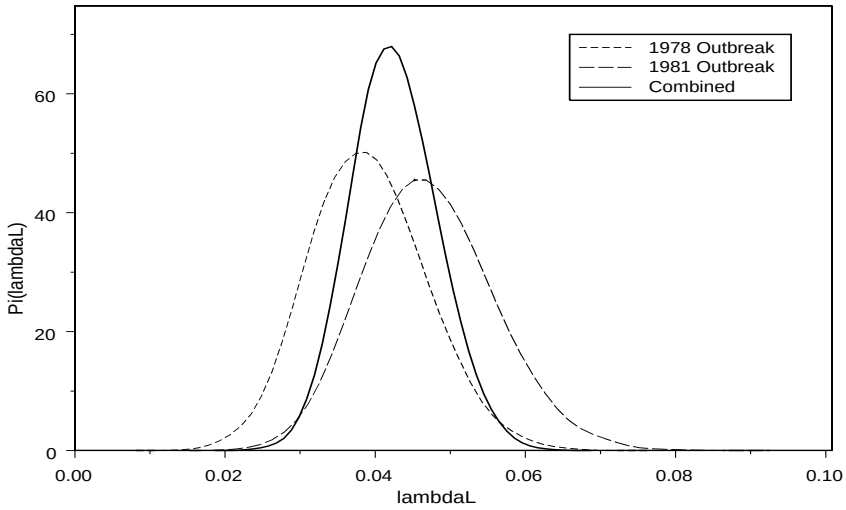

Figure 4.13: Posterior Density of λ_L for the two separate and the combined Tecumseh outbreaks.

two datasets which intuitively provides an average as can be seen from figures 4.12, 4.13 and 4.14 for λ_G, λ_L and R_* respectively. The contributions of the two datasets to the results of the combined dataset are approximately equal since they have roughly the same size.

The second observation is related to the variability of the posterior distribution. It is clear from figures 4.12 and 4.13 as well as the Tables 4.4, 4.5 and 4.6 that the combined dataset, which is approximately double the size of the individual sets, results in posteriors that have smaller variance. This is of course as expected since we would expect that more data provide us with more information and thus with greater posterior precision. However, this observation is related to a practical problem of great interest in epidemics. In particular, it is often difficult (and expensive) to create studies of disease spread over a popu-

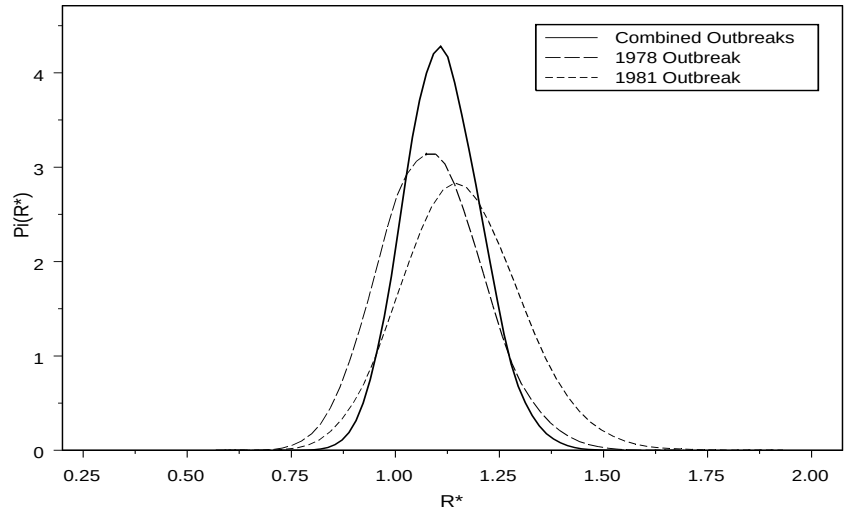

Figure 4.14: Posterior Density of R_* for the two separate and the combined Tecumseh outbreaks.

	Susceptibles per household			
No. infected	1	2	3	4
0	9	9	7	5
1	2	2	2	2
2		1	1	1
3			1	1
4				0
Total	11	12	11	9

Table 4.7: The Artificial dataset

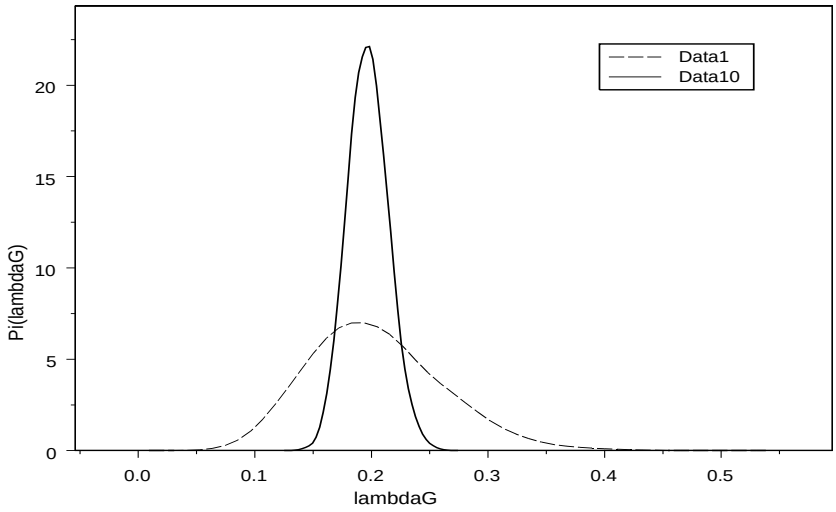

Figure 4.15: Posterior Density of λ_G for the two artificial datasets. We denote with Data10 the dataset where all the values are multiplied by 10.

lation. In contrast, it may be feasible to study the epidemic propagation over a random sample, say $\alpha \times 100\%$, of the population. This issue was addressed in chapter 3 under the assumption that the whole population behaves (with respect to global infections) as our sample. Here we shall present an empirical approach to the solution of a related problem. The method we describe based on the random graph representation becomes slower as the final size gets large. A practically useful alternative in the case of very large datasets consists of analysing a fraction of the actual data and we shall explore the consequences of this approach in the following subsection.

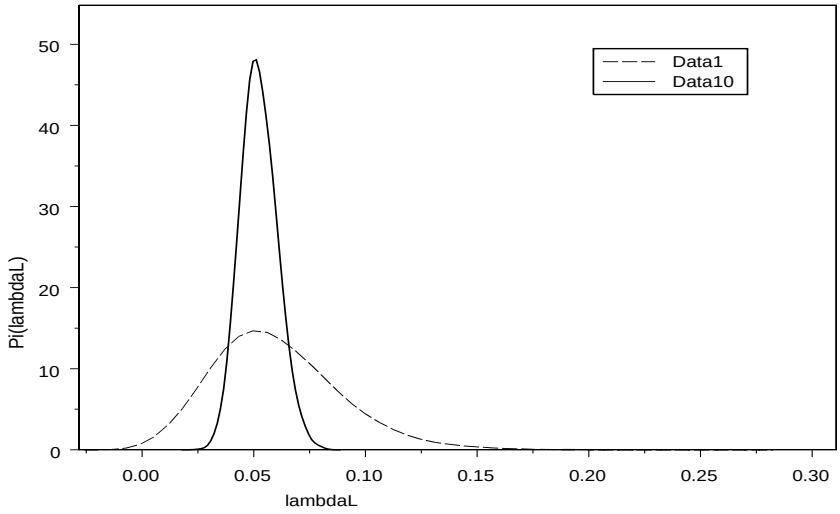

Figure 4.16: Posterior Density of λ_L for the two artificial datasets. We denote with Data10 the dataset where all the values are multiplied by 10.

4.5.3 Artificial Data

Based on the above observation we have created a number of artificial datasets. A dataset of this kind is presented in Table 4.7. Consequently, we have analysed this dataset, say Data1, and another dataset where we multiplied all the data values by 10. This experiment was carried out in order to evaluate the effect on the posterior estimates when observing a sample of the population. The effect of observing a sample with $\alpha = 0.1$ is visualised in figures 4.15 and 4.16 for λ_G and λ_L respectively. It is clear from the figures that the location estimates are very close while again the posterior dispersion of the larger dataset is smaller. Quantitatively the posterior point estimates of λ_L and λ_G agree in both cases up to the third decimal point while the posterior variances for the large data set are approximately one tenth those of the smaller data set. This certainly is

not surprising. In fact it seems reasonable when we consider a simple example. In particular, we know from standard classical statistics that the variance of the mean of an i.i.d. sample is $\frac{1}{\alpha}$ times larger than the standard error of the mean of an i.i.d. sample that is $\frac{1}{\alpha}$ larger than the original sample. Additionally, under regularity conditions that we cannot verify in our approach, the sample mean and the maximum likelihood estimator (MLE) converge to the true value. There is a Bayesian equivalent to this approach, see for example the appendix of Gelman *et al.* (1995) and the references therein. The basic result of large-sample Bayesian inference is that when the prior influence diminishes (which is the case for "weak" priors or "strong" data) the posterior distribution of the parameter vector converges to a multivariate normal centred (if the likelihood model is correct) to the true parameter value θ_0. The asymptotic posterior variance can be shown to be $(nJ(\theta_0))^{-1}$ where $J(\theta_0)$ is the Fisher information

$$J(\theta_0) = E\left[\left(\frac{d\log L(x \mid \theta)}{d\theta}\right)^2\Bigg|_{\theta_0}\right] = -E\left[\left(\frac{d^2\log L(x \mid \theta)}{d\theta^2}\right)\Bigg|_{\theta_0}\right],$$

and $L(x \mid \theta)$ is the likelihood. Hence, after a number of implicit and explicit assumptions, it does seem plausible that the posterior variance of a dataset that consists of a random $\alpha \times 100\%$ sample of the population under study is $\frac{1}{\alpha}$ times larger than the variance that we would obtain if we were collecting data over the whole population. We emphasize that all these results hold under assumptions that we cannot verify but it is relatively easy to see why the empirical results hold in simpler settings. These findings remain when we apply our algorithm to different datasets. Hence, when we possess a huge dataset and we analyse a fraction of it it is probably reasonable to rescale the results based on the fraction we used. Additionally, if there is a good reason to believe that a larger part of the population behaves locally and globally (as opposed to globally only in chapter 3) like in our sample, then the posterior distribution of λ_G and λ_L may be appropriately rescaled as well.

	Parameter	
	λ_L	λ_G
Mean	0.0593	0.2291
Median	0.0594	0.2292
S. dev.	0.0065	0.0134
95% C. I.	(0.046,0.072)	(0.202,0.254)

Table 4.8: Posterior parameter summaries for the perfect data divided by 10 and rounded to the closest integer. The true values are $\lambda_L = 0.06$ and $\lambda_G = 0.23$.

4.6 Evaluation using Exact Results

We shall now attempt to assess the accuracy of our algorithm using two different tools. Firstly we shall apply our results to the "perfect" dataset presented in Table 3.9. Subsequently, we shall apply our algorithm to the homogeneous case i.e., the generalised stochastic epidemic, for which we have exact inference results based on the multiple precision solution of the triangular equations presented in chapter 2.

4.6.1 Perfect Data

A slight restriction of the methods presented in this chapter is that the algorithm is designed for data that are integers. This of course is never a problem for real data analysis but for a computational exercise of the kind that we consider here it does introduce a small bias. In particular, the perfect dataset from Table 3.9 was divided by 10 and was rounded to the closest integer. Hence we would expect a small deviation from the exact λ's. Indeed, the results summarised in Table 4.8 show that the point estimates of λ_L and λ_G are reassuringly close to

	Threshold	
	Triangular	Random Graph
Mean	1.218	1.221
Median	1.193	1.195
S. dev.	0.271	0.274
95% C. I.	(0.767,1.825)	(0.767,1.835)

Table 4.9: Posterior parameter summaries for the smallpox data.

the expected values.

4.6.2 Homogeneous Case

We applied the birth-death algorithm to the smallpox data that we have also used in the second chapter where 30 out of 120 individuals of a Nigerian village are getting ultimately infected. The application to a subcase of the two level mixing model, the generalised stochastic epidemic, has the advantage that we can compare the outcome with "exact" results that we derived in the second chapter using multiple precision arithmetic for the solution of the final size equations. The actual results for the basic reproduction number R_0 are presented in Table 4.9 and it is obvious that there is very close agreement between the two methods. An easier way to evaluate the accuracy of the random graph method is the visualisation of the two outcomes and the two posteriors are almost identical as can be seen from figure 4.17. Hence, it appears that the results obtained using the random graph method are practically equivalent to the exact ones.

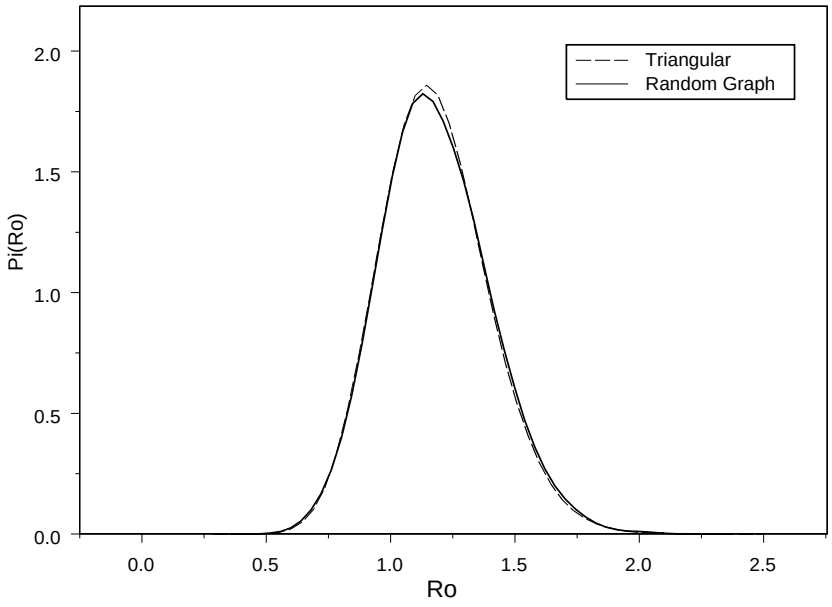

Figure 4.17: Posterior Density of R_0 using the random graph algorithm and the algorithm based on the multiple precision solution of the triangular equations.

4.7 Discussion

We have proposed methods for Bayesian inference for stochastic epidemic models using random graphs. The approach adopted in this chapter overcomes a number of previously unsolved problems in the area of statistical inference for stochastic epidemics. Additionally, the methodology can be extended to more complex, and realistic, models.

The methods presented in this chapter are related to methods used in statistical genetics, e.g. Stephens and Donnelly (2000). In particular, there is a number of similarities from the inverse problem perspective. Both problems are concerned with statistical inference for a stochastic process with complex structure when only the final state of the process is actually observed. Thus, using a sampling based approach, the underlying stochastic model is simulated conditionally on the data. Inference then proceeds according to these simulated realisations of the stochastic process of interest. However, the actual processes, despite a large number of similarities have a number of distinct features. Specifically, a coalescent looks like a tree while an epidemic can have links "backwards". Of course these edges will not result in infection, since these individuals have previously been infected and are removed during the future generations. However, the model is defined this way and the random graph should permit behaviour of this kind for (among others) the correct derivation of the likelihood. Additionally, the probabilistic analysis of epidemic models can reveal multiple levels of dependence, particularly when the outbreak has taken off. In contrast, in the coalescent there is mutation, a feature of the process that is of main interest in genetics. In practice this fact makes the analysis of the underlying process rather involved. In summary, the stochastic processes used to describe the mechanism that underlies disease propagation can be thought of as similar in nature to the coalescent that has been the ob-

ject of intense interest in statistical genetics recently. However, there remain a large number of distinct features that create the need for development of methodology that is suitably tailored towards the characteristics of epidemics.

An appealing characteristic of the methods described in this chapter, compared to the final severity approach of chapter 3, lies in the fact that the method is unconditional (on $R_* > 1$) and does not involve approximations. It is quite common in statistical methods for the analysis of stochastic processes that display threshold behaviour (like epidemics and branching processes) to condition upon non-extinction. In this chapter we have presented an alternative approach that may be extended to the analysis of different stochastic processes. A second feature of the random graph approach is that part of the output contains a decomposition of the local and global infections. This information may be of use from the applied viewpoint and further exploration of this issue is required.

The method we used can incorporate very general assumptions about the distribution of the infectious period. However, when only final size data are available there is very limited information from the statistical inference perspective (Rhodes *et al.* (1996)). In practice, as we have seen in both chapter 2 and chapter 4, it is only the variability of the estimates that is slightly affected. Hence, in extensions of the methodology presented here to more complex senarios it might be suitable to perform the initial analysis with a model with a constant infectious period since it might be easier to evaluate the augmented likelihood. Additionally, a constant infectious period results in simpler, and typically more numerically robust algorithms since in that case the number of links becomes a sufficient statistic.

Another attractive feature of the method presented in this chapter is its generality. In particular, the illustration was obtained using the two-level-mixing epidemic but extensions to three or more levels of mixing seem rel-

atively straightforward. Additionally, it might be of theoretical and applied interest to incorporate alternative ways of disease propagation like spread in a population with overlapping groups (Ball and Neal (2002)), additional spatial spread or multitype models (Ball and Lyne (2001)) that allow for the inclusion of covariates. Hence, the effort should go to the collection (and analysis) of more detailed data that can provide us with new quantitative insights on the transmission mechanism.

An alternative development could be the extension of our method to incorporate random population structures like the network models that received a considerable amount of attention during the last few years, see for example the review by Albert and Barabasi (2002). This approach could be tackled in two ways. One could obtain the (network) community structure using e.g. the algorithm of Girwan and Newman (2002). Then, it is natural to extend our methods to the (essentially fixed after the initial analysis) resulting population structure. A second, more challenging, approach is the simultaneous estimation of the population structure and the infection rates. Both approaches appear to deserve further exploration.

Chapter 5

Discussion and Future Work

In this thesis we presented methods of Bayesian statistical inference for stochastic epidemic models. The purpose of this analysis is twofold. We analysed stochastic models which describe the actual disease propagation. Hence these models can be used for a variety of different diseases. Additionally, we developed general methodology that can be extended to more elaborate, and realistic, epidemics.

In the first chapter we introduced the stochastic epidemic models of interest and Bayesian statistical methods as well as the modern computational tools that are necessary for the implementation of the Bayesian paradigm. Additionally, we reviewed existing methods of statistical inference for epidemic models.

In the second chapter we developed algorithms that use multiple precision arithmetic for the calculation of the final size probabilities. This novel algorithm allows us to assess the accuracy of the two most commonly used asymptotic results for epidemics, the branching process approximation for the initial stages of the epidemic and the normal approximation for the distribution of the final size in the event of a major epidemic. Our algorithm can also be

used for the evaluation of a third asymptotic theorem, the Poisson approximation for the number of individuals who ultimately escape infection, see Ball (1986) and Lefèvre and Utev (1995). We also used the final size distribution for statistical inference with respect to the threshold parameter R_0. Our approach is exact and it does not, in contrast with Rida (1991), assume that the epidemic is above threshold or that there is a large number of individuals.

The generalised stochastic epidemic that is the object of interest in the second chapter assumes that the population is homogeneously mixing. However, real life populations display particularly complex structures. Thus, we proceeded in the third chapter to develop methods of statistical inference for an epidemic model with two levels of mixing where, in a population that is partitioned into groups, an individual is allowed to have both within-group and population-wide infectious contacts. Assuming that the epidemic is above threshold and that we have a large number of groups, we approximated the likelihood using a previously derived central limit theorem for the final severity of the epidemic. We showed that previously derived inferences for epidemics among households arise as special cases in our framework and we discuss the limitations and the implicit assumptions of the different approaches. This approach can be extended to other variants of the basic model like the multitype case (Ball and Lyne (2001)) and models with overlapping subgroups (Ball and Neal (2002)). However, extensions of this kind require the development of appropriate asymptotic results (like those given in Ball and Lyne (2001)) in order to conduct approximate inference.

In the fourth chapter of this thesis we imputed detailed information about the infection spread in the form of a random graph. This approach is appealing for a number of reasons. The method is not approximate in the sense that we do not require an infinite number of households since we do not utilise asymptotic

results. Also, we do not condition upon non-extinction as is commonly assumed in stochastic and deterministic epidemic processes. Additionally, the output of the MCMC algorithm contains posterior information about the individual local and global contacts. Hence, we succeeded in achieving a decomposition that has been impossible to obtain before. This extra information can be particularly useful for applications like the design and assessment of different vaccination strategies. In contrast to our approximate method, the random graph approach can be easily extended to complex population structures like overlapping groups with any number of individuals residing in the part that the groups overlap. Additionally it would be particularly interesting to extend our methods to populations with random structures like the small-world and scale-free networks that have been the subject of intense interest in statistical mechanics recently.

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**, 961-974.

Albert, R. and Barabàsi, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**, 47-97.

Anderson, R.M. and May, R.M. (1991). *Infectious Diseases of Humans; Dynamics and Control.*, Oxford University Press, Oxford.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*, Springer Lecture Notes in Statistics, New York.

Bailey, N. T. J (1975). *The Mathematical Theory of Infectious Diseases and Its Application*, 2nd edn. Griffin, London.

Bailey, N. T. J and Thomas, A.S., (1971) The estimation of parameters from population data on the general stochastic epidemic. *Theor. Pop. Biol.*, **2**, 53-70.

Ball, F. G. (1983) The threshold behaviour of epidemic models. *J. Appl. Probab.*, **20**, 227-241.

Ball, F. G. (1986) A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. Appl. Probab.*, **18**, 289-310.

Ball, F. G., Britton T. and O'Neill, P. D., (2002) Empty confidence sets for epidemics, branching processes and Brownian motion. *Biometrika*, **89**, 211-224.

Ball, F. G. and Clancy, D. (1993) The final size and severity of a generalised stochastic multitype epidemic model. *Adv. in Appl. Probab.* **25**, 721-736.

Ball, F. G. and Lyne, O. D. (2001) Stochastic multitype SIR epidemics among a population partitioned into households. *Adv. Appl. Prob.*, **33**, 99-123.

Ball, F. G. and Lyne, O. D. (2002) Optimal vaccination policies for stochastic

epidemics among a population of households. *Math. Biosci.*, **178**, 333-354.

Ball, F. G. and Lyne, O. D. (2003) Statistical inference for epidemics among a population of households. In preparation.

Ball, F. G., Mollison, D. and Scalia-Tomba, G. (1997) Epidemics with two levels of mixing. *Ann. Appl. Probab.*, **7**, 46-89.

Ball, F. G. and Neal, P. J., (2004) Poisson approximations for epidemics with two levels of mixing. *Ann. Prob.*, **32**, 1168-1200 .

Ball, F. G. and Neal, P. J., (2003) A general model for stochastic SIR epidemics with two levels of mixing. *Mathematical Biosciences*, **180**, 73-102.

Ball, F. G., and O'Neill, P. D., (1999) The distribution of general final state random variables for stochastic epidemic models. *J. Appl. Prob.*, **36**, 473-491.

Barbour, A.D. and Mollison, D. (1989). Epidemics and Random Graphs. In *Stochastic Processes in Epidemic Theory. Lecture Notes in Biomath.*, **86**, 86-89. Springer, Berlin.

Bartlett, M.S. (1949) Some evolutionary stochastic processes. *J. Roy. Statist. Soc. Ser. B*, **11**, 211-229.

Bartlett, M.S. (1957) Measles periodicity and community size. *J. Roy. Statist. Soc. Ser. A*, **120**, 48-70.

Bartoszynski, R. (1972) On a certain model of an epidemic. *Applicationes Mathematicae*, **13**, 159-171.

Becker N. G., (1989). *Analysis of Infectious Disease Data*, Chapman and Hall, London.

Becker N. G., and Britton T. (1999) Statistical studies of infectious disease incidence. *J. R. Statist. Soc. B*, **61**, 287-307.

Becker N. G., and Britton T. (2001) Design issues for studies of infectious diseases. *J. Stat. Plan. Inf.*, **96**, 41-66.

Becker N. G., Britton T. and O'Neill, P.D. (2003) Estimating vaccine effects on transmission of infection from Household outbreak data. *Biometrics*, **59**,

467-475.

Becker N. G., and Dietz, K. (1999) The effect of the household distribution on transmission and control of highly infectious diseases. *Math. Biosci.*, **127**, 207-219.

Becker N. G., and Hopper, J.L. (1983) The infectiousness of a disease in a community of households. *Biometrika*, **70**, 29-39.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.*, Wiley, New York.

Bollobás, B. (2001). *Random Graphs*. Cambridge University Press, Cambridge.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley Classics, New York.

Brent, R.P. (1978) A Fortran multiple precision arithmetic package. *ACM Trans. Math. Soft.*, **4**, 57-70.

Britton T. and Becker N. G. (2000) Estimating the immunity coverage required to prevent epidemics in a community of households. *Biostatistics*, **1**, 389-402.

Britton T. and O'Neill, P.D. (2002) Bayesian inference for stochastic epidemics in populations with random social structure. *Scand. J. Stat.*, **29**, 375-390.

Clancy, D. and O'Neill, P.D. (2002) Perfect simulation for stochastic models of epidemics among a community of households. Research Report.

Congdon, P. (2001). *Bayesian Statistical Modelling*. Wiley, Chichester.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Demiris, N. and O'Neill, P.D. (2004) Bayesian inference for epidemics with two levels of mixing. *Scand. J. Stat*, to appear.

Diekmann, O. and Heesterbeek, J.A.P. (2000). *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Wiley.

Dietz, K. (1993) The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research*, **2**, 23-41.

Fearnhead, P. and Meligkotsidou, L. (2004) Exact filtering for partially-observed, continuous-time models. *J. R. Statist. Soc. B.*, To appear

Ferguson NM, Donnelly CA, Anderson RM. (2001) Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature*, **413**, 542-548.

Ferguson NM, Keeling, M Edmunds, Gani, A. Grenfell, B. Anderson, R. and Leach (2003) Planning for smallpox outbreaks. *Nature*, **425**, 681-685.

Gelfand, A.E. Lee, T.M. and Smith, A.F.M. (1990) Bayesian analysis of constrained parameter and truncated data problems. *J. Am. Stat. Ass.*, **87**, 523-532.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Ass.*, **85**, 398-409.

Gelman, A. Carlin, J.B. Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis.*, Chapman and Hall, London.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Tr. Pat. An. Mach. Int.*, **6**, 721-741.

Gibson, G.J. (1997) Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *Appl. Stat.*, **46**, 215-233.

Gibson, G.J. 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. Richardson, S. and Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice.* Chapman and Hall, London.

Girvan, M. and Newman, M. E. J. (2002) Community structure in social and biological networks. *Proc. Nat. Acad. Sci.*, **99**, 7821-7826.

Grenfel, B. and Dobson, A. (1995). *Ecology of infectious diseases in natural populations.*, Cambridge University Press, Cambridge.

Halloran, M.E., Longini, I.M., Nizam, A. and Yang, Y. (2002) Containing Bioterrorist Smallpox. *Science*, **298**, 1428-1432.

Halloran, M.E., Longini, I.M., and Struchiner, C.J. (1999) Design and interpretation of vaccine field studies. *Epidem. Rev.*, **21**, 73-88.

Hastings, W.K. (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. *Austr. N. Z. J. Stat.*, **45**, 491-502.

Higham, N.J. (1989) The Accuracy of Solutions to Triangular Systems. *SIAM J. Matrix Anal. Appl.*, **13**, 162-175.

Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, **19**, 293-325.

Isham, V. and Medley, G. (1996). *Models for infectious human diseases: Their structure and relation to data.*, Cambridge University Press, Cambridge.

Jagers, P. (1975). *Branching Processes with Biological Applications*. Wiley, New York.

Kaplan, E.H. Craft, D.L and Wein, L.M (2002) Emergency response to a smallpox attack: The case for mass vaccination. *Pr. Nat. Acad. Sci.*, **99**, 10935-10940.

Keeling, Woolhouse, Shaw, Matthews, Chase-Topping, Haydon, Cornell, Kappey, Wilesmith and Grenfell (2001) Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape. *Science*, **294**, 813-817.

Kulasmaa, K. (1982) The spatial general epidemic and locally dependent random graphs. *J. Appl. Probab.*, **19**, 745-758.

Lefèvre, C. and Picard, Ph. (1990) A non-standard family of polynomials and the final size distribution of Reed-Frost epidemic processes. *Adv. in Appl. Probab.*, **22**, 25-48.

Lefèvre, C. and Utev, S. (1995) Poisson approximation for the final state of a generized epidemic process. *Ann. Probab.*, **23**, 1139-1162.

Li, N., Qian, G., and Huggins, R., (2002) Analysis of between-household heterogeneity in disease transmission from data on outbreak sizes. *Aust. N. Z. J. Stat.*, **44**, 401-411.

Liggett, T.M. (1999). *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*. Springer, New York.

Lipsitch, M Cohen, T Cooper, B Robins, J.M Ma, S James, L. Gopalakrishna, G. Chew, S. Tan, C. Samore, M.H. Fisman, D. and Murray M. (2003) Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *Science*, **300**, 1966-1970.

Longini, I. M. and Koopman, J. S. (1982) Household and community transmission parameters from final distributions of infections in households. *Biometrics*, **38**, 115-126.

Marion, G., Gibson, G. J. and Renshaw, E. (2003). Estimating likelihoods for spatio-temporal models using importance sampling. *Statistics and Computing*, **13**, 111-119.

Mengersen, K. L. and Tweedie, R. L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, **24**, 101–121.

Metropolis, N. Rosenbluth, A. W. Rosenbluth, M.N. Teller, A.H Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087-1091.

Mollison, D. (1977) Spatial contact models for ecological and epidemic spread. *J. R. Statist. Soc. B.*, **39**, 283-326.

Mollison, D. (Ed.) (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. Am. Stat. Ass.*, **78**, 47-55.

Neal, P.J. Roberts G. O., and Viallefont, V. (2003) Robust MCMC algorithms for inference for stochastic epidemic models. submitted

Newman, M. E. J. (2004) Detecting community structure in networks. *Eur. Phys. J. B*, to appear.

O'Neill, P. D., (2002) A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences*, **180**, 103-114.

O'Neill, P.D. (2003) Perfect simulation for Reed-Frost epidemic models. *Stat. Comp.*, **13**, 37-44.

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. *Appl. Stat.*, **49**, 517-542.

O'Neill, P. D., and Becker, N. G., (2002) Inference for an epidemic when susceptibility varies. *Biostatistics*, **2**, 99-108.

O'Neill, P. D., and Roberts G. O., (1999) Bayesian inference for partially observed stochastic epidemics. *J. Roy. Stat. Soc. A.*, **162**, 121-129.

Propp, J.G and Wilson, D.B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Str. Alg.*, **9**, 223–252.

Rhodes, P. Halloran, M.E. and Longini, I.M. (1996) Counting process models for infectious disease data: distinguishing exposure to infection from susceptibility. *J. R. Statist. Soc. B.*, **58**, 751–762.

Rida, W.N. (1991) Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic. *J. R. Statist. Soc. B.*, **53**, 269-283.

Riley, S. Fraser, C. Donnelly, C. Ghani, A. Abu-Raddad, L. Hedley, A. Leung, G. Ho, L. Lam, T. Thach, T. Chau, P. Chan, K. Lo, S. Leung, P. Tsang, T. Ho,

W. Lee, K. Lau, E. 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**, 1961-1966.

Roberts G. O., (1996) Markov chain concepts related to sampling algorithms. In *Markov chain Monte Carlo in practice*. (Gilks, W. Richardson, S. and Spiegelhalter, D. ed.) Chapman and Hall, London.

Roberts G. O., (2003) Linking theory and practise of MCMC. In *Highly Structured Stochastic Systems* . (Green, P. J., Hjort, N. L., and Richardson S, ed.) Oxford University Press.

Roberts G. O., Gelman, A. and Gilks, W.R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms.. *Ann. Appl. Prob.*, **7**, 110-120.

Rubin, D.B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied Bayesian. *Ann. Statist.*, **12**, 1151-1172.

Rushton, S. and Mautner, A. J. (1955) The deterministic model of a simple epidemic for more than one community. *Biometrika*, **42**, 126-132.

Scalia-Tomba, G. (1985). Asymptotic final size distribution for some chain-binomial processes. *Adv. in Appl. Probab.*, **17**, 477-495.

Schinazi R. (2002) On the role of social clusters in the transmission of infectious diseases. *Theoretical Population Biology*, **61**, 163-169.

Shanks, D. and Wrench, J.W. (1962) Calculation of π to 100000 places. *Math. Comp.*, **16**, 76-99.

Smith, D.M. (1989) Efficient Multiple-Precision Evaluation of Elementary Functions. *Math. Comp.*, **52**, 131-134.

Smith, D.M. (1991) A Fortran Package For Floating-Point Multiple-Precision Arithmetic. *ACM Trans. Math. Soft.*, **17**, 273-283.

Smith, A.F.M., and Roberts, G.O.(1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J.*

R. Statist. Soc. B, **55**, 3-1020.

Spiegelhalter, D. Thomas, A. Best, N. and Gilks, W. (1996). *BUGS: Bayesian infecrence Using Gibbs Sampling, Version 0.60.* MRC Biostatistics Unit, Cambridge.

Stephens, M. and Donnelly, P. (2000) Inference in Molecular Population Genetics. *J. R. Statist. Soc. B*, **62**, 605-655.

Streftaris, G. and Gibson, G. J. (2004) Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling*, **4**, 1-13.

Strogatz, S.H (2001) Exploring complex networks. *Nature*, **410**, 268-276.

Watts, D.J. and Strogatz, S.H (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440-442.

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). *Adv. Appl. Prob.*, **3**, 223.