Epidemic Models and MCMC Inference

by

Ashley Ford

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

November 2014

THE UNIVERSITY OF
WARWICK

Contents

List of Figures

Acknowledgments

I have been fortunate to have Gareth Roberts as a supervisor and thank him for his support and guidance. In particular for providing rapid insight into difficulties when I encountered them and for giving me freedom to explore ideas, some of which did not make it into this thesis.

I also thank the statistics department at Warwick for providing such a stimulating environment in which to conduct research and to the other PhD students for helping me to feel younger than I am.

Most importantly thanks to my wife for agreeing to let me give up a well paid job to follow my interests and study.

Declarations

I hereby declare that this thesis is the result of my own work and research, except where otherwise indicated. This thesis has not been submitted for examination to any institution other than the University of Warwick.

Abstract

Statistical inference and model choice for partially observed epidemics provide a variety of challenges both practical and theoretical. This thesis studies some related aspects of models for epidemics and their inference.

The use of the matrix exponential to facilitate exact calculations in the General Stochastic Epidemic (GSE) is demonstrated, most usefully in providing the exact marginal likelihood when infection times are unobserved.

The bipartite graph epidemic is defined and shown to be a flexible framework which encompasses many existing models. It also provides a way in which a deeper understanding of the relation between existing models could be obtained.

The Indian buffet epidemic is introduced as a non-parametric approach to modelling unknown heterogeneous contact structures in epidemics. Inference for the Indian buffet epidemic is a challenging problem, some progress has been made. However the algorithms which have been studied do not yet scale to the size of problem where significant differences from the GSE are apparent.

Evidence confirming and demonstrating the importance of understanding the tail behaviour of proposals in importance sampling is presented. The adverse impact of heavy tailed proposals on the Grouped Independence Metropolis-Hastings (GIMH) and Monte Carlo within Metropolis (MCWM) algorithms is demonstrated.

A new algorithm, the Kernel Metropolis Hastings (KMH), is proposed as an approximate algorithm for low dimensional marginal inference in situations where the GIMH algorithm fails because of sticking. The KMH is demonstrated on a challenging 2-d problem.

Chapter 1

Introduction

Inference for partially observed epidemic models provides a variety of challenges. These include an appropriate choice of model, in particular for the contact process which has a major effect on the infection process. A second challenge is that existing Markov chain Monte Carlo (MCMC) algorithms for inference on epidemics have difficulty scaling to large populations. This thesis provides new results in three themes which form the basis of three planned papers which are:

- Exact calculation in epidemic models and inference
- Indian buffet epidemics and their inference
- Analysis of a related group of MCMC algorithms introduced in section 1.3.

These are presented as an integrated thesis, with the necessary background, in the chapters:

- 2 Epidemic Models
- 3 MCMC
- 4 Epidemic Inference

A detailed introduction to each chapter is given at its beginning, first a summary and introduction to the three themes is given, highlighting the original contributions of this thesis.

1.1 Exact Calculation in Epidemic Models and Inference

Many asymptotic results have been obtained for epidemic models which provide valuable insight into their behaviour in large populations. Advances in computer power mean that calculations using the exact Markov chain representation of an epidemic and the matrix exponential are now feasible. The representation was first noted by Bailey (1953) but not pursued in detail. More recently it has been described by several authors including Allen (2008) and Keeling and Ross (2008). In this thesis it is used to perform exact numerical computations and identify some previously unreported features of the epidemic threshold, it is also used to compare a regularly observed continuous time epidemic model with a binomial based model. The main motivation for investigating the Markov chain representation is to demonstrate its use in inference, this is done in two ways. The potential for using the exact transition matrix for inference in regularly observed general stochastic epidemics (GSEs) is demonstrated. Exact calculation of the marginal likelihood for the removal times of the GSE is developed and demonstrated to be feasible on a population of 120. An approximation that will allow scaling to larger populations is described in an appendix. The potential for a bi-modal posterior distribution for the parameters of an in progress GSE has been identified.

1.2 Indian Buffet Epidemics and their Inference

The class of epidemic models on bipartite graphs provides a powerful way of constructing new models and comparing existing models. The bipartite graph epidemic model is defined which provides a foundation for the new model, the Indian buffet epidemic, which is developed and studied through simulation. Inference for the Indian buffet epidemic using a variety of MCMC algorithms is studied, the most succesful algorithm does not scale to interesting problem sizes. The reasons for difficulties with other approaches are identified.

1.3 Analysis of GIMH and MCWM via SEMH and SAMH and the Kernel MH algorithm.

The grouped independence Metropolis-Hastings (GIMH) algorithm which was introduced by Beaumont (2003) and generalised by Andrieu and Roberts (2009) is a potentially useful MCMC algorithm, for inference in many situations including epidemics. However when a poor proposal is used it can suffer from ”sticking”. A novel analysis of some aspects of the GIMH and the bias of the closely related approximate algorithm the Monte Carlo within Metropolis algorithm (MCWM) (O’Neill et al., 2000) is presented. The analysis is presented in more general terms, to distinguish the original algorithm from the generalised algorithm the name stochastic exact Metropolis-Hastings (SEMH) is introduced. The analysis of GIMH is based on an analysis of some aspects of importance sampling in tractable situations which is given in section 3.2. The analysis in section 3.5.3 shows that the variance of the weights distribution can explain the sticking of the GIMH and the bias of the MCWM.

A new approximate algorithm, the Kernel Metropolis-Hastings (KMH) is proposed in section 3.6 which is expected to overcome the difficulties encountered in applying the GIMH algorithm in practice which are described in chapter 4. The KMH is demonstrated on a multimodal heavy tailed target distribution.

Chapter 2

Epidemic Models

2.1 Introduction

This chapter considers two distinct aspects of stochastic models for epidemics, the Markov representation of the general stochastic epidemic (GSE) and bipartite graph epidemics from which the Indian buffet epidemic is introduced.

The Markov representation of the general stochastic epidemic (GSE) model is used to derive exact distributions of various quantities which are of interest in their own right and are used as the basis for exact inference in chapter 4. Some of the known limitations of the simple models are used as motivation for considering models based on bipartite graphs for epidemics with heterogeneous contact processes.

As a pre-requesite in section 2.1.1 the basic epidemic models and the epidemic threshold are introduced. In section 2.3 the transition matrix of the GSE is used to perform exact numerical computations and identify some previously unreported features of the GSE.

The embedded Markov chain (EMC) of the Markov representation is introduced and used in section 2.3.3 to calculate the final size distribution in the GSE. This is then used to investigate the threshold between “minor” and “major” epidemics and identify the regions of the parameter space where the classic bi-modal behaviour is present.

The EMC is also used to calculate the joint distributions of final size and the number of infectives immediately after the first removal, also the joint distribution of final size and the maximum number of infectives at any time is calculated.

Continuous features of the GSE are studied in section 2.3.4 where the continuous time Markov process is used to calculate the exact distribution of duration conditioned on final size. A plot based on a well known martingale is introduced in section 2.3.5 which is later used to portray the difference between the GSE and other models. An exact comparison between the regularly observed GSE and a very similar binomial discrete time model is made in section 2.4.1.

As motivation for the development in sections 2.6 and 2.7 of the Indian buffet epidemic section 2.5 reviews epidemic models incorporating heterogeneous contact structures. As further motivation section 2.6.4 shows how some of these models can be considered as particular instances of the bipartite graph epidemic.

The Indian buffet epidemic is introduced and studied in section 2.7. After describing the Indian Buffet Process (IBP) section 2.7.2 defines the new epidemic model. The remainder of the chapter studies features of this new epidemic model, presenting results of simulations showing the aspects of the variability that emerge.

2.1.1 Basic Epidemic Models

A number of different stochastic and deterministic models for epidemics have been proposed, in both continuous and discrete time. The models are all an approximation in medical or veterinary terms to the complex processes involved in acquiring, incubating and transmitting a disease. All models must balance the often competing aspects of simplicity, realism and tractability for analytic results. A good qualitative understanding of fundamental behaviour of both simple and complex models can be obtained from analytic results obtained from simple models and their asymptotic behaviour. It should be noted that a model that is intractable for analytic results may be amenable to modern computational techniques of statistical inference and Mollinson in chapter 2 of Mollison (1995) states

> "The realistic detail of a stochastic model, specifying such things as the probability that one individual will infect another at a particular time and place, has long been recognised as a strength from the point of view of understanding and fitting models, but has generally been regarded as a grave handicap when it comes to analysis; even stochastic analyses have traditionally dealt whenever possible with massed variables such as the total number of infectives. However, in recent years there has been an increasing recognition that the unnecessary detail of a stochastic model framed in terms of individuals and their interactions can in many cases allow insights not possible from a higher level stochastic or deterministic model."

Several books are available that consider a wide range of models, for example Brauer et al. (2008), while other books focus on a more detailed analysis of a subset of models e.g. Daley and Gani (1999).

There is always a trade off between the realism of assumptions in the model and ease of analysis and identifiability of parameters. All the models considered in this thesis categorise individuals as being in one of a small number of states, with transitions between them at well defined but often unobserved times. This approximation can be justified by the lack of data in most situations[1] on which any more complex model could be based. The three states in the basic model considered are:

Susceptible The individual is uninfected and could be infected.

Infectious Capable of transmitting the disease, they have been infected with the disease and are capable of spreading the disease to those in the susceptible state.

Removed is the compartment used for those individuals who have been infected and then recovered from the disease or been isolated or died. In all the models considered those in this category are not able to be infected again or to transmit the infection to others.

The initial letters of the names of the states Susceptible,Infectious,Removed provide the name for this class of models: SIR models. The most notable omission in this model, particularly for some diseases such as smallpox is the lack of an exposed state, this is frequently added as an additional state giving rise to the class of SEIR models which incorporate this additional state:

Exposed The individual is infected but not yet infective, they are in a latent period and will progress to the Infectious state.

A more realistic model might have non-constant infectivity during the infectious period or include modelling of symptoms which are usually taken as coincident with the infectious period. Diseases such as malaria involving a parasite and those with variants and/or partial immunity require additional states, other states such as asymptomatic but infectious may be needed for some diseases. Long lasting diseases and childhood diseases such as measles, require consideration of births and deaths. These aspects are not considered further in this thesis.

A range of observed or unobserved co-variates for individuals, in particular age or location, can have significant effects on some or all of contact patterns,

[1] except in a few laboratory experiments

infectiousness, susceptibility and durations of phases. Some aspects of this are considered in section 2.5.

Approaches to the modelling of the transitions between the states can be categorised as deterministic or stochastic and continuous or discrete time. The distinction between a discrete time model and discrete time observation of a continuous model is considered in section 2.4. Deterministic models usually give rise to a system of differential equations which are closely related to an equivalent stochastic model, and often shed light on the non-linear evolution of expected values in the stochastic model and the presence of thresholds, they are not considered explicitly.

2.2 Continuous Time Epidemic Models

2.2.1 The General Stochastic Epidemic (GSE)

This standard SIR homogeneous mixing model is generally called the general stochastic epidemic (GSE) and was first described in the papers by Kermack and McKendrick in 1927 reprinted in Kermack and McKendrick (1991a,b,c). Diekmann has pointed out (Diekmann et al., 1990) that their General Stochastic Epidemic model was more flexible than the basic Markov SIR model often attributed to them and described here as the GSE. This model provides a good starting point for more complex models, most of which can be considered as generalisations of the GSE. This seemingly simple model also provides interesting challenges for inference, some of which are studied in chapter 4. A good understanding of this seemingly simple model is necessary for fully understanding more complex models and to give insight into difficulties that can be encountered in inference.

Two equivalent approaches to presenting SIR models are possible, either based on counts of individuals in the categories S,I,R and rates of transition or based on the state of individuals and their duration in each state. The latter approach is more easily extended to non-exponential distributions of time in the infectious state and also has some advantages in inference, both in the GSE and more complex models. The counts approach is used here as it naturally gives the Markov process representation which is used below. Although the duration of infectiousness is clearly not exponentially distributed, it can be shown that most bulk properties of the model are only affected by the mean of the distribution (Andersson and Britton, 2000) and so the implied use of an exponential distribution can be justified.

This widely studied SIR epidemic model considers the progress of an epidemic in an initially susceptible population of n_p individuals, after an initial infection from outside the population of one individual. Each individual is in one of the three states

*S*usceptible, *I*nfective, *R*emoved[2], and the epidemic progresses via two independent transitions:

infection $S \to I$ one susceptible individual changes from *S*usceptible to *I*nfective,

removal $I \to R$ one infected individual changes from *I*nfective to *R*emoved.

The transition rates between the states are taken to be Markov, dependent only on the counts of individuals in the two states *S*usceptible and *I*nfective at time t denoted $(S(t), \ I(t))$. In this model the population transitions are Poisson with time and state dependent rates:

$$\begin{array}{llll} S \to I & \text{at rate} & \lambda S(t) I(t), & \text{the state changes to} \quad (S(t)-1, \ I(t)+1) \\ I \to R & \text{at rate} & \rho I(t), & \text{the state changes to} \quad (S(t), \ I(t)-1). \end{array}$$

and can be shown in equivalent terms of transition probabilities as

$$P(S(t+dt) = S(t) - 1 \,\&\, I(t+dt) = I(t)+1) = \lambda S(t) I(t) \ dt \quad (2.2.1)$$

$$P(I(t) = I(t) - 1) = \rho I(t) \ dt \quad (2.2.2)$$

Authors differ in notation, sometimes the population is described in terms of N initial susceptibles and a initial infectives, giving a total population of $n_p = N+a$, and sometimes a factor of $1/N$ is extracted from the infection rate, each choice simplifies the notation for some calculations and complicates others, also the use of $N+a$ can clarify most asymptotic results. Some modellers argue over the presence of the $1/N$ and the meaning of infection rates with and without it, a clarification is given by Begon et al. (2002).

The number removed is denoted $R(t)$ and $S(t) + I(t) + R(t) = n_p$ for all t. When it is necessary to refer to the cumulative number that have been infected at any time in the interval $[0,t]$ then $C(t) = R(t) + I(t)$ is used.

When the GSE is presented in the form give by equation 2.2.1 it is apparent that it belongs to the class of density dependent jump processes studied in the monograph by Kurtz (1981) and so a rich set of asymptotic results are applicable, such as the convergence to a diffusion process, which is described briefly below.

The epidemic threshold and major epidemics The most significant feature of the GSE, in common with most SIR epidemic models is the existence of an

[2] including recovered, quarantined or death

epidemic threshold. Two typical example simulations of a GSE with the same parameters $\lambda = .015, \rho = 5, n_p = 1000$ are shown in figure 2.2.1, with very different outcomes, the epidemic is labelled "major" or "minor" dependent on whether the final size is a significant proportion of the population. In this example over many simulations approximately 67% of the epidemics would be "major". As $n_p \to \infty$ the distinction becomes clearer and the stochastic epidemic converges to the equivalent deterministic model. If λ and ρ vary with n_p such that $\lim_{n_p \to \infty} n_p \lambda / \rho$ exists then the limit is called the basic reproduction number $\mathcal{R}_0$. A common assumption is that $\lambda = n_p \beta$ for some constant β (see previous page) and ρ is constant, in this case$\mathcal{R}_0$ controls the asymptotic behaviour as $n_p \to \infty$. In particular in an infinite population if $\mathcal{R}_0 \leq 1$ then with probability 1 only a finite number will be infected whereas if $\mathcal{R}_0 > 1$ there is a positive probability of an infinite number of infections (Whittle, 1955). In finite populations and more complex models there can be some ambiguity over the definition of both a "major" epidemic and of $\mathcal{R}_0$, it is however a useful concept both theoretically to guide deeper understanding and practically to guide interventions such as vaccination. Many results, particularly in inference, condition on the epidemic being "major" often without defining it, an exception is Demiris and O'Neill (2006) where a clear definition and analysis is made. An investigation on the boundaries between "major" and "minor" epidemics is given below. The usual definition of $\mathcal{R}_0$ is the expected number of infections directly caused by the initial infective, see Pellis et al. (2012) for a study of $\mathcal{R}_0$ in household models.

In the GSE the expected number of infections caused by the initial infective is $(n_p - 1)\lambda/\rho$, for simplicity in notation the simpler formula $\mathcal{R}_0 = n_p \lambda / \rho$ is used to define $\mathcal{R}_0$ in the finite GSE considered here.

The total number of infections when the epidemic terminates, with zero infectives, is called the final size which is denoted by $\mathfrak{R}_\infty$. The distribution of this and its relation to $\mathcal{R}_0$ is investigated in section 2.3.3 using the Markov chain representation developed below.

Diffusion approximation In large populations the variability of the GSE can be represented by a diffusion approximation, which is presented in several books for example Allen (2008). The mean and variance (to order Δt) of the increments $\Delta X_t = X_{t+\Delta t} - X_t$ of the stochastic process $X_t = (S(t), I(t))^T$ follows from the definition in equation 2.2.1 as

$$\mathbb{E}(\Delta X_t) = \begin{pmatrix} -\lambda S(t) I(t) \\ \lambda S(t) I(t) - \rho I(t) \end{pmatrix} \Delta t$$

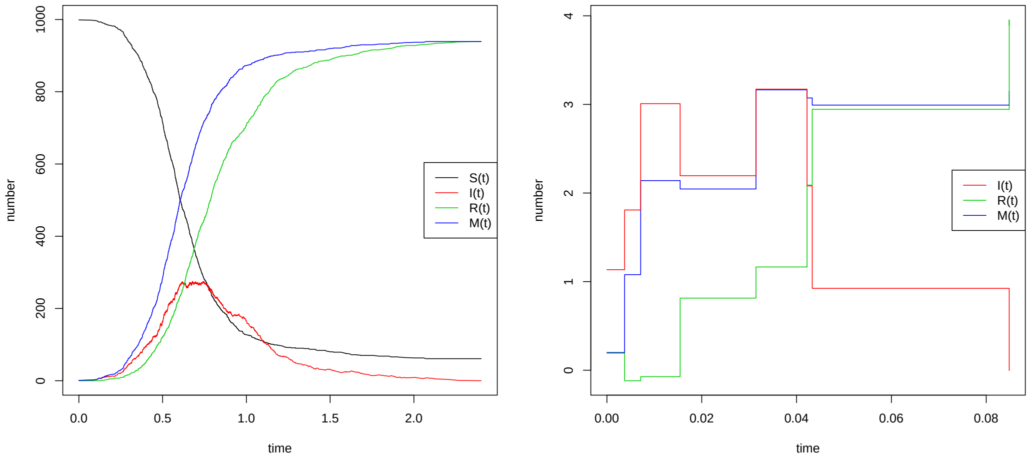


Figure 2.2.1: Simple examples of SIR GSE $\mathcal{R}_0 = 3, \lambda = .015, \rho = 5, n_p = 1000$ (the counts in the right hand plot have been "jitter-ed" to separate the lines). The line labeled M(t) in the plot is the cumulative number of infections C(t).

and

$$\mathrm{Var}\left(\Delta X_t\right) = \begin{pmatrix} \lambda S\left(t\right) I\left(t\right) & -\lambda S\left(t\right) I\left(t\right) \\ -\lambda S\left(t\right) I\left(t\right) & \lambda S\left(t\right) I\left(t\right) + \rho I\left(t\right) \end{pmatrix} \Delta t$$

Because the covariance matrix is symmetric and positive definite it has a unique square root and an equivalent Itô SDE representation is possible. When considering large populations the scaled process X_t/n_p is usually studied.

The results of Kurtz (1978) provide a more rigorous derivation and give asymptotic results for where the approximation is valid. The most important restriction is that the approximation is only valid away from the absorbing state of $I\left(t\right) = 0$, the exact calculations below investigate the relation between the early states when $I\left(t\right)$ is small and the eventual outcome.

2.3 The Markov Representation of the General Stochastic Epidemic (GSE)

The Markov representation of the GSE $X_t = \left(S\left(t\right), I\left(t\right)\right)^T$ was first noted by Bailey (1953) but not pursued in detail. More recently, as readily available computational power has become available it has been described by several authors including Allen (2008) and Keeling and Ross (2008). The Kolmogorov forward equations of this process were used by Bailey (1964) to derive differential equations for the moment generating function of $I(t)$. Here after presenting some well known matrix techniques, a matrix representation of the transition matrix of the GSE is used to perform exact numerical computations and identify some previously unreported features of the GSE.

2.3.1 Matrix Techniques for Markov Processes and Chains

The theory of Markov processes and chains is well developed and covered in many books for example Parzen (1962) or Norris (1998), on which this section is based. A general formulation is in terms of a stochastic process $\{X_t, t \in \mathcal{T}\}$, where $\mathcal{T}$ is an ordered set and X_t is a family of random variables $X_t : \Omega \to \mathcal{S}$, where $\mathcal{S}$ is some measurable space. The sets $\mathcal{S}$ and $\mathcal{T}$ are the state space and an index space, usually time, the mathematical level is significantly reduced by only considering finite[3] state spaces $\mathcal{S}$ and distinguishing two time index sets $\mathcal{T}$, when $\mathcal{T} \subset \mathbb{Z}$ we refer to a Markov chain and when $\mathcal{T} \subset \mathbb{R}$ we refer to a Markov process, terminology still differs and though common this usage is not universal. The fundamental property that distinguishes a Markov process (or chain) from other stochastic processes is that given the current state the future is independent of the past, this can be expressed as

$$\mathbb{P}\left(X_{t_n} = x_{t_n} | X_{t_1} = x_1, X_{t_2} = x_2, \ldots X_{t_{n-1}} = x_{n-1}\right) = \mathbb{P}\left(X_{t_n} = x_{t_n} | X_{t_{n-1}} = x_{n-1}\right)$$

for any ordered set of n times $t_1 < t_2 < \ldots < t_n$, all $\in \mathcal{T}$.

A time homogeneous Markov chain X_n, with $\mathcal{T} = \mathbb{Z}^+ = \{0, 1, 2, ..\}$ is defined by its transition probability matrix $\mathbf{P} = (p_{ij} : i, j \in \mathcal{S})$ where

$$\mathbb{P}\left(X_{n+1} = j | X_n = i\right) = p_{ij} \, \forall \, i, j \in \mathcal{S}$$

and its initial distribution $\mathbb{P}\left(X_0 = i\right) = p_i^{\text{init}} \, \forall \, i \in \mathcal{S}$, where $\sum_j p_{ij} = 1 \, \forall \, i \in \mathcal{S}$ and $p_{ij} \geq 0$, the initial distribution $p^{\text{init}} = (p_i^{\text{init}}, i \in \mathcal{S})$ is taken to be a row vector. The Chapman-Kolmogorov equation then gives

$$\mathbb{P}\left(X_{n+2} = k | X_n = i\right) = \sum_{j \in \mathcal{S}} p_{ij} p_{jk} \, \forall \, i, k \in \mathcal{S} \text{ and } n \in \mathcal{T}$$

[3] This chapter only considers finite spaces, the chapter on MCMC where larger state spaces are used does not make use of this section.

and so the n step transitions are obtained using matrix multiplication as

$$\mathbb{P}\left(X_{n+m} = j | X_m = i\right) = \left[\mathbf{P}^n\right]_{ij} \forall\, i, j \in \mathcal{S},$$

and the distribution at time n is

$$\mathbb{P}\left(X_n = i\right) = \left[p^{\text{init}}\mathbf{P}^n\right]_i \forall\, i \in \mathcal{S}.$$

The structure and long term behaviour of a chain is governed by the communicating classes of the chain. A state k is said to be accessible from j if $\left[\mathbf{P}^n\right]_{jk} > 0$ for some $n \in \mathcal{T}$ and if both k is accessible from j and j is accessible from k we say that j and k communicate. This symmetric and transitive relation gives the communicating classes, $C(j) \subset \mathcal{S}$ where $k \in C(j)$ if and only if j and k communicate, a chain with a single communicating class is called irreducible. A state j is absorbing if $C(j) = \{j\}$ and $p_{jj} = 1$. The communicating class of a state j will be empty, if the states accessible from j and those from which j is accessible are disjoint. A state is recurrent if eventual return to it occurs with probability 1, otherwise it is transient.

A Markov process, with $\mathcal{T} = \mathbb{R}^+ = [0, \infty)$, can be defined by the distribution of the initial state $\mathbb{P}\left(X_0\right)$ (as in the discrete case) and a matrix of transition rates $\mathbf{Q} = (q_{ij}, i, j \in \mathcal{S})$ where $q_{ij} \geq 0$ for all $i \neq j$ and $\sum_{j\in\mathcal{S}} q_{ij} = 0 \forall\, i \in \mathcal{S}$ it is convenient to introduce $q_i = -q_{ii}$ and noting that $q_i \geq 0$ we have

$$\mathbb{P}\left(X_{t+\delta} = j | X_t = i\right) = \begin{cases} \delta q_{ij} + o(\delta) & \forall\, i \neq, j \in \mathcal{S} \\ 1 - \delta q_i + o(\delta) & i = j \end{cases}.$$

The right hand side can be written in matrix form as $(\mathbf{I} + \delta\mathbf{Q})_{ij}$ and the transition probability matrix for any $t \geq 0$, is given by $\mathbf{P}(t) = e^{t\mathbf{Q}}$ where $\left[\mathbf{P}(t)\right]_{ij} = \mathbb{P}\left(X_{t+s} = j | X_s = i\right)$, see for example Norris (1998) chapter 2. The matrix exponential is defined by $e^{t\mathbf{Q}} = \sum_{j=0}^{\infty} \frac{(t\mathbf{Q})^k}{k!}$, methods for numerical calculation and some properties of it are given in Appendix A. Term by term differentiation gives

$$\frac{d}{dt}\mathbf{P}(t) = \sum_{k=1}^{\infty} \frac{t^{k-1}\mathbf{Q}^k}{(k-1)!} = \mathbf{P}(t)\mathbf{Q} = \mathbf{Q}\mathbf{P}(t)$$

the Kolmogorov forward and backward equations.

Embedded Markov Chain

Associated with any continuous time Markov process on a finite state space is the embedded Markov chain (EMC) which describes the path through the states, without regard to time. It is sometimes called the jump chain, (§2.6 of Norris, 1998). This chain is used to derive expressions for several quantities in the GSE. The EMC is a discrete time Markov chain, denoted $\mathfrak{X}_i$ on the same state space as the full Markov process $\{X_t, t \in \mathcal{T}\}$. The subscripts i, j, k are used to indicate discrete time and s, t continuous time. The transition matrix (p_{ij}) for $\mathfrak{X}_i$ is obtained from $\mathbf{Q}$ the transition rate matrix of X_t as

$$p_{ij} = \begin{cases} q_{ij}/\sum_{k\neq i} q_{ik} & i \neq j \text{ and } q_{ii} \neq 0 \\ 0 & i = j \text{ and } q_{ii} \neq 0 \\ 0 & i \neq j \text{ and } q_{ii} = 0 \\ 1 & i = j \text{ and } q_{ii} = 0 \end{cases} \tag{2.3.1}$$

or equivalently $\mathbf{P} = \mathbf{I} - \text{diag}(\mathbf{Q})^{-1}\mathbf{Q}$ (taking 0/0 as 1 on the diagonal, for any absorbing states).

The Fundamental matrix of an absorbing Markov chain

The fundamental matrix of an absorbing Markov chain Kemeny and Snell (1976) is a powerful way of calculating some moments and probabilities on an absorbing Markov chain. Consider a discrete time absorbing chain with transition matrix $\mathbf{P}$ with m absorbing states and n transient states, with a suitable ordering of states we have

$$\mathbf{P} = \begin{pmatrix} \mathbf{S} & \mathbf{R} \\ 0 & \mathbf{I} \end{pmatrix}$$

where $\mathbf{S}$ is an $n \times n$ matrix of transition probabilities within the transient states, $\mathbf{R}$ is an $n \times m$ matrix of transition probabilities from the transient to the absorbing states and I is an $m \times m$ identity matrix. Sometimes it is convenient to merge the absorbing states into a single absorbing state.

Definition 1. The fundamental matrix $\mathbf{N}$ of an absorbing Markov chain has entries $[N]_{ij}$ which are the expected number of visits to state j before absorption starting from state i.

Theorem 1. *The fundamental matrix of an absorbing Markov chain can be calculated as* $\mathbf{N} = (\mathbf{I} - \mathbf{S})^{-1}$.

Proof. The proof is given by Kemeny and Snell (1976) (theorem 3.2.1) based on $\mathbf{N} = \sum_{k=0}^{\infty} \mathbf{S}^k$ giving the probability of a visit to a state at step k. The matrix inverse always exists. □

Corollary 1. *The expected number of steps before absorption starting at state i is $\sum_{j=1}^{n} \mathbf{N}_{ij}$ or $\mu_s = \mathbf{N1}$.*

Proof. Kemeny and Snell (1976) (theorem 3.2.4) □

Corollary 2. *The probability of being absorbed in absorbing state j when starting from transient state i is $[\mathbf{B}]_{ij}$ where $\mathbf{B} = \mathbf{NR}$.*

Corollary 3. *The variance of the number of steps before absorption, starting at state i, is the ith term of* $(2\mathbf{N} - \mathbf{I})\mu_s - \mu_s \circ \mu_s$, where $\circ$ indicates a Hadamard product.

The Fundamental matrix of an absorbing Markov process

The fundamental matrix of an absorbing Markov process, and similar results to those above, are derived from the Markov chain results as follows. Consider a continuous time absorbing Markov process with transition rate matrix $\mathbf{Q}$ with m absorbing states and n transient states, with a suitable ordering of states we have

$$\mathbf{Q} = \begin{pmatrix} \mathbf{S} & \mathbf{R} \\ 0 & 0 \end{pmatrix}$$

where $\mathbf{S}$ is now an $n \times n$ matrix of transition rates within the transient states, $\mathbf{R}$ is an $n \times m$ matrix of transition rates from the transient to the absorbing states. Consider the Markov chain with transition probability matrix $\mathbf{P} = e^{\delta \mathbf{Q}}$ for some $\delta > 0$, as δ approaches 0 we can write $\mathbf{P}_\delta = \mathbf{I} + \delta\mathbf{Q}$, dropping the $o(\delta)$ terms, this is a stochastic matrix so long as $\delta < \min_i(q_i)$.

Lemma 1. *The fundamental matrix of an absorbing Markov process is $\mathbf{N} = -\mathbf{S}^{-1}$ and $[\mathbf{N}]_{ij}$ is the expected time spent in state j before absorption starting from state i.*

Proof. Denote the fundamental matrix of $\mathbf{P}_\delta$ as $\mathbf{N}_\delta$ where

$$\mathbf{P}_\delta = \begin{pmatrix} \mathbf{S}_\delta & \mathbf{R}_\delta \\ 0 & \mathbf{I} \end{pmatrix}$$

so $\mathbf{S}_\delta = \mathbf{I} + \delta\mathbf{S}$ and $\mathbf{R}_\delta = \mathbf{R}$ and $\mathbf{N}_\delta = (\mathbf{I} - \mathbf{S}_\delta)^{-1} = (-\delta\mathbf{S})^{-1}$ and $\delta\,[\mathbf{N}_\delta]_{ij}$ is the expected time spent in state j before absorption starting from state i and $\delta\mathbf{N}_\delta = -\mathbf{S}^{-1}$. Now consider the limit as $\delta \to 0$, it follows from theorem 2.8.2 of Norris (1998) that the Markov chains with transition probability matrix $\mathbf{P}_\delta = e^{\delta \mathbf{Q}}$ converge to the Markov process with transition rate matrix $\mathbf{Q}$. □

Corollary 2 for the discrete chain applies unchanged, and is repeated here.

Corollary 4. *The probability of being absorbed in absorbing state j when starting from transient state i is $[\mathbf{B}]_{ij}$ where $\mathbf{B} = \mathbf{NR}$.*

Corollary 1 requires a change of words to:

Corollary 5. *The expected time before absorption starting at state i is $\sum_{j=1}^{n} \mathbf{N}_{ij}$ or $\mu_d = \mathbf{N1}$.*

A small change is required to corollary 3 to give

Corollary 6. *The variance of the time to absorption, starting at state i, is the ith term of $2\mathbf{N}\mu_d - \mu_d \circ \mu_d$.*

Proof. The time to absorption for the $\mathbf{P}_\delta$ chain is $\delta\times$ the number of steps, so the mean is $\delta\mu_s$ and the variance of the time to absorption is $\delta^2\left((2\mathbf{N}_\delta - \mathbf{I})\mu_s - \mu_s \circ \mu_s\right)$ as $\delta \to 0$ $\delta\mu_s \to \mu_d$ and noting that $\mathbf{N}_\delta = \mathbf{N}/\delta$ the variance is $((2\mathbf{N} - \delta\mathbf{I})\delta\mu_s - \delta\mu_s \circ \delta\mu_s)$ which $\to 2\mathbf{N}\mu_d - \mu_d \circ \mu_d$.

□

Distribution of time to absorption of a Markov process

Lemma 2. *The joint probability density function (p.d.f.) of time to absorption and probability of final state j of a Markov chain starting at state i is $[\mathbf{Q}\exp(t\mathbf{Q})]_{ij}$, where $j \in \mathcal{I}_{abs}$ and $\mathcal{I}_{abs} \subset \mathcal{S}$ is the set of absorbing states.*

Proof. We have where $\mathbb{P}\left(X_{t+s} = j | X_s = i\right) = \left[e^{t\mathbf{Q}}\right]_{ij}$ for any i, j and as j is absorbing this is the cumulative distribution function (c.d.f.) of time to absorption in j times the probability that ultimate absorbtion is in j. The result follows by term by term differentiation of the power series for $\exp(t\mathbf{Q})$. □

Corollary 7. *The p.d.f. of time to absorption of a Markov process from an initial distribution p^{init} is $\sum_i p_i^{init} \sum_{j\in\mathcal{I}_{abs}} [\mathbf{Q}\exp(t\mathbf{Q})]_{ij}$*

2.3.2 Representation of the GSE as a Markov Process

We match the usual notation for state space models by denoting the state at time t as $X_t = (S\left(t\right),\ I\left(t\right))$ and the parameters by $\theta = (\lambda, \rho)$, with $\lambda > 0$ and $\rho > 0$. $X_t \in \mathcal{X} \subset \mathbb{Z}^2$ such that $S\left(t\right) + I\left(t\right) \le n_p$. $\mathcal{X}$ has size $n_s = (n_p+1)(n_p+2)/2$ which limits the sizes of problem that can be handled directly using the exact Markov transition matrix. However sizes up to $n_p = 120$ can be handled efficiently for all the calculations using the approaches developed below and the calculations based on the EMC have been performed for $n_p = 1200$.

Recall that the possible transitions are

$$\begin{array}{lllll} S \to I & \text{at rate} & \lambda S\left(t\right) I\left(t\right) & \text{the state changes to} & \left(S\left(t\right)-1,\ I\left(t\right)+1\right) \\ I \to R & \text{at rate} & \rho I\left(t\right) & \text{the state changes to} & \left(S\left(t\right),\ I\left(t\right)-1\right). \end{array}.$$

The transition rate matrix $\mathbf{Q}_\theta$ is readily computed from the transition rates above and contains $O(n_p^2)$ non zero entries from which the exact probability of any transition can be computed as:

$$\mathbb{P}(X_{t+s} = x_{t+s}|X_t = x_t) = \left[\exp(s\mathbf{Q}_\theta)\right]_{x_t,x_{t+s}} \tag{2.3.2}$$

and so for the given initial state $X_0 = (n_p - 1, 1)$ the probabilities of the GSE being in any state at time t are available. As the state space is finite, certain technical details that are needed for infinite but countable state spaces are not discussed.

We can see that this describes a random walk on $\mathbb{Z}^2$, and the dynamics are such that no state is visited twice. The state space and possible transitions for a very small example are shown in figure 2.3.1 where the green circles indicate the absorbing states and blue transient states and the label indicates the number in each of the S,I,R states.

In order to use standard 2-d matrices for computational purposes a mapping from the state space (in $\mathbb{Z}^2$) to $\mathbb{Z}$ is required, the ordering chosen is lexicographic on $\left(R\left(t\right),\ I\left(t\right)\right)$ as this makes the transition matrix upper triangular.

2.3.3 Final Size Distribution

The use of the embedded Markov chain (EMC) to calculate the final size distribution in the GSE has been known since Bailey (1953), but it is frequently overlooked in favour of a triangular set of equations derived by Whittle (1955) which has the advantage that it has been extended to non-Markovian distributions by Ball (1986). However this set of equations, which is suitable for small populations, is numerically unstable for populations of more than about 60. The population size at which the instability appears varies with the parameters. Demiris and O'Neill (2006) used multiple precision arithmetic to overcome these difficulties and made some comparisons of approximate and exact results, they also introduced a new definition of the epidemic threshold. The approach taken here produces identical results but

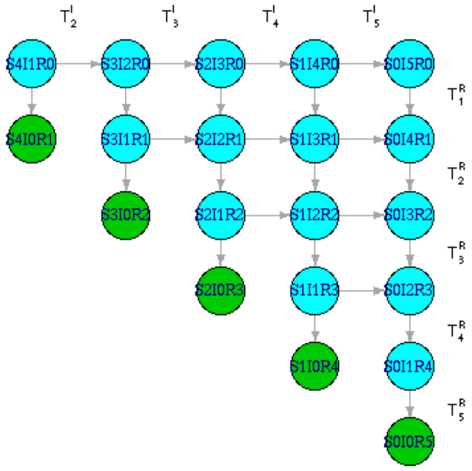


Figure 2.3.1: State space for SIR $n_p = 5$

also allows other distributions, including joint and conditional distributions to be computed.

The transition probability matrix of the EMC for the GSE is denoted G (or g_{jk} for a term where j and k are both $\in \mathbb{Z}^2$) and can be calculated using equation 2.3.1 from the generator matrix of the full process or directly. As there at most two transitions from each state the non-zero values of each row are simply obtained, the recovery term is $\frac{\rho n_I}{\rho n_I + \lambda n_I n_S} = \frac{\rho}{\rho + \lambda n_S}$ for $n_I > 0$ and as $\mathcal{R}_0 = n_p \lambda / \rho$ the terms of the matrix are obtained as

$$g_{(n_I,n_S)(n_I^{'},n_S^{'})} = \begin{cases} \frac{1}{1+n_S\mathcal{R}_0/n_p} & n_S = n_S^{'}, \quad n_I = n_I^{'} - 1, \; n_I > 0 \\ \frac{n_S\mathcal{R}_0}{n_p+n_S\mathcal{R}_0} & n_S = n_S^{'} - 1, \; n_I = n_I^{'} + 1, \;\; n_I > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2.3.3)$$

The maximum number of steps until absorption is $2n_p - 1$, which only occurs when all susceptibles are infected and so the distribution of final states is given by $p_{\text{fsz}} = p_0 G^{2n_p-1}$ where p_0 is the distribution of initial states, usually $(n_p - 1, 1)$. This is straightforward to calculate, using a sparse matrix library, for any value of $\mathcal{R}_0$, a further improvement in computational speed is available by noting that as p_{fsz} is invariant for G then $p_{\text{fsz}} = p_0 G^{2^l}$ for any $l > \log_2(2n_p - 1)$ and G^{2^l} is efficiently calculated by repeated squaring. Examples of p_{fsz} are shown in figure 2.3.2, the left hand plot indicates for a small population how the shape of the distribution changes with $\mathcal{R}_0$ and the bimodal nature of the distribution when $\mathcal{R}_0 > 1$. For larger populations the distinction between “minor” and “major” epidemics becomes more pronounced and the probabilities have a larger range, the right hand plot shows the final size distributions for $n_p = 1000$, with a log scale, and shows that for $\mathcal{R}_0 = 2$ there is a large range of final size values $(100 - 600)$ where the probability is negligible, in fact the cumulative probability only changes from 0.5020668 to 0.5020656 on this range. The curve for $\mathcal{R}_0 = 1.3$ is distinctly bimodal while the second mode on the curve for $\mathcal{R}_0 = 1.1$ is barely discernible. The possible shapes are generally described as “J” or “U” shaped for the unimodal and bimodal distributions, see for example Ball and Nåsell (1994), a further refinement is possible. The unimodal distibution for small $\mathcal{R}_0$ has its maximum at the boundary, corresponding to no further infections. If more than one infective is introduced at the start of the epidemic then the mode can move to a small value, away from the boundary, this is not considered further. The bimodal distributions can be further distinguished into cases where the mode corresponding to a “major” epidemic is at the boundary, corresponding to all susceptibles becoming infected or only a significant proportion. These two cases are described as “U” shaped or “S” shaped (consider the S rotated 90°), the examples on the left of figure 2.3.2 are “J” or “U” shaped and those on the right “S” shaped.

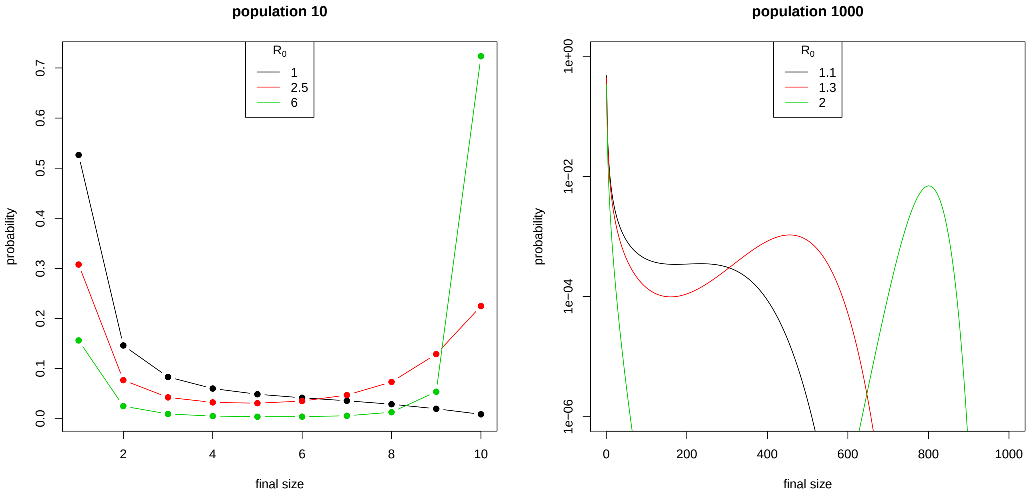


Figure 2.3.2: Final size distributions for the GSE. The left hand plot for $n = 10$ has a linear y-scale, the right hand plot for $n = 1000$ has a log y-scale.

The shape of the final size distributions for the GSE has been determined for

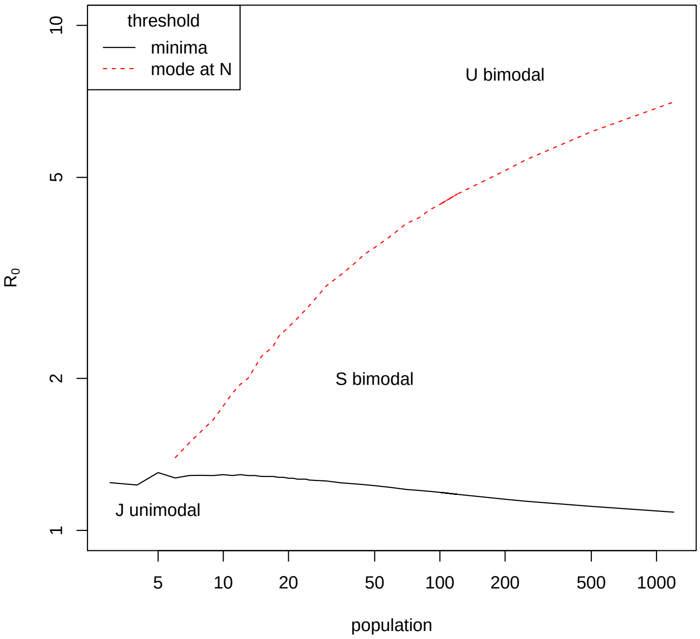


Figure 2.3.3: Shape of final size distributions for the GSE, with lines indicating the boundary between regions.

a range of $\mathcal{R}_0$ and populations n_p the results are plotted in figure 2.3.3. The mean, which can be calculated using the fundamental matrix described above, is a poor measure of a bimodal distribution, however for a uni-variate distribution the means of the two halves are usually good descriptive statistics, when the bi-modality is indistinct the choice of where the split should be can affect the results. For the final size distribution the choice of split is equivalent to defining "major" and "minor" epidemics, which in the case of large $\mathcal{R}_0$ or large n_p is unambiguous. A well known asymptotic result is that as $n_p \to \infty$ the distribution of the final size, conditioned on it being a "major" epidemic, converges to a normal distribution (see Andersson and Britton (2000) section 4.4 for example), of interest is how good the resulting approximation is for finite n_p. An exact study requires a clear understanding of the definition of a "major" epidemic and we therefore study the bi-modality in more detail.

Demiris and O'Neill (2006) introduced a definition of a "major" epidemic based on the first probability less than ε, which they take as 10^{-3} for their example calculations, for small n_p such as shown at the left of figure 2.3.2 ε would need to be changed. A new threshold $\mathcal{M}(\mathcal{R}_0)$ for $\mathcal{R}_0 > 1$ is introduced based on the asymptotic extinction probability $1/\mathcal{R}_0$, a “major” epidemic is defined as one with final size greater than $\mathcal{M}(\mathcal{R}_0)$ which is defined by

$$\mathcal{M}(\mathcal{R}_0) \doteq \left\{ M | \sum_{i=1}^{M-1} p_{\text{fsz}}(i) \leq 1/\mathcal{R}_0 < \sum_{i=1}^{M} p_{\text{fsz}}(i) \right\}. \tag{2.3.4}$$

.

Some examples, including those plotted in figure 2.3.2, of the thresholds are shown in table 2.3.1 together with the overall and conditional means and locations of the turning points, if any. The $\mathcal{M}(\mathcal{R}_0)$ threshold is plotted along with the more obvious local minimum in figure 2.3.4, the main difference is that for small $\mathcal{R}_0$ the local minimum is undefined, for $\mathcal{R}_0 = 1.2$ it only exists for $n_p \geq 74$ and for $\mathcal{R}_0 = 1.1$ it only exists for $n_p \geq 1000$ and no points are plotted. The increasing difference between $\mathcal{M}(\mathcal{R}_0)$ and the local minima as n_p increases is an indicator of the arbitrary nature of the choice of threshold in an area of close to zero probability. This highlights the importance of defining the threshold, especially for small n_p or small $\mathcal{R}_0$, when asymptotic results conditioned on a major outbreak are being used.

n_p	$\mathcal{R}_0$	$\mathcal{M}(\mathcal{R}_0)$	overall mean	mean minor	mean major	local minimum	local maximum
10	1.001	10	2.555	-	-	-	-
10	1.01	9	2.573	2.5039	10.00	-	-
10	1.3	5	3.177	1.8557	7.78	6	7
10	2.5	3	5.409	1.3792	8.42	4	10
10	6.0	2	8.162	1.1383	9.72	4	10
1000	1.1	106	27.883	6.7946	239.63	171	235
1000	1.3	51	94.371	3.7644	396.81	160	455
1000	2.0	18	397.070	1.9419	792.26	326	801

Table 2.3.1: sizes and thresholds for major and minor epidemics

An alternative way of viewing the final size distributions, as n_p and $\mathcal{R}_0$ vary, is to examine $\mathbb{P}\left(\mathfrak{R}_\infty \leq c\right)$ the probability of the final size being less than a threshold c, where c is itself a function of the distribution. As above we consider the three values $\mathcal{M}(\mathcal{R}_0)$ and the local minima and maxima, when they exist.

The figures 2.3.7,2.3.6,2.3.5 show two distinct families of distributions as $\mathcal{R}_0$ varies, with a transition region between them. Figure 2.3.5 shows that for $\mathcal{R}_0 = 2$ and

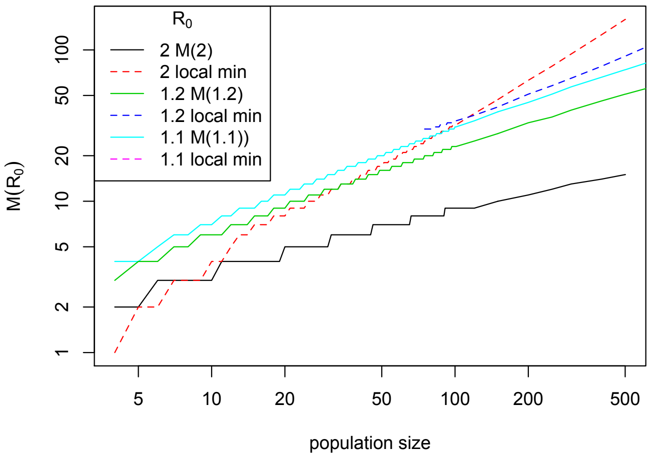


Figure 2.3.4: Epidemic thresholds $\mathcal{M}(\mathcal{R}_0)$ and local minima of final size distribution

4 there is a sensible threshold for a major epidemic, either the local minimum or $\mathcal{M}(\mathcal{R}_0)$, for all values of n_p and the local maxima increases from the asymptotic value as n_p decreases. However in figure 2.3.7 for $\mathcal{R}_0 = 1.25$ we can see that the local maxima and minima approach each other and for $n_p < 75$ the final size distribution is monotonically decreasing. For smaller $\mathcal{R}_0$ this change happens for larger n_p, there is a transitional range for $\mathcal{R}_0$ in $[1.28, 1.3]$ where the critical n_p is small. These critical values of n_p are shown in table 2.3.2.

$\mathcal{R}_0$	1.05	1.10	1.15	1.20	1.25	1.3
n_{crit}	> 1200	≈ 800	≈ 250	75	30	6

Table 2.3.2: Critical population sizes for $\mathcal{R}_0$, below which the final size distribution is "J" shaped.

The main impacts are that for $\mathcal{R}_0 > 1$ and close to 1.

- a very large population is needed to have a clear distinction between major and minor epidemics,
- simulated "major" epidemics will have a wide range of final sizes and dissimilar trajectories,
- care is needed in inference which is conditioned on a "major" epidemic.

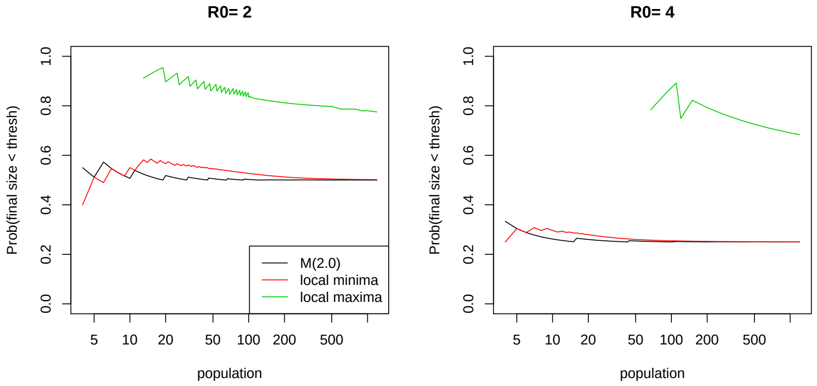


Figure 2.3.5: Final size thresholds, upper range of $\mathcal{R}_0$

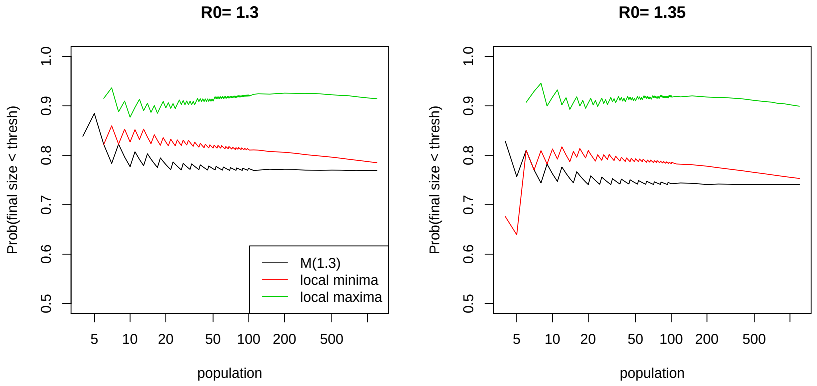


Figure 2.3.6: Final size thresholds, transition range of $\mathcal{R}_0$

Distribution of the Number of Infectives at the First Removal

Although simulated data have a time origin of the first infection, in analysis of real epidemics the time origin is usually taken as the first removal, at which time the number of infectives is unknown. Here the exact distribution of this number $I\left(T_1^R\right)$ is derived, which only depends on $\mathcal{R}_0$ and not directly on the removal rate ρ. Maximum likelihood estimation of this quantity has been considered by Kypraios (2009) using a different approach. The distribution is obtained directly by considering the Markov chain on a reduced state space with only 0 or 1 recoveries and making all the states with 1 recovery absorbing. The resulting continuous time Markov chain can be used to derive the distribution of the time to the first recovery, but this does not appear to have a practical use, except possibly to use as a prior in the Bayesian estimation discussed in chapter 4. The embedded chain provides the structure, which

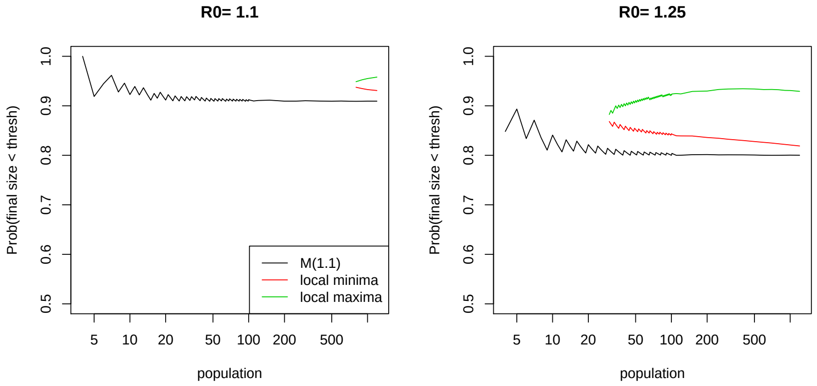


Figure 2.3.7: Final size thresholds, lower range of $\mathcal{R}_0$

can be used directly for calculation, or simplified to give:

Proposition 1. *In the GSE the distribution of the number of infectives immediately after the first removal* $p_l = \mathbb{P}\left(I\left(T_1^R\right) = l\right)$ *is given by*

$$p_l = \begin{cases} \frac{1}{1+\mathcal{R}_0(1-n_p^{-1})} & l = 0 \\ \frac{1}{1+(n_p-l-1)\mathcal{R}_0/n_p} \prod_{i=0}^{l-1}(1-p_i) & l \geq 1 \end{cases}$$

Proof. At each step the probability of removal before infection is $p_{n_S} = \frac{1}{1+n_S\mathcal{R}_0/n_p}$ (see equation 2.3.3) from which the result follows. □

Joint distribution of infectives after first recovery and final size

Potentially of more interest is the joint distribution of final size and the number of infectives immediately after the first removal from which the conditional distributions can be obtained. The calculation of the final size distribution (section 2.3.3) in fact calculates the full transition matrix which is $\mathbb{P}\left(\mathfrak{R}_\infty | I\left(T_1^R\right) = l\right)$, so the joint distribution is obtained from $\mathbb{P}\left(\mathfrak{R}_\infty, I\left(T_1^R\right)\right) = \mathbb{P}\left(I\left(T_1^R\right)\right)\mathbb{P}\left(\mathfrak{R}_\infty | I\left(T_1^R\right)\right)$ and hence the conditional distribution $\mathbb{P}\left(I\left(T_1^R\right) | \mathfrak{R}_\infty = j\right) = \mathbb{P}\left(\mathfrak{R}_\infty = j, I\left(T_1^R\right)\right)/p_{\text{fsz}}(j)$. The number infected at the first recovery and the final size are strongly correlated, as n_p increases $I\left(T_1^I\right)$ becomes a good predictor of final size. Examination of plots of these distributions for a fixed population size does not reveal any great surprises, they have previously been investigated by simulation, examples are shown in 2.3.8, these exact distributions could be of use in inference.

The conditional distribution of the number infected at the first recovery given

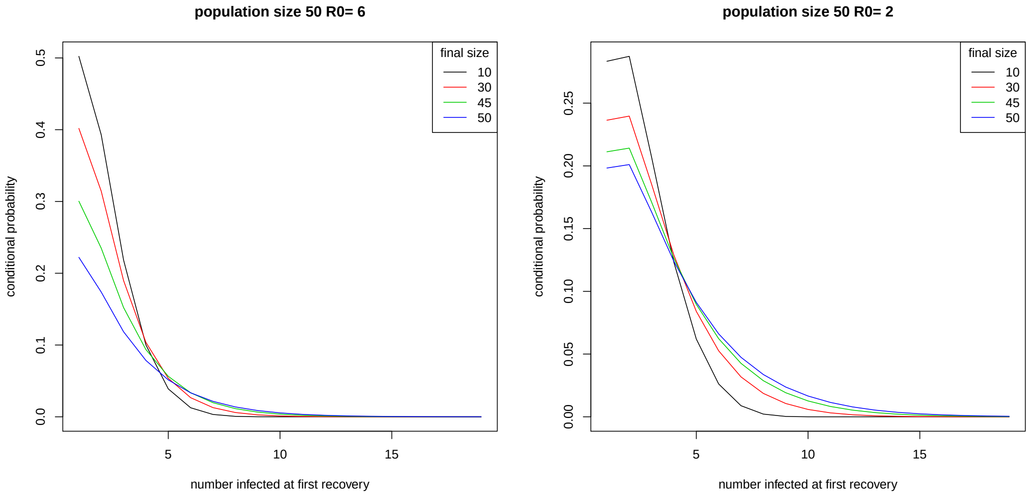


Figure 2.3.8: Conditional distribution of infectives after first recovery given final size

the final proportion of the population infected (this is often called the "attack rate", although it is a ratio not a rate) appears to be independent of n_p for large enough n_p, e.g. for $\mathcal{R}_0 = 2$ for $n_p \geq 100$, however there is a dependence on $\mathcal{R}_0$ which is illustrated by comparison of the two halves of figure 2.3.9.

Joint distribution of maximum number of infectives and final size

Also of interest is the joint distribution of final size and the maximum number of infectives at any time, for some diseases such as influenza this can be of as much interest as the final size, as if the proportion of the population infected at one time is too large, there will be difficulty maintaining essential services. This requires the construction of a Markov chain with an expanded state space. An additional random variable is defined as $W_t = \max_{\tau \leq t} I(t)$ and the expanded state at time t is $X_t = (S(t), I(t), W_t)$ where $X_t \in \mathcal{X} \subset \mathbb{Z}^3$ s.t. $S(t) + I(t) \leq n_p$ and $W_t \geq I(t)$, $\mathcal{X}$ and has size $O(n_p^3)$.

The possible transitions are now:

$$\begin{array}{cccccc} & \text{rate} & \text{from state} & \text{to state} & \text{condition} & \\ S \to I & \lambda S(t) I(t) & (S(t), I(t), W_t) & (S(t)-1,\ I(t)+1, W_t+1) & \text{if } I(t) = W_t & \\ S \to I & \lambda S(t) I(t) & (S(t), I(t), W_t) & (S(t)-1,\ I(t)+1, W_t) & \text{if } I(t) < W_t & . \\ I \to R & \rho I(t) & (S(t), I(t), W_t) & (S(t),\ I(t)-1, W_t). & & \end{array}$$

The generator matrix $\mathbf{Q}_\theta$ is readily computed from either the transition rates

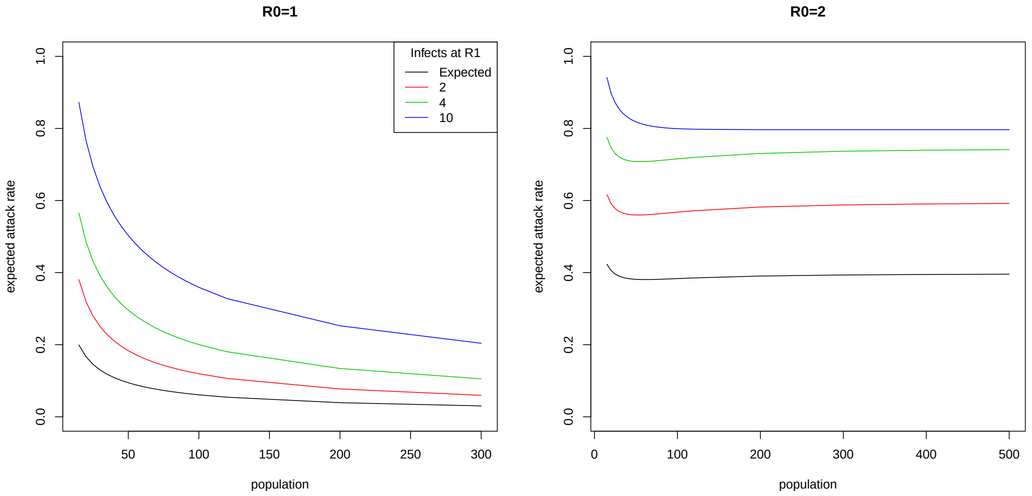


Figure 2.3.9: Expected attack rate conditional on the number infected at the first removal, for $\mathcal{R}_0 = 1, 2$

above or from the generator for the unexpanded model and still contains $O(n_p^2)$ non zero entries as in the unexpanded state.

The marginal distribution of W_t is immediately available and an example for $\mathcal{R}_0 = 2$, $n_p = 100$ is shown in figure 2.3.10, where all the distributions have been conditioned on it being a "major" outbreak based on the local minimum. The conditional distributions are also available and the right hand plot compares the final size distribution only conditioned on it being "major" (i.e. > 31) with two distributions only conditioned on the maximum taking the values 12 or 30 (note the supports differ).

2.3.4 Distribution of the Duration of the GSE

The calculations so far presented have been based on the EMC and have not used the times of transitions of the Markov chain, here the duration of the epidemic is considered. The duration, $T_{\text{durn}} = T_m^R$, where $m = \mathfrak{R}_\infty$ is the final size, is the time until the last removal, which in the Markov representation is the time to absorption in one of the states with no infectives and so is immediately available using the standard results given in section 2.3.1. These results also provide the joint distribution of final size and duration and hence the distribution of duration conditioned on final size. The general formula in lemma 2 applied to the GSE gives the joint probability of final size and the p.d.f. of the duration as in the case of the GSE each of the absorbing states corresponds to a particular final size which we

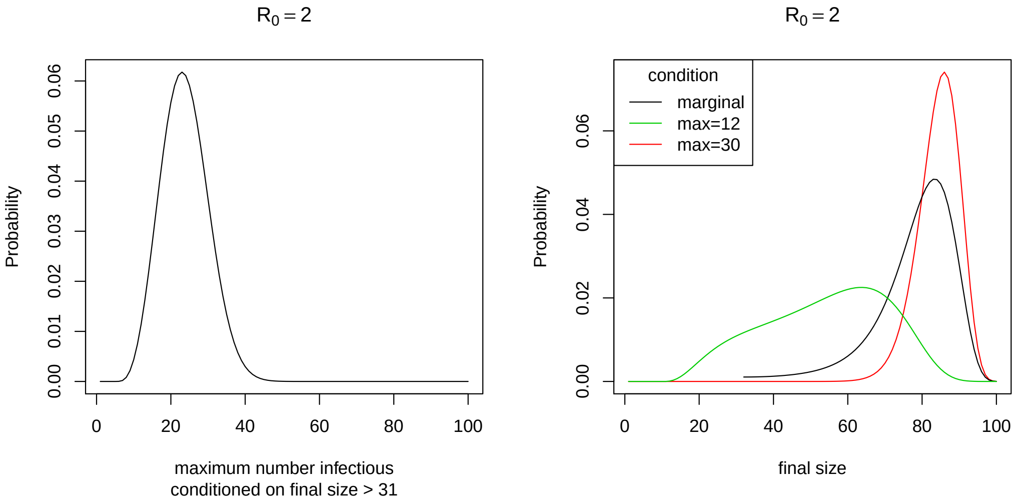


Figure 2.3.10: Distribution of maximum number of infectives and final size conditioned on max for $\mathcal{R}_0 = 2$, $n_p = 100$

identify.

Three examples are examined with $(n_p, \mathcal{R}_0) = (81, 2.025)$,$(81, 1.3)$ and $(120, 1.3)$, the first is the model examined by Barbour (1975) and has been chosen to be "well behaved", with clear local minimum and maximum in the final size distribution, the second has a weaker bi-modality. For each of a set of 50 duration times t, chosen to cover most of the support of the distribution, and all sizes $m = 1 \ldots n_p$ calculate the joint probability/p.d.f. $f_{\text{durn}}(t, m)$ where $f_{\text{durn}}(t, m)dt = \mathbb{P}\left(T_{\text{durn}} \in [t, t + dt) \text{ and } \mathfrak{R}_\infty = m\right)$, contour plots of $\log(f_{\text{durn}}(t, m))$ are shown in the right half of figure 2.3.12, summing over m gives the marginal distribution of duration shown in the left half of 2.3.11 and the conditional distributions from $f_{\text{durn}}(t, m)/p_{\text{fsz}}(m)$ where p_{fsz} is obtained as described previously in section 2.3.3.

The left hand plot of figure 2.3.11 matches figure 2 of Barbour (1975) with a change of timescale, as he uses a recovery rate of 2 or 0.5 whereas 1 has been used here.

Comparison of the results for $n_p = 81, \mathcal{R}_0 = 1.3$ and $n_p = 120, \mathcal{R}_0 = 1.3$ shown in figure 2.3.12 shows the conditional distributions of duration given final sizes of 20 (out of 81) and 30/120 are very similar. This is an example of a general observation, from other examples not shown here, that the distribution of duration conditioned on attack rate is independent of population size for n_p greater than quite small values and that there is little dependence on $\mathcal{R}_0$.

The expected duration $\mathbb{E}(T_{\text{durn}})$ could be calculated by numerical integra-

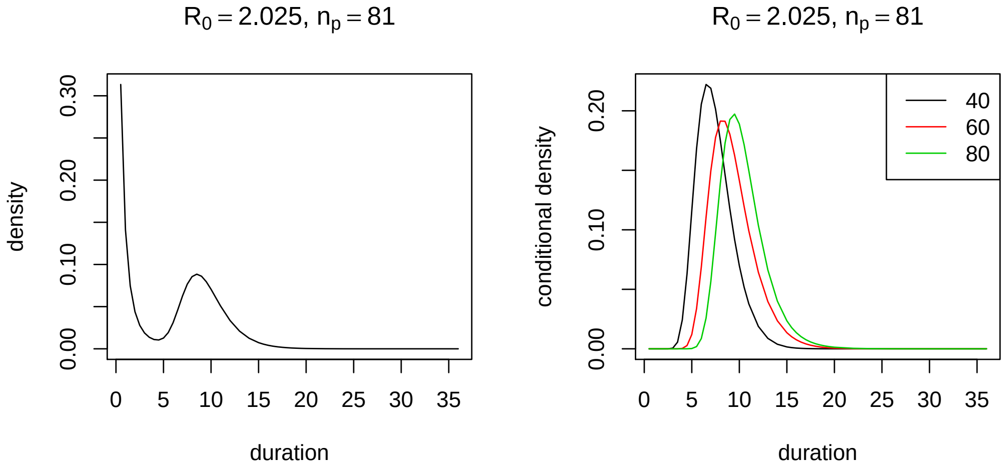


Figure 2.3.11: Distribution of duration, marginal (left), conditional on final size (right) $n_p = 81, \mathcal{R}_0 = 2.025$

tion of the p.d.f. obtained in the previous paragraph but a faster and more direct method is to use the fundamental matrix which gives the mean and variance directly using the formula in equation 5, this has also been suggested by Keeling and Ross (2008). However the bi-modality of the distribution as shown in section 2.3.3 again means that it is not a useful statistic, for that example $\mathbb{E}(T_{\text{durn}}) = 4.917603$, and we are more interested in the mean conditioned on a "major" outbreak. The calculation using the fundamental matrix also gives $\mathbb{E}(T_{\text{durn}})$ starting from each transient state, from which the expected duration of example epidemics with different initial numbers of infectives is obtained, for this example it increases up to 5 initial infectives, then decreases, this is because the probability of a "major" outbreak increases rapidly with the number of infectives but once 5 is reached the "major" outbreak is almost certain and increasing the number of initial infectives removes the time needed to infect them. More useful expectations can also be obtained, for example the expected remaining duration after the nth removal, ignoring any known removal times before it. The distribution of number of infectives at the nth removal is available from the EMC, which is used as initial distribution over states, and then calculated from the same fundamental matrix. An alternative approach is to condition on the final size using a technique described by Kemeny and Snell (1976) p64, conceptually a modified Markov chain is constructed with just one absorbing state, they show how the fundamental matrix of the modified chain is obtained from that of the original chain.

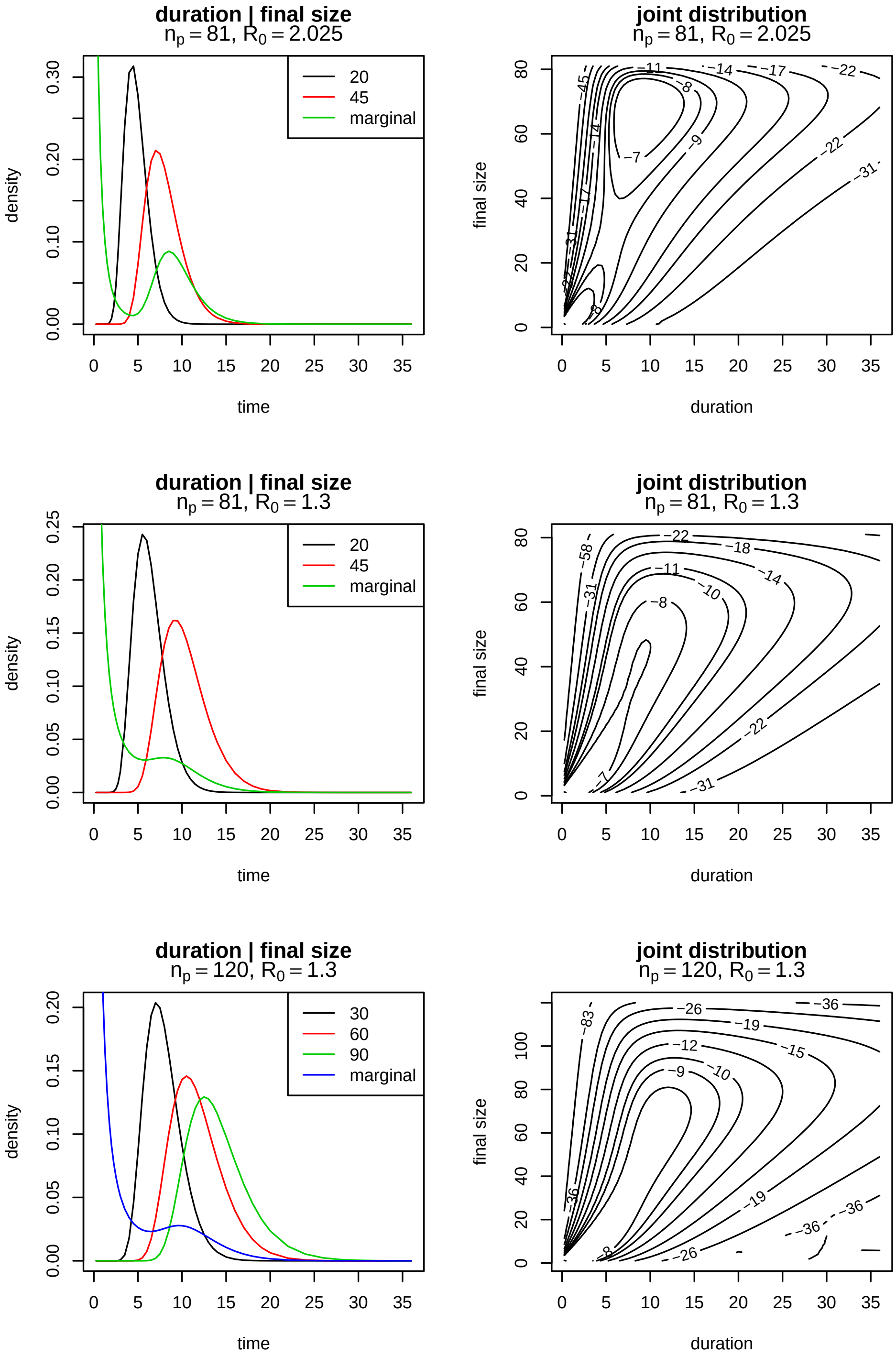


Figure 2.3.12: Joint and conditional distribution of duration and final size $n_p = 81, \mathcal{R}_0 = 2.025, 1.3$ and $n_p = 120, \mathcal{R}_0 = 1.3$

I0	1	2	3	4	5	6	7	8	81
$\mathbb{E}(T_{\text{durn}})$	4.92	6.88	7.64	7.90	7.93	7.86	7.75	7.63	4.98

Table 2.3.3: Expected duration of example epidemics with different initial numbers of infectives

Barbour has shown that the final size and duration are asymptotically independent and his formula for the distribution of the duration, based on asymptotics, is good even for moderate values of n_p. An example of the departure from independence is shown in the contour plots in figure 2.3.12. The main advantage of using the exact Markov representation is to investigate conditioning on other events, an example question is: given we observe $R(t_c)$ at t_c, what is the distribution of remaining duration and final size ?

2.3.5 Martingale Plots of the GSE Trajectory

The visualisation of possible trajectories of the GSE through the state space is useful for understanding the variability of the GSE and comparison of other models with the GSE. A plot based on a well known martingale is introduced which assists in interpretation of variability and is used below to illustrate some features of the bipartite graph epidemic models.

In Becker and Hasofer (1997) several martingales derived from the GSE are used to develop estimators of the parameters. One of these $M(t) = S(t)(1 + \mathcal{R}_0/n_p)^{R(t)}$ can be used to study the trajectory, the EMC provides a simple proof that $\{M(t) : t \geq 0\}$ is a martingale.

Theorem 2. *In the GSE* $\{M(t) = S(t)(1 + \mathcal{R}_0/n_p)^{R(t)} : t \geq 0\}$ *is a martingale.*

Proof. Consider the sequence $M_j = M(T_j)$ where T_j is a transition time, infection or removal, the two possible changes from the state (s, i, r) with $M_j = s(1+\mathcal{R}_0/n_p)^r$ and their probabilities are:

	next state	M_{j+1}	probability	$M_{j+1} - M_j$
infection	$(s-1, i+1, r)$	$(s-1)(1+\mathcal{R}_0/n_p)^r$	$\frac{s\mathcal{R}_0/n_p}{1+s\mathcal{R}_0/n_p}$	$-(1+\mathcal{R}_0/n_p)^r$
removal	$(s, i-1, r+1)$	$s(1+\mathcal{R}_0/n_p)^{r+1}$	$\frac{1}{1+s\mathcal{R}_0/n_p}$	$s\mathcal{R}_0/n_p(1+\mathcal{R}_0/n_p)^r$

and so $\mathbb{E}(M_{j+1}) = M_j$ and $\mathbb{E}(|M_{j+1}|) < \infty$ and as $M(t)$ is constant between M_j for each j then $\{M(t) : t \geq 0\}$ is a martingale. □

On simulated data from the GSE it is straightforward to calculate M_j and plot against the number of removals, this is done in figure 2.3.13 for a set of 1000 major

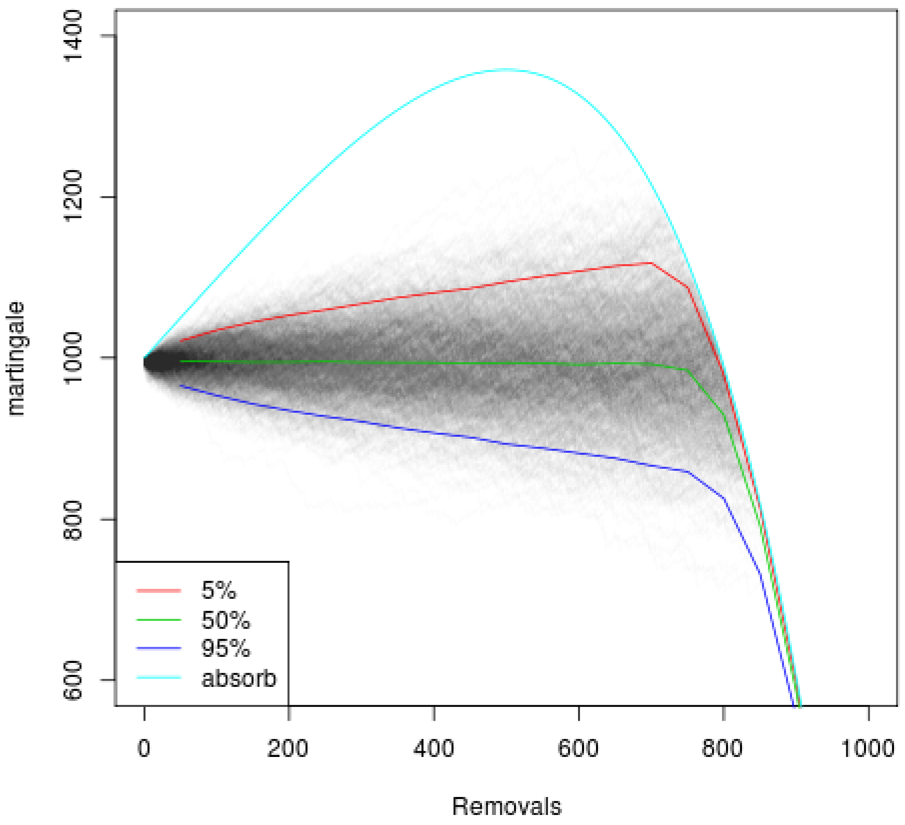


Figure 2.3.13: Martingale plots of 1000 simulated GSE $\mathcal{R}_0 = 2$, a shaded density highlights the most likely trajectories, single trajectories are not all visible. Exact quantiles of the distrbution of M_j and the absorbing boundary at zero infectives are included.

epidemics simulated on a population of 1000 with $\mathcal{R}_0 = 2$; also shown are the absorbing boundary with zero infectives and the median and two quantiles calculated using equation 2.3.3. For most of the epidemic the curves are as might be expected: fairly close to a horizontal line, however at the right hand side when close to the absorbing boundary the probability is close to 1 of a removal with a resulting small increase in M_j and there is a small probability of a large decrease in M_j. It should be noted that if an estimated $\mathcal{R}_0$ is used then M is no longer a martingale, in particular if we use the estimator (see chapter 4) based on the same martingale equation, the end point is fixed as a function of final size. It would be possible to derive a goodness of fit test from the calculated bounds, however this would require full data which are rarely available. Similar plots are used below to portray the difference between some models incorporating heterogeneity and the GSE.

2.4 Discrete Time Epidemic Models

Usually complete data are unavailable and the only data available are regular, i.e. daily, counts of those *R*emoved. The observation interval is often one day, sometimes more and rarely less. There are two approaches to modelling these data, one is to consider a continuous time model with regular observation, the other is to use a discrete time stochastic process. The first approach is typified by considering the GSE with daily counts of those *R*emoved, so the observations are $Y_t = n_p - S(t) - I(t)$ at $t = j\Delta_t$ for $j = 1, 2 \ldots T$ where Δ_t is the fixed observation interval. We can consider this as a Hidden Markov model (HMM) where the observation is a deterministic many to one mapping $\mathcal{X} \mapsto 0 \ldots n_p$ and the transition probability matrix $P_\theta = e^{\Delta_t \mathbf{Q}_\theta}$. This approach is considered further in chapter 4.

The differences between this and more usually considered HMM are that the process is acyclic, and the transition matrix is sparse. Other characteristics are that the observations have limited information about θ and that calculating the exact transition matrix $P_\theta = e^{\Delta_t \mathbf{Q}_\theta}$ is slow. Also a significant feature of the SIR epidemic is that at all stages, particularly the very early and late stages there is a significant chance that the epidemic dies out. In terms of the HMM it enters an absorbing state, although there may be further data indicating this is wrong, or in terms of the support of the posterior, states with zero infectives are excluded, except the final state.

Discrete time models also have a long history, the Reed-Frost model was developed in 1928 and the Greenwood in 1931 (see for example sect §3.8.1 of Allen (2008)). These models work in terms of generations of infection rather than calendar time and are appropriate for modelling diseases with either long latent periods, so that generations of infection can be identified, or in very small populations such as households. They can also be used to obtain useful analytic results e.g. Scalia-Tomba (1985), collectively these and other variants are called chain binomial models. Becker (1989) gives a description of this class of model for generations as: with probability q_i an individual escapes infection when there are i infectives, the Reed-Frost model has $q_i = q^i$, and the Greenwood model $q_i = q$ for $i > 0$ and a single parameter $q \in (0, 1)$. The number infected each generation is $N^I_{k+1} = N^S_k - N^S_{k+1}$ where $N^S_{k+1} | N^S_k$ is binomially distributed as

$$\Pr(N^S_{k+1}|N^S_k = s, N^I_k = i) \sim \text{bin}(s, q_i), \tag{2.4.1}$$

the epidemic stops the first time $N^I_{k+1} = 0$.

In the Reed-Frost model the parameter $p = 1 - q$ is interpreted as the probability that a given infective individual infects any given susceptible individual during the former's infectious period.

2.4.1 A Binomial Model for Regularly Observed Epidemics

An alternative binomial model is described here which could be considered as an approximation to the regularly observed continuous time GSE or considered as a distinct model. Again this is a Markov chain on the same state space as the GSE, and in the layout of figure 2.2.1 motion is again only to the right or down but whereas the GSE only moves 1 step, here the size of moves is only limited by the boundary of the state space. Analytic asymptotic results on this model are harder to obtain than in the continuous time GSE, however computation of the likelihood is significantly easier. Standard HMM techniques for inference could be applied to this model or it could be used as a proposal distribution in MCMC for the regularly observed GSE.

At each step k of the chain the number of new infections V_k and removals W_k of current infectives are independently binomial distributed with parameters p_R and $p_I(.)$ as

$$V_k \sim \text{binomial}(S\,(k)\,, p_I(I\,(k)))$$

and

$$W_k \sim \text{binomial}(I\,(k)\,, p_R).$$

The next state of the Markov chain is then determined by $S\,(k+1) = S\,(k) - V_k$ and $I\,(k+1) = I\,(k) + V_k - W_k$. The parameter p_I is dependent on the number of infectives and the obvious approach is to link them to the parameters of the GSE by $p_R = 1 - \exp(-\rho)$ and $p_I(\nu) = 1 - \exp(-\lambda\nu)$, this model differs from the GSE in that recoveries can not occur on the same day as infection, which in many situations will be more realistic. This model can now be compared to the related regularly observed continuous time model. The distributions of V_k and W_k are compared with the equivalent exact marginal distributions of the GSE, in figure 2.4.1. The example chosen has $\mathcal{R}_0 = 3, \Delta_t = 0.5$, the distributions are plotted for transitions from two states, one near the beginning, one in the middle. Examination of those and other plots suggests that the marginal distributions are close except when $N^I_t = 1$ or 2, the differences are mainly that the binomial model ignores the correlation present in the GSE and in the early stages newly infected individuals don't contribute to the infectious pressure until the next day.

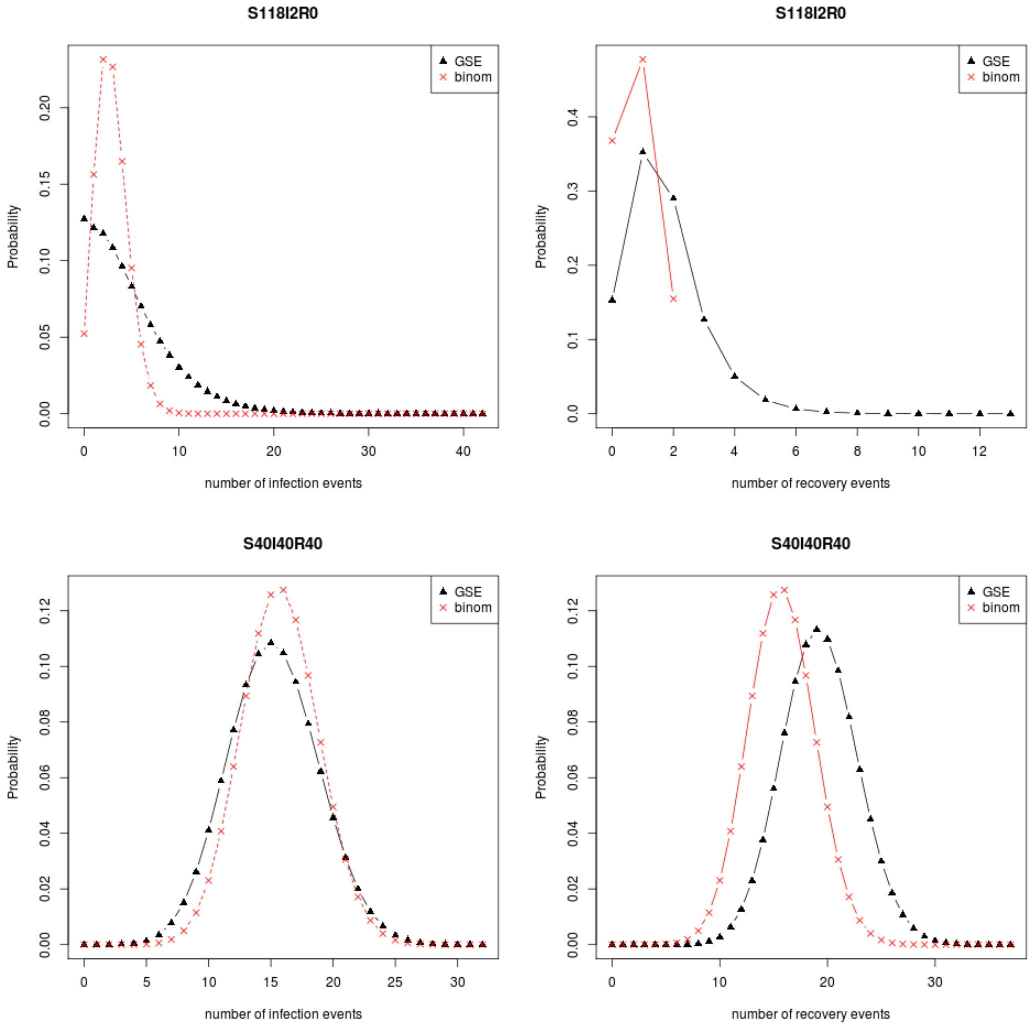


Figure 2.4.1: Comparison of GSE and binomial models, $n_p = 120$. The upper pair are at the start of the epidemic S=118,I=2,R=0. The lower pair are in the middle at S=40,I=40,R=40. $\mathcal{R}_0 = 3,\ \Delta t = 0.5$.

2.5 Models for Epidemics Incorporating Heterogeneity

The crucial term in the SIR model is the non-linear term for the infection process $\lambda S(t) I(t)$ which incorporates the least justifiable assumptions, that of homogeneous mixing of the population, equal infectiousness and equal susceptibility. Much work has been done developing models that incorporate different aspects of the true heterogeneity. A range of observed or unobserved co-variates for individuals, in particular age or location, can have significant effects on some or all of contact patterns, infectiousness, susceptibility and durations of phases, some aspects of some models for heterogeneity are considered in this section.

Much work has been done developing models that incorporate different aspects of the true heterogeneity, many established results are in books such as Mollison (1995) which includes sections on heterogeneity and grouped populations. Models which allow for variation in either infectiousness or susceptibility are common, for instance (Becker and Yip, 1989) point out that "variation in susceptibility of individuals can give the impression that the infection rate is declining over time, because highly susceptible individuals tend to be infected earlier" or it could give rise to other incorrect assumptions. Another approach is that of Severo (1969b) who introduces a heuristic model where the term $\lambda S\,(t)\,I\,(t)$ is replaced by $\lambda(S\,(t))^{1-b}(I\,(t))^{a}/N$ with $a > 0$ and $b < 1$.

Different aspects of heterogeneity will provide departures from the homogeneous model in a variety of ways which may operate at different timescales. The effects may manifest themselves as increased variance or as a mixture of epidemics with different time origins. Consideration must also be given to whether they should be modelled as random effects or unknown parameters.

The most general model incorporating heterogeneity in infectivity allows a different infection rate $\lambda_{i,j}$ between each pair of individuals, denoting the state of individual j at time t as $X_{j,t} \in \{\mathsf{S}, \mathsf{I}, \mathsf{R}\}$ for each j

$$\begin{aligned} \mathbb{P}(X_{j,t+dt} = \mathsf{I} \,|\, X_{j,t} = \mathsf{S}) &= \sum \lambda_{i,j} \mathbf{1}(X_{i,t} = \mathsf{I}) dt \qquad (2.5.1) \\ \mathbb{P}(X_{j,t+dt} = \mathsf{R} \,|\, X_{j,t} = \mathsf{I}) &= \rho dt \end{aligned}$$

the resulting population rate of infection is $\sum_{i \in S} \sum_{j \in I} \lambda_{i,j}$. With $O(N^2)$ parameters and $O(N)$ data points, clearly this is not identifiable. A wide variety of models have been proposed that assume a plausible structure for the infectious contact rates. The unsolved problem from both a practical and a theoretical perspective is how to choose an appropriate model and measure if it is a plausible fit for observed data in a principled manner. Co-variate data can be modelled by using a linear or log linear model for the $\lambda_{i,j}$ in equation 2.5.1, an example of such a model is described in Jewell and Roberts (2012). Imposing more structure enables a deeper analysis and the most widely studied classes of models are briefly described.

2.5.1 Household Epidemic Models

For most human diseases there is a much greater chance of infection within a household than outside, and data are often available at this level. A large body of work has studied both inference and asymptotic results, a recent example is Ball et al. (2010a) which combines variable infectiousness with a household model and shows that it is possible to discriminate between the models by comparing the Kullback-Leibler divergence for the fitted models to data. The usual formulation is that each individual is subject to two independent sources of infection, global and within the house. The overall infection rate on a susceptible individual is $\lambda_G I_G(t) + \eta(I_H(t))$ where $I_G(t)$ is the global number of infected individuals, $I_H(t)$ the number infected within the same household, λ_G the global infection rate and $\eta(n)$ a function for the within household infection rate when n are infected. The usual form is either $\eta(n) = n\lambda_H$ or sometimes $\eta(n) = \mathbf{1}[n > 0]\lambda_H$.

2.5.2 Spatial models

On a large scale the progression of many diseases is dominated by the spatial aspects and so the full range of spatial analysis techniques can often be applied. An example is the analysis of the 2001 foot and mouth epidemic in the UK (Diggle, 2006) using partial likelihoods for the spatio-temporal spread of the disease. For human diseases in industrialised countries and some diseases of intensively farmed animals, the variation in the contact process is often non-Euclidean and alternatives include network models which are described below.

2.5.3 Multitype population models

Populations can frequently be partitioned into groups based on location such as town or school or on categorical co-variates such as sex or age group e.g. preschool, school age, adult. A commonly used model assumes homogeneous mixing within a group and different infectiousness within and between groups. The basic model is that if the population in group j is n_j and counts of individuals in the two states Susceptible and Infective at time t are denoted by $(S(j,t),\ I(j,t))$ infection occurs within that group at rate $\sum_i \psi_{i,j} S(j,t) I(j,t)$ and $\psi_{i,j}$ is the rate of infection between an infective in group i and a susceptible in group j, usually the matrix $\psi_{i,j}$ is assumed symmetric. These are sometimes called meta-population models. An example of statistical inference on a simple version of this model is that of Becker

(1989, chapter 5) who analyses an epidemic of a respiratory disease on Tristan da Cunha and shows that from final size information alone it is possible to identify a higher rate of transmission in school children. These models are important as outbreak control measures are often based around structures within populations (e.g. school closures).

The difference from a homogeneous model depends on the infection rate, at high infection rates the epidemic rapidly spreads to all groups and then behaves similarly to a heterogeneous model. While at lower infection rates the epidemic may be delayed or absent in some groups, so making the probability of multiple modes in the distributions of summary statistics of the epidemic.

2.5.4 Epidemics on Networks or Graphs

In several animal diseases a contact network[4] based on known animal movements between farms and other potential infectious contacts is known. An analysis based on these networks is possible, examples include Jewell et al. (2008, 2009a); Jonkers et al. (2010). Human movement and contact patterns are considerably more complex and difficult to identify or model, further consideration of human contact networks is given in section 2.5.5. Standard homogeneous models including SIR, SEIR, SIS and Reed-Frost have all been extended by using a graph (or network) to model the allowable infection paths, we continue to concentrate on the SIR model. Epidemic models are constructed on a graph $G = (V, E)$ where V is the set of n_p individuals, labelled w.l.o.g. with integers e.g. i, j, and E the set of edges indicates that an infectious contact is possible. This could be considered as a special case of equation 2.5.1 where $\lambda_{i,j} = 0$ if there is no edge between i and j, the case considered most frequently of undirected, unweighted graphs corresponds to $\lambda_{i,j} = \lambda > 0$ if there is an edge between i and j. The SIR epidemic on a graph can be simply defined via the adjacency matrix A, $A = a_{i,j}$ $(n_p \times n_p)$ where $a_{i,j} = 1$ if there is an edge between i and j else 0. Denote the state of individual j at time t as$X_{j,t} \in \{\mathsf{S}, \mathsf{I}, \mathsf{R}\}$ and the set of individuals in each state at t as $\mathcal{S}(t), \mathcal{I}(t)$ and $\mathcal{R}(t)$ where $|\mathcal{S}(t) \cup \mathcal{I}(t) \cup \mathcal{R}(t)| = n_p$. Then for each j we have

$$\mathbb{P}(X_{j,t+dt} = \mathsf{I} \,|\, X_{j,t} = \mathsf{S}) = \lambda \sum a_{i,j} \mathbf{1}(X_{i,t} = I) dt \qquad (2.5.2)$$

$$\mathbb{P}(X_{j,t+dt} = \mathsf{R} \,|\, X_{j,t} = \mathsf{I}) = \rho dt \; . \qquad (2.5.3)$$

Recent interest in models for random networks in other contexts has gener-

[4] We use the terms graph and network interchangeably, mainly using the former when discussing mathematical aspects and the latter for actual examples.

ated considerable interest in the study of epidemics on random graphs. Results typically investigate the existence of and conditions for epidemic thresholds on graphs chosen from a distribution, usually uniform, over a family of graphs. These models can be considered as random effects models where the $\lambda_{i,j}$ in equation 2.5.1 are chosen from a binary distribution, for example the Erdös-Rényi random graph corresponds to choosing $\lambda_{i,j} = \lambda_{j,i} = \lambda$ with probability p, otherwise 0. A recent summary of models for random graphs in the context of epidemics is given in section 2.7 of Danon et al. (2011). Much of the recent literature considers the effect of degree distribution and clustering of the graph on the epidemic, but as pointed out by Eubank et al. (2004)

> Both degree distribution and clustering are relevant to short-term propagation in a network, but longer time dynamics will be driven by global graph properties. It is thus natural to consider estimation schemes for global topological measures, such as expansion. Informally, the higher the expansion, the quicker is the spread of any phenomenon (such as disease, gossip or data).

One approach to modelling both local and large scale structure in graphs, that provides the option of incorporating relevant geography is the geographical threshold graph Masuda et al. (2005). The bipartite approach taken below also naturally incorporates both large and small scale features.

Recently several significant results have been obtained for the evolution of epidemics on graphs chosen randomly from particular distributions on a subset of all possible graphs. The results of Volz proved by Decreusefond et al. (2010) and of Ball et al. (2010b) are particularly significant while that of Ball et al. (2014) is considered in section 2.6.

2.5.5 Data on Human Contact Networks

Obtaining accurate data on human contact networks is difficult because of several fundamental problems of definition of a contact and appropriate sampling, these are considered in more depth in Danon et al. (2011). There is still a gap between theory and application, two approaches to obtaining contact data relevant to the spread of human epidemics are described below.

POLYMOD

The POLYMOD study Mossong et al. (2008) studied the social interactions within 8 European countries of 7,290 participants, recording characteristics of 97,904 contacts, including age, sex, location, duration, frequency, and occurrence of physical contact. They found that mixing patterns and contact characteristics were remarkably similar across different European countries, (reproduced in figure 2.5.1). The POLYMOD matrices have been widely used as a significant improvement on the assumption of homogeneous mixing, for example in Medlock and Galvani (2009) the cost-effectiveness of vaccination of different age-classes during the H1N1 pandemic in 2009 is studied. However the usual approach is to use the matrices directly as if true, a more flexible approach regarding the data as a sample from an underlying non-parametric structure such as Gaussian random field could be usefully investigated.

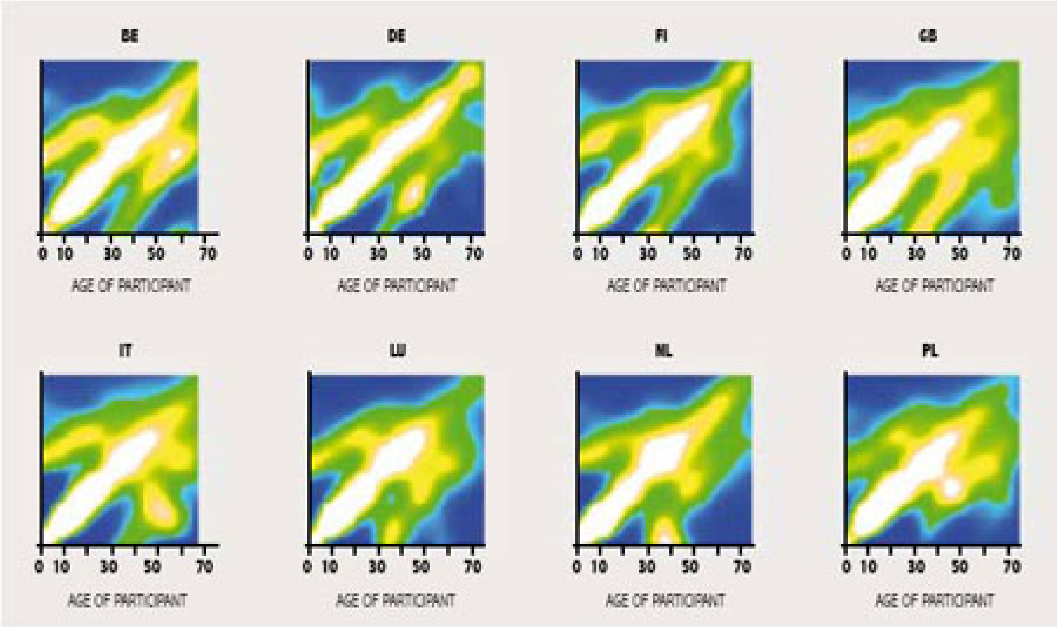


Figure 2.5.1: POLYMOD contact rates by age and country. Each plot is for one country and shows the reported contact rates between pairs of individuals of different ages. White is high, blue low, the age of the participant is on the x-axis and age of reported contact on the y-axis.

EpiSims

The EpiSims project (Eubank et al., 2004) has developed a large scale simulation of epidemic spread in a metropolitan area, the example used in their paper is a hypothetical smallpox outbreak in Portland, Oregon. Using census and transportation data a dynamic bipartite contact graph is constructed for 1.5 million people and 180,000 locations, which is a synthesis of realistic models of disease propagation, human behaviour and available data. They use the bipartite graph to construct a weighted graph which is used within the simulation to provide contact rates. Although they provide their derived contact graph for 1,515,271 individuals, they do not provide the underlying bipartite graph. The marginal distributions of the bipartite graph are given in figures 2a and 2b of the reference which are similar to those shown in figure 2.7.7.

2.6 Epidemic Models on Bipartite Graphs

Epidemics on bipartite graphs have received much less attention than epidemics on graphs. This section shows that a wide variety of other epidemic models can be formulated as bipartite graph models and so provides a unifying framework for comparison of models.

2.6.1 Bipartite Graphs

An alternative model for the possible contacts between individuals is to consider a set of locations at which individuals mix homogeneously. The locations may be of different types such as schools, households, work places. Individuals visit one or more of these locations. This can be represented by a bipartite graph such as that shown in the upper half of figure 2.6.1 where the upper nodes represent locations and the lower nodes individuals. A link indicates visits or association with the location and so a potential contact. A range of other epidemic models for heterogeneity can be considered as bipartite graph models, and some are presented below, these alternative representation are largely of use in understanding the relation between models. Software for bipartite graph epidemics can also be used on these representations.

A bipartite graph $G = (U, V, E)$ consists of two disjoint sets U and V comprising the nodes or vertices[5], and a set of edges E where each edge is a pair of nodes (u, v), $u \in U, v \in V$. Both directed and undirected bipartite graphs can be studied, here only undirected graphs are considered. In general the two sets U and V can be of the same type so that all vertices in U and V are in some larger set of vertices, here U and V are distinct with U representing individuals who may become infected and V an abstract set of possible contacts, which may include physical premises such

[5] the terms are used interchangeably

as schools, houses or work places and can include a temporal aspect. A convenient representation is the adjacency matrix $\mathbf{A} = (a_{ij}, i \in U,\, j \in V)$ where $a_{ij} = 1$ if and only if $(i, j) \in E$ and $a_{ij} = 0$ otherwise.

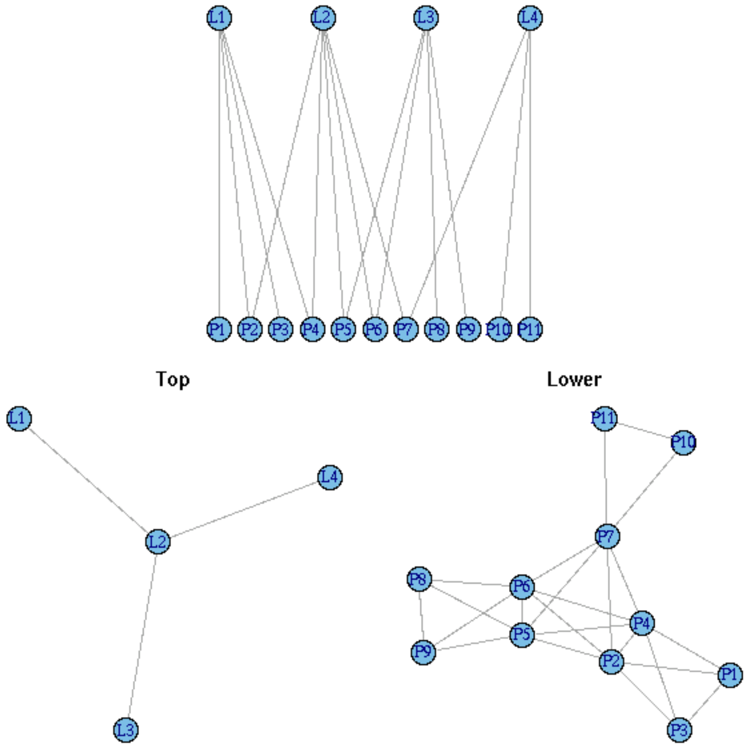


Figure 2.6.1: Example bipartite graph (Newman)

Two projections from a bipartite graph $G = (U, V, E)$ to two unipartite graphs $G_U = (U, E_U)$ and $G_V = (V, E_V)$ are possible (see Latapy et al. (2008) for details), they correspond to paths of length 2 in the bipartite graph. A small example bipartite graph, which is taken from Newman (2003), is shown in the upper part of figure 2.6.1 (where $U = \{P1, P2, \ldots, P11\}$ and $V = \{L1, L2, L3, L4\}$) and the two projections are shown in the lower part. The upper projection, shown on the left, corresponds to connections between places, the lower projection to connections between people i.e. the standard graph widely considered and described briefly in section 2.5.4. The adjacency matrices of the upper and lower projections are obtained from the adjacency matrix of the bipartite graph $\mathbf{A}$ from $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ by replacing the diagonals with 0 and entries ≥ 1 with 1. Properties of the two projections may provide information that can provide insight into epidemics on the graphs for example their diameters provide information on how fast an epidemic might spread.

The following result links the two diameters :

Theorem 3. *The diameters of the upper and lower projections of a connected bipartite graph differ by at most 1.*

Proof. For any connected graph G with adjacency matrix $\mathbf{A}$ and diameter d, G^d is a complete graph. As $A_{low}^{n+1} = AA_{top}^{n}A^T$ and all $a_{ij} > 0$, where A_{low} is the adjacency matrix of the lower projection and A_{top} that of the upper projection.

□

Graph Decomposition

Any graph can be decomposed into cliques which corresponds to a bipartite representation, however the decomposition is not unique and determining an optimal representation, with the minimal number of columns, is known to be a hard problem (NP complete [6]), however much research has been done into finding approximations to the minima, for example Barber (2008). Except in special cases, such as where the graph has been obtained from the projection of a bipartite graph, this mapping from a contact graph to a bipartite graph is unlikely to be useful.

A combined spatial bipartite graph epidemic model The importance of spatial aspects of the contact process has been mentioned above, one approach to the study of spatial epidemics is to study epidemics on regular grids, however a regular grid lacks both the heterogeneity and strong local clustering that real contact processes exhibit. A flexible model based on the bipartite graph epidemic, which combines aspects of a spatial model with more local contacts is readily constructed. By taking a finite grid, either rectangular or triangular and placing a number of individuals at each node and one or more individuals in both groups along each edge, a flexible model can be constructed. Simulations of some examples of these models show interesting properties, including final size distributions that are close to uniform combined with epidemic curves that are similar to those from the GSE. Further investigation of their properties is planned.

.

2.6.2 Random Bipartite Graphs

Recently many models for unipartite graphs have been proposed, that give rise to distributions over subsets of the space of possible graphs, see for example Durrett

[6] `http://en.wikipedia.org/wiki/Clique_cover`

(2007). In a similar way a variety of models can be proposed for bipartite graphs that give rise distributions over subsets of the space of possible graphs, those that relate to epidemics are described below. The criteria for choice will depend on the purpose of the study but analytic tractability is often the principal concern.

Random Intersection Graph

The simplest model is the random intersection graph (RIG), which is the lower projection of an Erdös-Renyi bipartite graph in which each link occurs independently with probability r. Two examples of their use in studying epidemics are Britton et al. (2008) who describes the RIG as:

> Random intersection graphs were introduced in Singer (1995)[7] and Karonski et al. (1999). In its simplest form, the model is defined as follows: Given a set V of n vertices and a set A of m auxiliary vertices, construct a bipartite graph $B_{n,m,r}$ by letting each edge between vertices $v \in V$ and $a \in A$ exist independently with probability r. The random intersection graph $G_{n,m,r}$ with vertex set V is obtained by connecting two vertices $v, w \in V$ if and only if there is a vertex $a \in A$ such that a is linked to both v and w in $B_{n,m,r}$.

A recent paper by Ball et al. (2014) derives expressions for a threshold parameter R_* in a class of RIG, so that in a large population an epidemic with few initial infectives can give rise to a large outbreak if and only if $R_* > 1$ and shows that a law of large numbers can be derived. The extension to other distributions of graphs would not be straightforward. Their model differs from our model based on the same bipartite graph, they have a constant infection rate for individuals linked by one or more edges, whereas in our model the infection rate is greater for individuals with more contacts.

Fixed margin random graphs In the context of the analysis of contingency tables the distribution and simulation of binary matrices with fixed marginals has been investigated by Besag and Clifford (1989) and Chen et al. (2005). These are the bipartite equivalent of the widely studied “configuration network” for unipartite graphs.

[7] see Britton for these references

2.6.3 Epidemics on Bipartite Graphs

The extensions of standard epidemic models, both discrete time and continuous time, with homogeneous mixing to the bipartite network are straightforward to define and simulate. Here the focus is again on the continuous time SIR model, the extensions to other models are obtained in a similar way.

Two approaches to defining the infection rate in continuous time models are possible, a single infection rate could apply to all pairs of individuals connected through one or more locations, this approach is used in the papers described in section 2.6.2. We use an alternative that has an increased infection rate between pairs of individuals that have more than one class[8] in common, as well as reflecting reality, that increased potential routes of infection are likely to give an increased infection rate, it permits the representations of other models in this framework as described in section 2.6.4, subsequently any reference to an epidemic on a bipartite graph should be taken to be of this form.

Epidemic Thresholds on Bipartite Graphs

Newman (2003) has considered epidemics on a class of random bipartite networks using results from percolation theory to derive asymptotic thresholds, he asserts that the position of the epidemic threshold decreases with increasing clustering. The implication that this applies to all graphs is unproven. Britton et al. (2008) have considered a Reed-Frost model in a similar way saying:

> The approximation gives rise to expressions for the epidemic threshold and the probability of a large outbreak in the epidemic. It is investigated how these quantities varies with the clustering in the graph and it turns out for instance that, as the clustering increases, the epidemic threshold decreases.

The interpretation of existing results on epidemic thresholds on graphs requires care as the result combines both the probability of selecting a graph with the probability of an epidemic on that graph.

Definition of a Bipartite Graph Epidemic

A bipartite graph epidemic on a bipartite graph $G = (U, V, E)$ with adjacency matrix $\mathbf{A} = (a_{ij}, i \in U, j \in V)$ is defined as follows. Denote the state of individual $j \in U$ at time t as $X_{j,t} \in \{\mathsf{S}, \mathsf{I}, \mathsf{R}\}$ and the set of individuals in each state at t

[8] the terms class, group and column are used interchangeably

as $\mathcal{S}(t), \mathcal{I}(t)$ and $\mathcal{R}(t)$[9] where $\mathcal{S}(t) \cup \mathcal{I}(t) \cup \mathcal{R}(t) = U$. Each class $k \in V$ has an associated infection rate $\lambda_k \geq 0$, possibly constant across classes or drawn from a specified prior distribution such as a gamma. Then for each susceptible individual $j \in \mathcal{S}(t)$ the rate of infections at time t is

$$\eta_j(t) = \sum_{k \in V} a_{jk} \lambda_k I_k(t) \tag{2.6.1}$$

where

$$I_k(t) = \sum_{l \in U} a_{lk} \mathbf{1}\left[X_{l,t} = \mathsf{I}\right] = \sum_{l \in \mathcal{I}(t)} a_{lk} \tag{2.6.2}$$

is the number of individuals that are in class k and infective at time t. The removals happen independent of class at a rate of ρ for each individual in $\mathcal{I}(t)$. The initial infective is chosen with distribution $\mathbb{P}_\iota$, a distribution on U, and the complete stochastic process for the epidemic is represented as $\text{BipE}\left(\mathbf{A}, \boldsymbol{\lambda}, \rho, \mathbb{P}_\iota\right)$ where $\boldsymbol{\lambda}$ is the vector of infection rates $\lambda_k \, k \in V$.

The likelihood for the bipartite graph epidemic is presented in section 4.5.1 where it is used for inference.

Algorithm to simulate a bipartite graph epidemic

The linear structure across columns of the infection rate permits an efficient algorithm for simulation, which keeps track of the state of each individual $X_{j,t} \in \{\mathsf{S}, \mathsf{I}, \mathsf{R}\}$ and the counts of infectives $I_k(t)$ and susceptibles $S_k(t)$ in each column k. This is shown in algorithm 2.1 where without loss of generality we take $U = \{1 \ldots n_p\}$ and $V = \{1 \ldots n_K\}$.

[9] note that $\mathcal{R}(0)$ and $\mathcal{R}_0$ are distinct

Algorithm 2.1 Simulation of a bipartite graph epidemic

```
1. sample initial infective $i \sim \mathbb{P}_\iota$
2. initialise:
    (a) set $t = 0$;
    (b) for $k = 1 \ldots K$ set $I_k(0) = a_{ik}$; $S_k(0) = n_p - a_{ik}$ ; $I(0) = 1$; $S(0) = n_p - 1$
    (c) set $X_{i,0} = \mathsf{I}$, for $j \in \{1 \ldots n_p\} \setminus \{i\}$ set $X_{j,0} = \mathsf{S}$
3. repeat the remaining steps for each event time $t$,
    (a) a maximum of $2n_p - 1$steps will be taken
4. calculate infection rates for each column
    (a) $r_k = \lambda_k I_k(t) S_k(t)$
5. calculate total event rate $\lambda = \rho I(t) + \sum_{k=1}^{n_k} r_k$
6. sample $\Delta t$ the time to the next event exponential rate $\lambda$.
7. set $t' = t + \Delta t$
8. sample the event type:
    (a) with probability $\rho I(t) / \lambda$ it is a removal
        i. sample the individual to be removed uniformly from $\{j : X_{j,t} = \mathsf{I}\}$
        ii. set $X_{j,t'} = \mathsf{S}$; $I(t) = I(t) - 1$;
            for each $k$ set $I_k(t) = I_k(t) - a_{jk}$
    (b) with probability $r_k/\lambda$ it is an infection in column $k$
        i. sample the individual to be infected uniformly from $\{j : X_{j,t} = \mathsf{S} \,\&\, a_{jk} = 1\}$
        ii. set $X_{j,t'} = \mathsf{I}$; $I(t) = I(t) + 1$ ;$S(t) = S(t) - 1$;
        iii. for each $k$ set $I_k(t) = I_k(t) + a_{jk}$ and $S_k(t) = S_k(t) - a_{jk}$
9. set $t = t'$, record $t$ and its event type
10. if $I(t)$ is zero return the epidemic times and stop
```

2.6.4 Bipartite Representations of Standard Epidemic Models

In this section several commonly used epidemic models which incorporate some form of heterogeneity are shown to have a bipartite graph based representation.

Household model

The widely studied household model, described briefly above, with a within household infection rate $\eta(n) = n\lambda_H$, is readily represented as a bipartite graph epidemic. When the number of households is n_h this has a bipartite graph representation with an adjacency matrix of size $n_p \times (n_h + 1)$ where n_p is the sum of all the household sizes. With a fixed household size m then $n_p = n_h m$.

The bipartite graph epidemic of equation 2.6.1 is obtained by setting $\lambda_1 = \lambda_G$ and $\lambda_{j+1} = \lambda_H$ for $1 \leq j \leq m$ and the adjacency matrix is $a_{i,1} = 1$ for all $i \leq n_p$ and $a_{i,j+1} = 1$ for $N_{j-1} < i \leq N_j$, where $N_0 = 0$ and N_j is the sum of household sizes $1 \ldots j$. For example for 4 houses of sizes 2,3,3,4 the adjacency matrix and associated infection rates is shown in table 2.6.1.

λ_G	λ_H	λ_H	λ_H	λ_H
1	1			
1	1			
1		1		
1		1		
1		1		
1			1	
1			1	
1			1	
1				1
1				1
1				1
1				1

Table 2.6.1: bipartite representation of a household model

Multi-type model

The frequently used multi-type model mentioned in section 2.5.3 with infection rates $\psi_{i,j}$ between an infective in group i and a susceptible in group j, can be considered as a bipartite graph epidemic model subject to conditions on $\psi_{i,j}$.

Theorem 4. *If $\Psi = (\psi_{i,j}\, 1 \leq i, j \leq m)$ is symmetric and $\psi_{i,i} \geq \sum_{j \neq i} \psi_{i,j}$ for all i then a multi-type epidemic model with m types and infection rates Ψ has an equivalent representation as a bipartite graph epidemic model with $m(m+1)/2$ groups.*

Proof. By construction, set $\lambda_k = \psi_{i,i} - \sum_{j \neq i} \psi_{i,j}$ for $k = 1 \ldots m$ and assign the elements of $\psi_{i,j}$ where $i < j$ to λ_k for $k = m + 1 \ldots m(m+1)/2$. Construct an adjacency matrix with columns $k = 1 \ldots m$ for the within type infections each being an indicator vector for type k and the remaining $m(m-1)/2$ columns for the between type infections being a 'logical or' of columns i and j. □

The condition will usually apply if the groups are geographically separate but may not if the groups are split by ages or if varying susceptibility and infectiousness is modelled by a product form for $\psi_{i,j}$. For example for 3 types of sizes 3,2,4

$\psi_{1,1} - \psi_{1,2} - \psi_{1,3}$	$\psi_{2,2} - \psi_{2,1} - \psi_{2,3}$	$\psi_{3,3} - \psi_{3,1} - \psi_{3,2}$	$\psi_{1,2}$	$\psi_{1,3}$	$\psi_{2,3}$
1			1	1	
1			1	1	
1			1	1	
	1		1		1
	1		1		1
		1		1	1
		1		1	1
		1		1	1
		1		1	1

.

Bipartite graph epidemics with asymmetric infection rates A frequently studied model has both susceptibility and infectiousness varying between groups so that $\psi_{i,j} = c_i d_j$ where c_i is the infectiousness and d_j the susceptibility of individuals in groups i and j. An extension to the bipartite graph epidemic defined in section 2.6.3 could be considered: where the infection rates $\lambda_k\, k \in V$ are replaced by two sets of rates λ_k^c and λ_k^d and equations 2.6.1 and 2.6.2 are replaced by

$$\eta_j(t) = \sum_{k \in V} a_{jk} \lambda_k^d \sum_{l \in \mathcal{I}(t)} a_{lk} \lambda_k^c I_k(t) \tag{2.6.3}$$

as the rate of infections at time t for each susceptible individual $j \in \mathcal{S}(t)$. This would give an immediate correspondence between this extended model and the important subset of multitype models, but at the expense of extra complexity and is not considered further here.

Other models

Random graph Although any graph can be represented in a bipartite form using the clique decomposition, the usual model with a constant infection rate along each edge does not in general have a bipartite graph epidemic representation. The exceptions include the set of graphs formed from the lower projection of a bipartite graph with an adjacency matrix that contains no repeated rows with more than one 1. This set is composed of graphs composed of cliques with the overlap between cliques containing at most one vertex.

Spatial models A frequently used spatial model for epidemics is to have the infection rate between two individuals depend inversely on a spatial kernel, a very similar set of infection rates can be obtained by combining a bipartite graph epidemic, with overlapping spatial tilings.

The simplest example is to choose a small number m of tilings, 3 for example and construct the first tiling with vertices at (im, jm) for $i, j \in \mathbb{Z}$ and subsequent ones at $(im + 1, jm + 1)$ $(im + 2, jm + 2)$ etcetera. Now take the spatial locations, suitably scaled, and for each tiling and each individual determine which square contains the location, each square on each tiling corresponds to a group/column. Any pair of rows/individuals will be in the same square for 0,1,2 or 3 tilings and so have 0,1,2 or 3 columns in common. The infection rates of the spatial kernel model and the proposed bipartite graph model will be approximately proportional. Increasing m will bring the models closer but increases the computational burden.

A three level model A model incorporating households, schools and workplaces is considered by Britton et al. (2011) which can also be represented as a bipartite graph epidemic model. Their example has 500 households of size 4, where the 2 adults in a house each attend one of 40 work places and the 2 children attend the same school of size 100. A straightforward extension of the household representation above is used to represent this with 551 columns 500+40+10 + 1 for a global infection possibility. This model goes a long way to capturing the most obvious heterogeneities in urban life and has been used to simulate examples from their model.

2.7 Indian Buffet Epidemic (IBufE)

When the contact structure underlying the epidemic is unknown, but believed to be both non-homogeneous and possibly having significant effects, a non-parametric approach to modelling the contact structure is appropriate as it provides more flexibility and less unexpected consequences than choosing a model from a restricted set which may inadvertently imply unexpected features. The Indian Buffet Epidemic model provides such an approach, it has been developed to provide a model that fits a wide range of heterogeneity in the contact process with two parameters that describe the departure from homogeneity but which does not require detailed knowledge of individuals contact behaviour. An Indian Buffet Processes is used to provide a distribution over the space of possible bipartite graphs which provides the contact distribution for the epidemic, this original concept is described in the remainder of this chapter.

2.7.1 Indian Buffet Processes

The Indian Buffet Process (IBP) was introduced by Griffiths and Ghahramani (2005) as a generalisation of the Dirichlet process and the Chinese restaurant process. In the original application it provides a distribution for a latent class membership matrix **Z** which is indirectly observed via a linear observation process. A recent review article is Griffiths and Ghahramani (2011) where the IBP is described as a stochastic process defining a probability distribution over equivalence classes of sparse binary matrices with a finite number of rows and an unbounded number of columns. The IBP is a possible distribution in any situation requiring a binary matrix with unidentifiable columns and so after describing the IBP it will be combined with the bipartite graph epidemic model to give the Indian Buffet Epidemic (IBufE), where the IBP matrix **Z** is used as a bipartite adjacency matrix for the epidemic. The finite number of rows correspond to the population and the columns to unspecified locations and/or times of potential contacts. It was pointed out in section 2.6.1 that any graph or contact structure can represented as a bipartite graph and so as the IBP provides a distribution over all binary matrices it can be used as a distribution over all contact structures.

The IBP has several representations all of which yield equivalent distributions over slightly different equivalence classes of binary matrices and which allow a variety of different approaches to MCMC for an IBP based problem. The name comes from a metaphor for a sequential process where each new individual selects classes (dishes) from an infinite Indian Buffet and the choice is recorded as the elements $z_{i,k}$ of the matrix **Z**. The original IBP has one parameter $\alpha > 0$, which governs the expected number per row. The two parameter IBP is mentioned as a possible extension in several papers and as explained later is necessary to give realistic contact structures for an epidemic. The reason is that with high probability the one parameter IBP has one or more columns which are nearly all 1 and so nearly indistinguishable from homogeneous mixing. The clearest definition of the two parameter IBP is in Ghahramani et al. (2007) which is followed here, the single parameter distribution is obtained by setting $\beta = 1$. As β is increased the matrices become sparser with more columns and a lower maximum column count.

The limit definition of the IBP is the limit of a distribution over matrices with K columns, as $K \to \infty$, first a distribution for finite K is introduced. For finite K and N consider the distribution over all 2^{NK} binary $N \times K$ matrix, with entries $z_{i,k}$ where for each column or class k, the entries $z_{i,k} \sim \text{Bernoulli}(\psi_k)$ independently for $i = 1 \ldots N$ and for $k = 1 \ldots K$ ψ_k has a beta$(\alpha\beta/K,\ \beta)$ distribution[10]. The probability of $\mathbf{Z}$ is

$$\mathbb{P}(\mathbf{Z}) = \prod_{k=1}^{K} \int_0^1 \prod_{i=1}^{N} \psi_k^{z_{i,k}} (1-\psi_k)^{1-z_{i,k}} \text{beta}(\psi_k; \alpha\beta/K,\ \beta) d\psi_k$$

The independence of the $z_{i,k}$ within a column means the probability of $\mathbf{Z}$ depends only on $m_k = \sum_{j=1}^{N} Z_{j,k}$ and is a product of beta-binomial distributions

$$\mathbb{P}(\mathbf{Z}) = \prod_{k=1}^{K} \frac{\mathrm{B}(m_k + \frac{\alpha\beta}{K},\ N - m_k + \beta)}{\mathrm{B}(\frac{\alpha\beta}{K},\ \beta)} \tag{2.7.1}$$

where the usual term $\binom{N}{m_k}$ from the beta-binomial distribution is absent because of the many $\mathbf{Z}$ which have the same m_k. The beta-binomial distribution is described in appendix B together with some results on IBP distributions.

As we are interested in unlabelled columns it is appropriate to consider equivalence classes of matrices that permit the study of the distributions of infinite matrices. The “left ordered form” function $\text{lof}(\mathbf{Z})$ is a many to one mapping of all binary matrices which is defined in terms of the binary representation of each column. Each column is considered as a binary number $< 2^N$ with the first row as the most significant bit. The function $\text{lof}(\mathbf{Z})$ orders the columns of $\mathbf{Z}$ in decreasing order of their binary representation, two examples are shown in figure 2.7.1 (a) has $\beta = 1$ while (b) has $\beta > 1$. We denote the set of equivalence classes of matrices with distinct left ordered forms as $\mathcal{Z}_{\text{lof}}$.

Definition 2. The Indian Buffet Process with parameters α, β defines a distribution on $\mathcal{Z}_{\text{lof}}$ which is denoted $\text{IBP}(\alpha, \beta, N)$. If $\mathbf{Z} \sim \text{IBP}(\alpha, \beta, N)$ and$[\mathbf{Z}]$ is the lof equivalence class containing $\mathbf{Z}$ then

[10] The notation for the the beta function $\mathrm{B}(x, y)$ and beta distribution are defined in appendix D

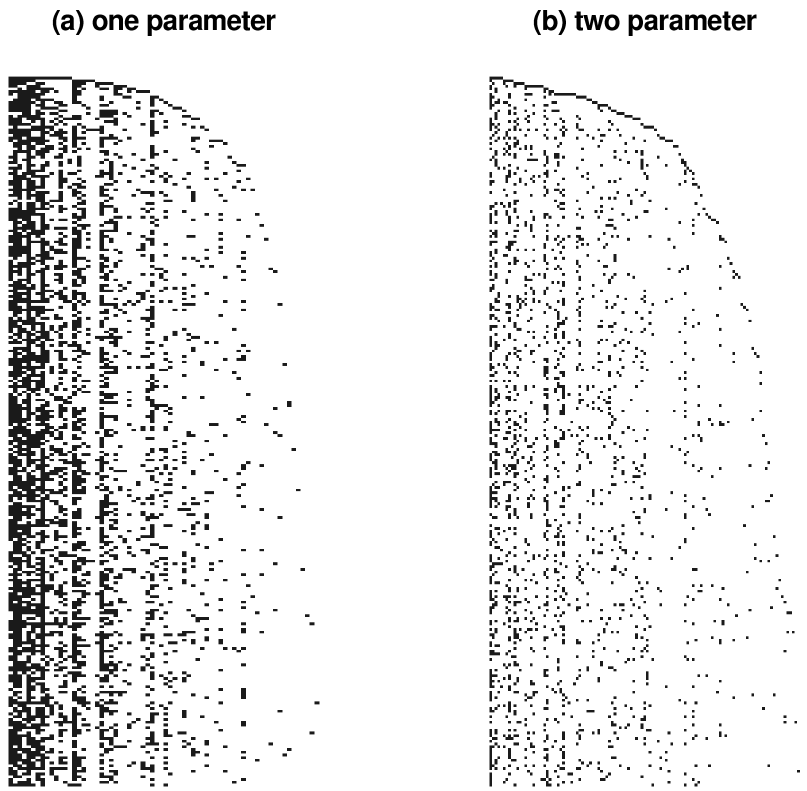


Figure 2.7.1: Simulations of Indian Buffet process, both in left ordered form (lof) $N = 260$, (a) $\alpha = 15$, $\beta = 1$, (b) $\alpha = 8$, $\beta = 4$

$$\mathbb{P}\left([\mathbf{Z}]\right) = \frac{(\alpha\beta)^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} e^{-\overline{K}_+} \prod_{k=1}^{K_+} B(m_k,\, N - m_k + \beta) \qquad (2.7.2)$$

where K_h is the number of columns with binary representation h, $m_k = \sum_{j=1}^{N} Z_{j,k}$ and K_+ is the number of non-zero columns, $m_k > 0$, these three terms are functions of $\mathbf{Z}$, hence the non appearance of $\mathbf{Z}$ on the r.h.s.. The expected value of K_+ is represented as $\overline{K}_+ = \mathbb{E}\left(K_+\right) = \alpha \sum_{j=1}^{N} \frac{\beta}{\beta+j-1}$.

It is easily shown (see Griffiths and Ghahramani (2005)) that the number of classes for an individual/row $\sum_{k=1}^{K} z_{i,k}$ has a Poisson distribution with mean α independently of K.

Figure 2.7.1 (a) shows an example $\mathbf{Z}$ ordered in left-ordered binary form (*lof*), it can be seen that a few columns on the left contain nearly all individuals. The reason that the number in the largest class in the IBP is close to N can be seen by noting that it will usually coincide with the largest ψ_k. As when $\beta = 1$ ψ_k ~Beta$(\alpha/K, 1)$ with p.d.f. $\propto x^{\alpha/K-1}$ and c.d.f. $= x^{\alpha/K}$ the distribution of the maximum of K such random variables $\psi_{(K)}$ is therefore x^{α} which is concentrated near 1.

IBP two parameter sequential form The sequential formulation in terms of the restaurant metaphor is that each customer chooses previously sampled dishes with probability dependent on the the number of times previous customers have chosen the dish. The first customer chooses Poisson(α) dishes, which sets $z_{1,k} = 1$ for k up to the sampled value. Subsequent customers, $i > 1$, choose each previously sampled dish with probability $P(z_{i,k} = 1) = m_{i,k}/(\beta+i-1)$ (i chooses dish k) where $m_{i,k} = \sum_{j=1}^{i-1} z_{jk}$ and also takes a number of new dishes, denoted $K_{i,E}$ sampled from a Poisson distribution, $K_{i,E} \sim \text{Poisson}\left(\alpha\beta/(\beta + i - 1)\right)$. The average number of dishes per customer is α, that is $\mathbb{E}\left(\sum_{k=1}^{\infty} z_{ik}\right) = \alpha$ for each each row i.

The sequential formulation does not generate matrices that are always in lof (.) form, and we denote the set of possible matrices by $\mathcal{Z}_{\text{seq}}$. The distribution is

$$\mathbb{P}\left(\mathbf{Z}\right) = \frac{(\alpha\beta)^{K_+}}{\prod_{i=1}^{N} K_{i,E}!} e^{-\overline{K}_+} \prod_{k=1}^{K_+} B(m_k,\, N - m_k + \beta) \tag{2.7.3}$$

which is very similar to the *lof* form and equation 2.7.2 is obtained by replacing $\prod_{i=1}^{N} K_{i,E}!$ with $\prod_{h=1}^{2^N-1} K_h!$, which corresponds to the combinations of columns in the *lof* map of $\mathcal{Z}_{\text{seq}}$ to $\mathcal{Z}_{\text{lof}}$.

Stick Breaking A third construction of the IBP based on stick breaking is given in Teh et al. (2007), however it is only presented for the one parameter form where use is made of the cumulative product of ν_k i.i.d. Beta$(\alpha, 1)$ random variables, giving $\psi_{(k)} = \nu_k \psi_{(k-1)} = \prod_{l=1}^{k} \nu_l$ where $\psi_{(k)}$ is the decreasing ordering of ψ_k. This cannot be extended to $\beta \neq 1$ so is not considered here.

Simulating the IBP Simulation of the IBP is a pre-requisite for simulating the Indian buffet epidemic described below and for MCMC inference on it which is described in chapter 4. It is straightforward to simulate the IBP using any of the formulations, for $\beta > 1$ the finite K method would require allocating a large matrix, which is avoided in the sequential method. Converting a simulated matrix to *lof* form is usually unnecessary, when it is required it is relatively time consuming.

2.7.2 Definition of the Indian Buffet Epidemic

In defining the Indian buffet epidemic we can consider two scenarios, which differ in the unlikely case of observing two epidemics with identical contact structures. We can consider the matrix $\mathbf{Z}$ to represent some real but unknown and unobservable contact matrix, where the IBP is a Bayesian prior on $\mathcal{Z}$. Alternatively we can consider $\mathbf{Z}$ to be part of a random effects model where the distribution of $\mathbf{Z}$ reflects the randomness inherent in contact processes. In the former case multiple epidemics would use the same $\mathbf{Z}$, in the latter case each epidemic would sample a new $\mathbf{Z}$.

Definition 3. The Indian Buffet Epidemic on a population of size n_p with parameters $\theta = (\alpha, \beta, \lambda, \rho)$, an infection rate scaling function $\xi(n, \lambda)$ and an initial infective distribution $\mathbb{P}_\iota$ is a two level stochastic process. A contact matrix $\mathbf{Z} \sim \text{IBP}(\alpha, \beta, n_p)$ is used as the adjacency matrix for a bipartite graph epidemic as defined in section 2.6.3, the infection rate within each column k of $\mathbf{Z}$ is $\lambda_k = \xi(m_k, \lambda)$ where $m_k = \sum_{j=1}^{n_p} z_{jk}$ and these are used in equations 2.6.1, 2.6.2 to define the epidemic. The complete stochastic process is denoted $\text{IBufE}(\theta, n_p, \xi, \mathbb{P}_\iota)$.

2.7.3 Analysis of the Indian Buffet Epidemic

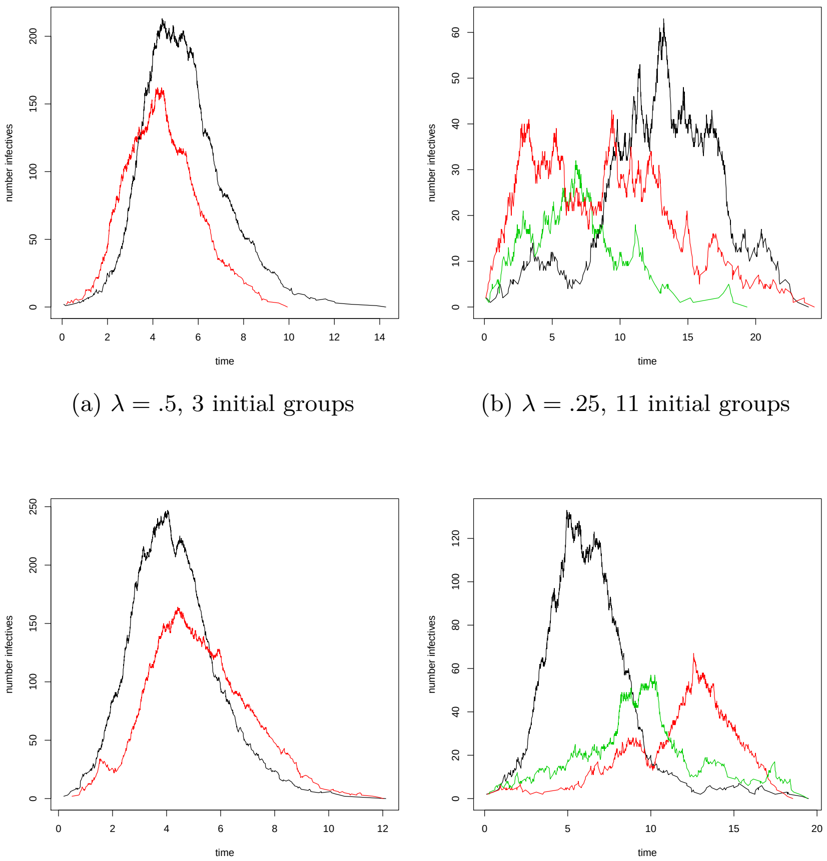
Understanding the characteristics of the Indian buffet epidemic and the differences from a homogeneously mixing epidemic is a challenging problem, extension of existing analytic results such as those for epidemics on graphs is hindered by the dependence within the IBP and in particular the existence of columns within the IBP that contain a single entry and have no impact on the epidemic. Analysis has largely relied on simulations which although simple to perform are complex to analyse. Analytic results are limited to analysis of the initial infection rate which is described below. Simulating the Indian buffet epidemic is in principle straightforward, first a $\mathbf{Z} \sim \text{IBP}$ is generated, then the epidemic is simulated using the bipartite algorithm given in section 2.6.3. The hierarchical nature of the model means that choices must be made of what to hold constant between simulations and what to vary, even for a fixed set of parameters.

At first sight simulations of the IBufE look quite similar to the GSE and as with the GSE, as the ratio of infection rate to recovery rate is varied different shapes of infective curves are obtained, some examples are discussed below.

Infection Rate Scaling Function The form chosen for the infection rate scaling function $\xi(n, \lambda)$ is also important, it reflects the way in which infectivity scales with population, and as mentioned in section 2.2.1 the paper by Begon et al. (2002) considers this topic. Initial experiments used $\xi(n, \lambda) = \lambda$, particularly when $\beta \leq 1$ one column contains most individuals, $m_k \approx n_p$ and this column dominates the epidemic, which is then indistinguishable from the GSE. Household models typically have very different global and household infection rates, in a similar vein a choice such as $\xi(n, \lambda) = \lambda/n$ seems appropriate, which in some sense makes each column have similar effect on the epidemic, this is used subsequently. Other choices such as $\xi(n, \lambda) = \lambda/\sqrt{n}$ have also been simulated, further choices could be made by consideration of studies such as that by Hethcote (1994) of five human diseases in communities with population sizes from 1,000 to 400,000. By fitting an incidence of the form $N^{\nu}SI/N$ he finds that ν is between 0.03 and 0.07. The communities he studies are themselves combinations of the smaller groups implicit in the Indian buffet epidemic so the value of ν is not necessarily appropriate for the choice of $\xi(n, \lambda)$. An analysis of the effect different choices of infection rate scaling function have on the epidemics requires a deeper understanding of the other components of the Indian buffet epidemic and unless explicitly mentioned $\xi(n, \lambda) = \lambda/n$ is used in the remainder of this thesis.

Example Simulations of the Indian Buffet Epidemic As an illustration of the variation that can arise, some examples are shown in figure 2.7.2 of simulations on a single $\mathbf{Z} \sim \text{IBP}(4, 25, 1000)$. This example has $K_+ = 374$ and 285 columns with more than one entry. The 1000 rows have between 0 and 11 entries, and the bipartite graph, excluding 8 empty rows, is connected. Each sub-figure shows 2 or 3 epidemics with the same initial infective and value for λ, although each infective curve could be from a GSE there is more variation. The plots in sub-figures (a) and (c) have the same value of $\lambda = 0.5$ and are fairly similar, differing in the number of groups the initial infective is in. The plots in sub-figures (b) and (d) have a lower infection rate and show a much larger variation.

Effects of Parameters on the Indian Buffet Epidemic The parameters $\theta = (\alpha, \beta, \lambda, \rho)$ and the initial infective distribution $\mathbb{P}_\iota$ of the Indian buffet epidemic obviously affect the epidemic distribution in different ways. The removal rate ρ only affects the scaling of time and can be taken w.l.o.g. as 1. The infection rate parameter λ can readily be understood in a qualitative way, for small values of λ the epidemic has a high probability of being very small and if not small, the peak incidence will probably be small. Large values of λ will give large epidemics and often will have a rapid increase in infectives, however in contrast to the GSE as $\lambda \to \infty$ the final size will converge on the size of the component containing the initial infective. Usually there will be a giant component which is smaller than the population n_p and so the final size will have a distribution with a mean strictly smaller than n_p.



(a) $\lambda = .5$, 3 initial groups

(b) $\lambda = .25$, 11 initial groups

(c) $\lambda = .5$, 6 initial groups

(d) $\lambda = .33$, 11 initial groups

Figure 2.7.2: Example simulations of Indian buffet epidemics, with the same $\mathbf{Z} \sim \text{IBP}\,(4, 25, 1000)$, each sub-figure contains epidemics with the same initial infective and value for λ. (Note the axes differ between sub-figures)

Understanding the influence of the other parameters of the Indian buffet epidemic, including the distribution of the initial infective, on the outcome is made more difficult by the increased variability described in the preceding paragraph and illustrated in figure 2.7.2 which must be accounted for in the analysis. The most obvious outcomes of a single outbreak are final size, peak incidence and duration which can all be analysed by simulation. When a set of epidemics is considered, results on the probability of a “major” epidemic are difficult to obtain as there are many cases with no clear threshold in the final size distribution. However consideration of the first infection after the initial infective permits some progress to be made.

Initial Infective An important difference from the GSE is the importance of the choice of initial infective from the distribution $\mathbb{P}_\iota$. In models for epidemics on graphs it is well known that the choice of initial infective has a strong influence on the probability of a major epidemic and also on the duration. In bipartite graph epidemics the choice of initial infective has a larger effect on the probability of a major epidemic as the initial infection rate is roughly proportional to the number of groups that the initial infective is in. Three simulations with the same $\mathbf{Z} \sim \text{IBP}\,(6, 1, 600)$, $\xi\,(n, \lambda) = \lambda/n$ and different initial infectives are plotted in figure 2.7.3, they have similar final sizes and the effect is visible of the number of groups which the initial infective is in on the time to get established and hence the duration. The epidemic which starts in 8 groups also shows a flattened peak, which arises from the total number of infectives being the sum of epidemics in two large groups with slightly different time scales.

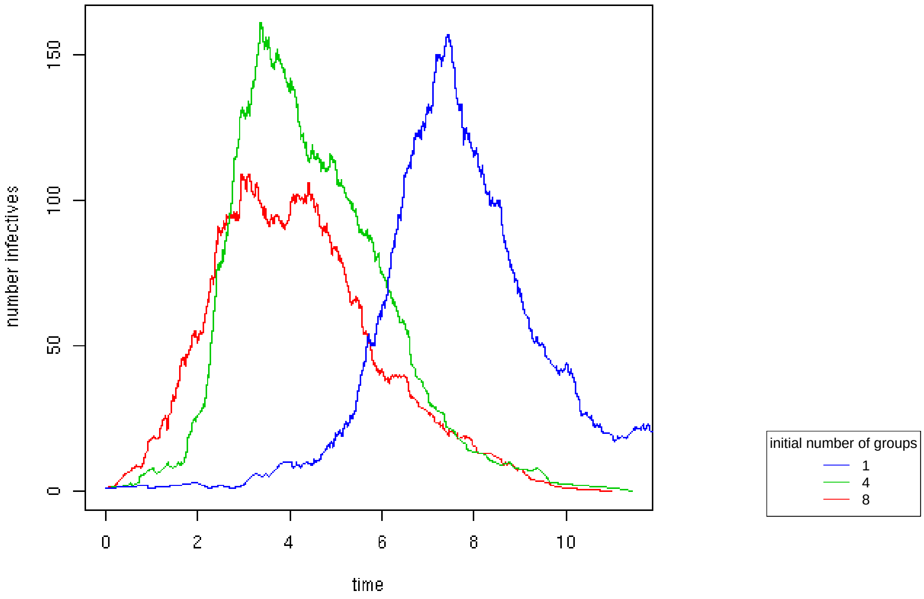


Figure 2.7.3: Three simulations showing the number of infectives, from the same $\mathbf{Z} \sim \text{IBP}\,(6, 1, 600)$, with initial infectives in 1,4 or 8 groups.

Properties of the IBP that Affect the Epidemic

An approach to investigating the choice of parameters for the IBP that reflect reality and their effect on the IBufE is to examine the two graph projections along with the marginal distributions of the adjacency matrix. Several statistics are available for the IBP, the distributions of some are known but others relevant to the connectivity of the lower projection are not. Here some empirical results on their distributions are given.

First to clarify the motivation for the relevant statistics an example of a small IBP matrix ($N = 15$) with two empty rows (13,15) and 6 columns (3,4,5,8,9,10) containing a single 1 is given.

$$
\begin{array}{lcccccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
1 & 1 & . & . & . & . & . & . & . & . & . \\
2 & 1 & 1 & 1 & . & . & . & . & . & . & . \\
3 & 1 & . & . & 1 & . & . & . & . & . & . \\
4 & 1 & 1 & . & . & 1 & . & . & . & . & . \\
5 & . & 1 & . & . & . & 1 & . & . & . & . \\
6 & . & 1 & . & . & . & 1 & . & . & . & . \\
7 & . & 1 & . & . & . & . & 1 & . & . & . \\
8 & . & . & . & . & . & 1 & . & . & . & . \\
9 & . & . & . & . & . & 1 & . & . & . & . \\
10 & . & . & . & . & . & . & 1 & . & . & . \\
11 & . & . & . & . & . & . & 1 & . & . & . \\
12 & . & . & . & . & . & . & . & 1 & . & . \\
13 & . & . & . & . & . & . & . & . & . & . \\
14 & . & . & . & . & . & . & . & . & 1 & 1 \\
15 & . & . & . & . & . & . & . & . & . & .
\end{array}
$$

The nodes corresponding to the empty rows can obviously not be infected and in terms of the connectivity of the lower projection are obviously isolated. The columns containing a single bit cannot affect the epidemic and can be deleted without effect, when these columns are deleted rows 12 and 14 become empty. The connectivity of this example, is a single connected component containing nodes 1 to 11 and 4 isolated nodes, this is the form that most simulated IBP matrices have for large N. The dependence of the structure and connectivity on the parameters α, β and N is non-linear. For example for β=1 and $\alpha \geq 4$ there is a single connected component with high probability for $N > 10$. Whereas for β=8 and $\alpha = 2$ the aymptotic region is only reached around $N \simeq 10^4$.

In the example above if row 15, column 10 contained a 1, an additional component containing nodes 14 and 15 would be formed, leaving two isolated nodes.

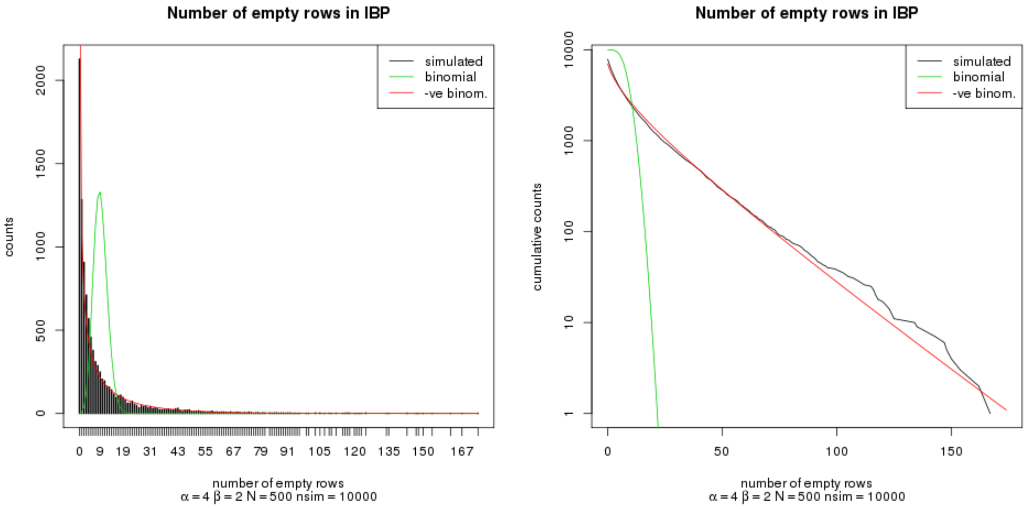


Figure 2.7.4: Number of empty rows in the IBP. The left hand plot compares the counts of simulated data for $\mathbf{Z} \sim \text{IBP}\,(4, 2, 500)$ with a binomial distribution resulting from an assumption of independence and an empirically fitted negative binomial distribution. The right hand plot shows the same data as a log hazard.

In a large number of simulations the vast majority have been of the former kind, a fully connected core and isolated nodes. Some are of the second type, with a giant component with most of the non-isolated nodes and a few small components each with 2 or 3 nodes. In no cases was the second largest component a significant fraction of the size of the giant component. The actual size of the giant component is strongly dependent on the IBP parameter α and the value of N at which the giant connected component emerges is strongly dependent on β.

Several statistics are available for the IBP, the distributions of some are known but others relevant to the connectivity of the lower projection are not. Here some empirical results on their distributions are given.

Recall that each row has a Poisson distribution with mean α but these are highly correlated and the number of empty rows has an over-dispersed distribution relative to the naive binomial distribution with probability $\exp(-\alpha)$ that would result from independence. Empirical study shows that a negative binomial distribution fitted to the known mean and observed variance gives a reasonable fit to all the examples studied. Some examples are shown in figures 2.7.4,2.7.5.

The connectivity of the remaining rows has also been studied. In a large number of simulations the vast majority have been of the former kind, a fully connected core and isolated nodes. Some are of the second type, with a giant component with most of the non-isolated nodes and a few small components each with 2 or 3 nodes.

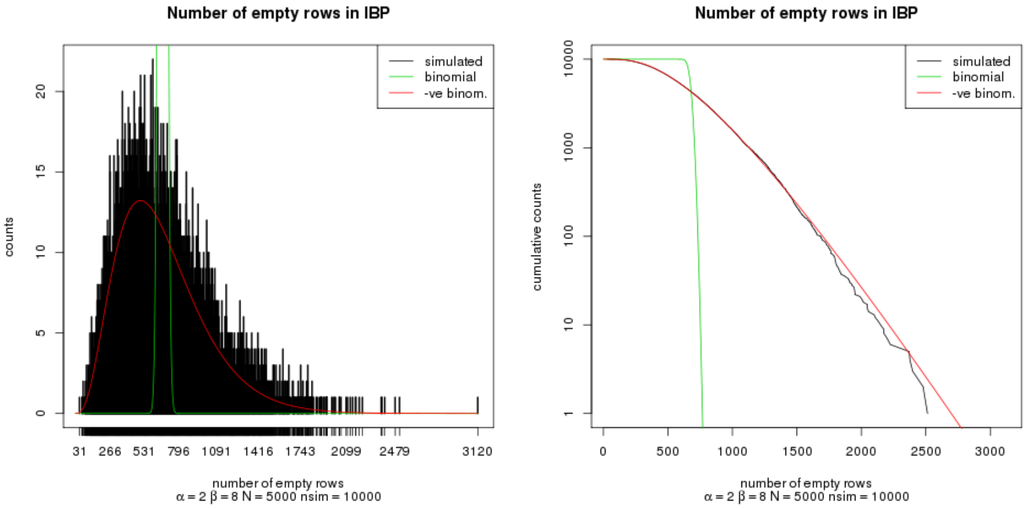


Figure 2.7.5: Number of empty rows in the IBP. The left hand plot, with $\beta = 8$ shows a much wider distribution than the previous plot with $\beta = 2$, compares the counts of simulated data for $\mathbf{Z} \sim \text{IBP}\,(2, 8, 5000)$ with a binomial distribution resulting from an assumption of independence and an empirically fitted negative binomial distribution. The right hand plot shows the same data as a log hazard.

In no cases was the second largest component a significant fraction of the size of the giant component. The actual size of the giant component is strongly dependent on the IBP parameters α and β through the distribution of empty rows as described above.

An example of the distibution of the connectivity is shown in figure 2.7.6 for 1000 simulations of $\mathbf{Z} \sim \text{IBP}\,(3, 2, 2000)$ the size of the giant component (`gc`), the number of components (`nc`) and the number of nodes with row sum > 0 (`notgc`) are shown together with other statistics.

In summary the graphs formed from the IBP with high probability have the following form:

- a set of individuals/rows that are completely disconnected, which are chosen from a distribution with mean $N \exp(-\alpha)$.
- a giant connected component, which usually contains all the remaining rows after the empty rows are removed.
- a set of groups/columns that are irrelevant to the epidemic, containing only a single entry.

This observation and the simulation results above leads to the following conjecture:

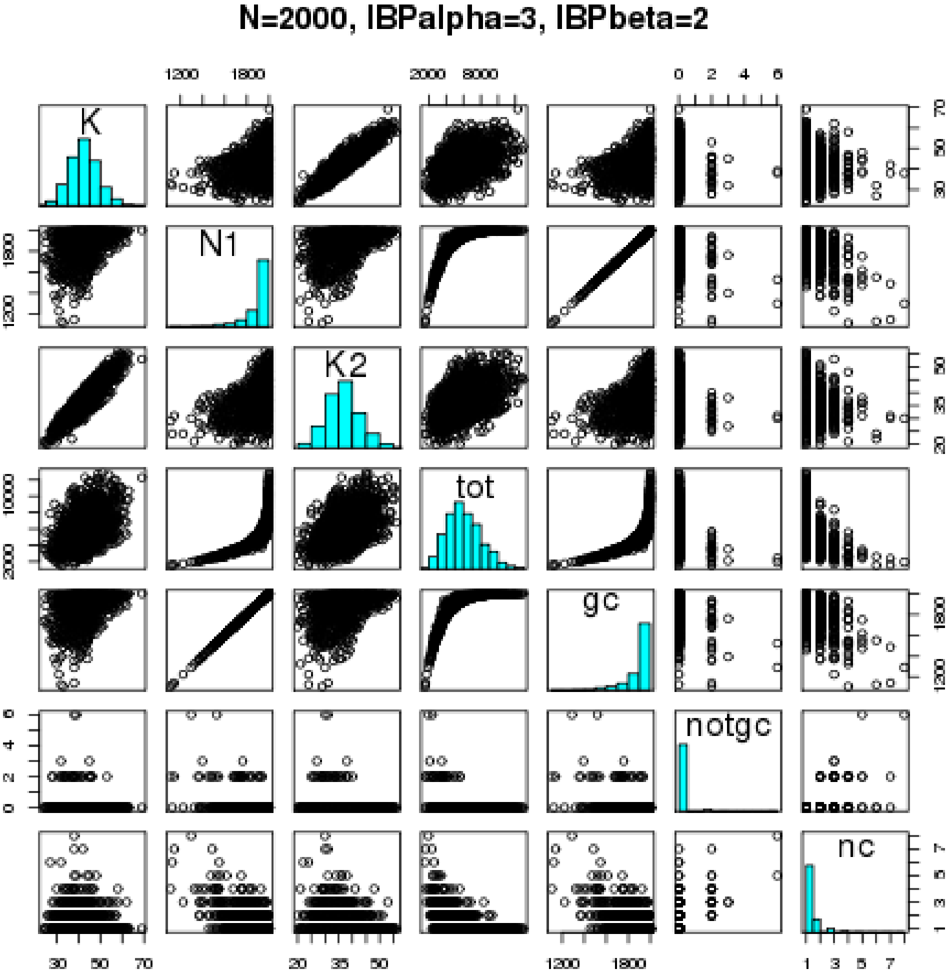


Figure 2.7.6: Connectivity of IBP lower projection. This pairs plot, displays the pairwise dependence and marginal distribution of 7 statistics from 1000 simulations of $\mathbf{Z} \sim \mathrm{IBP}\,(3, 2, 2000)$. The variables are numbers of: `K` columns, `N1` non-empty rows, `K2` columns with at least 2 bits, `tot` total number of bits, `gc` size of the giant component, `notgc` nodes with row sum > 0 and not in the giant component, `nc` number of components with size > 1

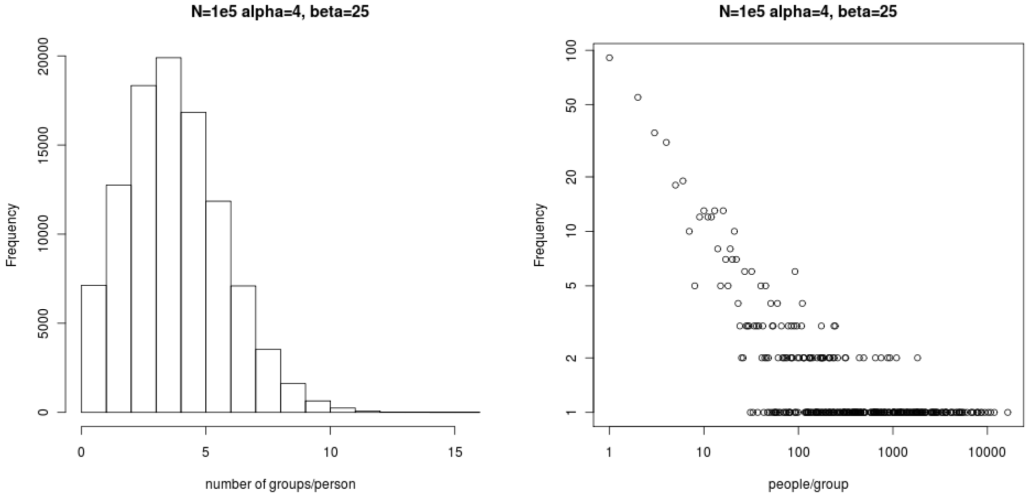


Figure 2.7.7: Distribution of marginal sums for an example IBP, $N = 10^5$, $\alpha = 4$, $\beta = 25$. The left hand plot is a standard histogram, the right hand a count of counts.

Conjecture 1. *As $N \to \infty$ the bipartite graph formed from the non-zero rows of $\mathbf{Z} \sim IBP(\alpha, \beta, N)$ has a giant connected component almost surely. The fraction of nodes in the giant component converges to $1 - \exp(-\alpha)$.*

Impact of Margins of the IBP on the Indian Buffet Epidemic The behaviour of the epidemic is strongly affected by the distribution of the two margins of the IBP both of which can be calculated, the row sums are available explicitly as a Poisson(α) distribution. The distribution of column sums can be approximated numerically from the finite K representation, details are given in appendix B.3. An example of the distribution of the margins for a large IBP with parameters ($N = 10^5$, $\alpha = 4$, $\beta = 25$), is shown in figure 2.7.7, which shows similarity with the distributions obtained in the EpiSims project for estimates of the contact matrices for several cities which was mentioned in section 2.5.5. This gives support to the belief that the IBP gives a plausible model for contact distributions.

The distribution of the column sums has a heavy tail in terms of group size, with many small values and a few large values. As the columns of $\mathbf{Z}$ are ordered in *lof* form the distribution is strongly correlated with the position of the column, however as interest is in the joint distribution of all the column sums consideration of the cumulative probability provides the desired distribution. The finite K distribution can be used to approximate numerical values of the IBP distribution arbitrarily accurately, examples are shown in figure 2.7.8 for three values of K, together with

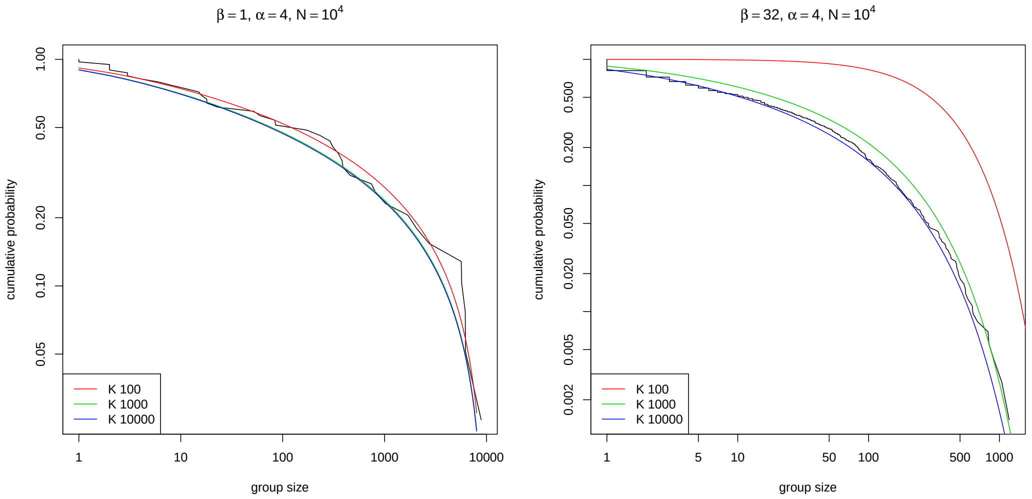


Figure 2.7.8: Distribution of IBP group size(people/group). Comparison of simulated examples with finite K approximations.

the distribution from a single simulated example. The closeness of the simulated curve to the calculated curves, for all except $K = 100$ in the right hand plot where $\beta = 32$, suggests that the approximations are good.

For epidemics, columns with a single entry, $m_k = 1$, have no influence on the epidemic and so can usually be ignored, they are included in the distribution of the IBP margins shown in figure 2.7.7, as the left hand column in the histogram of row sums and the point in the upper-left corner of the plot of column sums.

Final size of the Indian Buffet Epidemic

The distribution of the final size of an IBufE combines two parts, the final size conditioned on $\mathbf{Z} \sim \text{IBP}\,(\alpha, \beta, n_p)$ and the size of the component of the bipartite graph containing the initial infective, usually a giant component will be present, unless the population n_p is small (e.g. < 50). Both parts of the distribution are dependent on the distribution of the row-sums of $\mathbf{Z}$ which we denote here by $X_i = \sum_k z_{ik}$.

The size of the giant component is clearly less than the number of rows which contain at least one 1. The distribution of the number of isolated individuals is bounded by the distribution of the number of X_i which are 0.

Simulations show that the mean final size of epidemics conditioned on $\mathbf{Z}$ are approximately proportional to the sample mean of the X_i, $\bar{X} = n_p^{-1} \sum_i X_i$. From the definition of the IBP the expected value of $\bar{X}$ is α, but although the distribution of an individual X_i is Poisson(α) they are strongly correlated. The joint distribution of the row sums can be obtained recursively from the sequential representation as:

$$\begin{aligned} X_1 &\sim \text{Poisson}\,(\alpha) \\ X_2 &\sim \text{binomial}(X_1, 1/(\beta+1)) + \text{Poisson}\,(\alpha\beta/(\beta+1)) \\ &\cdots \\ X_i &\sim \text{binomial}(X_{i-1}, 1/(\beta+i-1)) + \text{Poisson}\,(\alpha\beta/(\beta+i-1)) \end{aligned}$$

from which the distribution can be evaluated numerically.

An example of two final size distributions for Indian buffet epidemics is shown in figure 2.7.9 both have a population of $n_p = 10^3$, the two IBufE curves are from more than 10^5 simulations with $\alpha = 8$ and $\beta = 4$ or 25. When compared with the exact GSE distribution, which is also plotted, it can be seen, bearing in mind the log scale of probability, that the distributions are much wider. The distributions shown are not conditioned on a “major” epidemic and although the local minima would give plausible thresholds for defining a “major” epidemic the simple results in the GSE linking the probability of a “major” epidemic, initial infection rate and final size distribution do not apply. The exact probabilities for the first few values of the final size (including the initial infective) are similar:

Final size	1	2	3	4	5
GSE	0.357	0.0822	0.0378	0.0218	0.0140
$\beta = 4$	0.363	0.0778	0.0351	0.0191	0.0122
$\beta = 25$	0.370	0.0812	0.0347	0.0202	0.0125

.

Initial Infection Rate of the Indian Buffet Epidemic The most obvious outcomes of a single outbreak are final size, peak incidence and duration which are difficult to analyse, however consideration of the first infection after the initial infective permits some progress to be made. It is easy to see the strong effect the choice of initial infective has by looking at the probability that at least one additional individual is infected. The total initial infection rate b_i when the initial infective is i is

$$b_i = \sum_{k=0}^{K} z_{ik}\,(m_k - 1)\,\lambda_k$$

which gives the probability of a further infection as $\frac{b_i}{\rho + b_i}$ and of immediate extinction

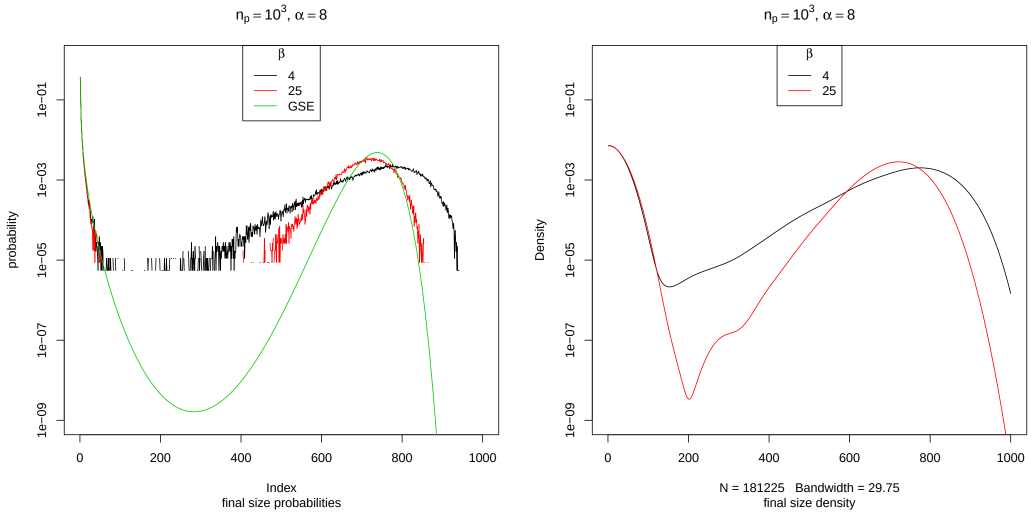


Figure 2.7.9: Final size distributions for two Indian buffet epidemics with population $n_p = 10^3$, $\alpha = 8$ and $\beta = 4$ or 25.
The two IBufE curves are from 181,225 and 114,710 simulations, the left hand plot shows the probabilties directly estimated from the counts. For comparison the exact probabilities for the GSE are also shown. The right hand plot shows kernel density estimates obtained from the simulated counts.

as $(1 + b_i/\rho)^{-1}$. If the initial infective is chosen to be i with probability g_i

$$\mathbb{P}\left(\text{second infection}\right) = \sum_{i=1}^{n_p} g_i \frac{b_i}{\rho + b_i}. \qquad (2.7.4)$$

In the Indian buffet epidemic as the number of connected columns for any row in the IBP has a Poisson(α) distribution and the distribution of m_k is beta-binomial, $\mathbb{E}\left(b_i\right)$ can be approximated by assuming that $Z_{i,\cdot}$ and m_k are independent. This is conjectured to be asymptotically true for large n_p.
If $\lambda_k = \xi\left(m_k, \lambda\right)$ is chosen to be $c/\left(m_k - 1\right)$ where c is a constant which is loosely related to $\mathcal{R}_0$ (e.g. λ/ρ)) then b_i is c times the number of groups which i is in and so if the initial infective is chosen uniformly the probability of no further infections is approximately $1/(1+c\alpha)$. So by analogy with the branching process result for the GSE it is conjectured that the probability of a “major epidemic” is approximately $1 - \frac{1}{c\alpha}$, further it is convenient to introduce $\mathcal{R}_{\text{IB}} = \alpha\lambda/\rho$ as a parameter that is expected to be strongly related to the outcomes and loosely connected to $\mathcal{R}_0$ for the GSE.

When $\xi\left(n, \lambda\right)$ is other than $\lambda/(n-1)$ the approximation for $\mathbb{E}\left(b_i\right)$ assuming the two marginals of the IBP are independent is obtained from

$$b_i = \sum_{k=0}^{K} z_{ik} \left(m_k - 1\right) \xi \left(m_k, \lambda\right).$$

Simulations suggest the importance of $\mathcal{R}_{\mathrm{IB}}$ in understanding the variation in epidemic outcomes, further work is needed to relate the distribution of outcomes of the epidemics to the parameters of the Indian buffet epidemic .

2.8 Concluding Remarks on Epidemic Models

Two complementary topics have been presented in this chapter Markov representations of the GSE and bipartite graph epidemic models and in particular the Indian buffet epidemic.

The Markov representation has been shown to be a mechanism for gaining insight into the GSE and closely related models. Exact calculations have revealed the boundaries for the existence of the classic bimodal final size distribution.

The class of epidemic models on bipartite graphs provides a powerful way of constructing new models and comparing existing models. The bipartite graph epidemic model defined in section 2.6.3 provides a foundation for the Indian buffet epidemic which is also developed above. Extensions of both bipartite graph and Indian buffet epidemic models to SEIR and discrete time models are straightforward and have not been described, in particular the combination of the Indian buffet epidemic and the binomial model described in section 2.4.1 should give a model capable of scaling to larger populations.

A graph theoretic conjecture on the existence of a giant component in the IBP has been posed based on simulated examples.

The Indian buffet epidemic has been shown to be an interesting model for an epidemic in a heterogeneously mixing population. Similarities, shown in section 2.7.3, between the distribution of the IBP and the best available estimates of contact structures for an urban human population obtained by the EpiSims project show the plausibility of the Indian buffet epidemic. Analytic results on the Indian buffet epidemic remain elusive but simulation of the epidemic is straightforward using the algorithm above and and numerical results for some of the relevant distributions have been obtained.

Chapter 3

MCMC

3.1 Introduction

The primary aim of this chapter is to present a novel analysis of some aspects of the grouped independence Metropolis-Hastings (GIMH) algorithm which was introduced by Beaumont (2003) and generalised by Andrieu and Roberts (2009). The bias of the closely related approximate algorithm the Monte Carlo within Metropolis algorithm (MCWM) (O'Neill et al., 2000) is also investigated. The analysis is presented in more general terms, to distinguish the original algorithm from the generalised algorithm the name stochastic exact Metropolis-Hastings (SEMH) is introduced. The analysis of GIMH requires a good understanding of some aspects of importance sampling, so a new study of the properties of importance sampling in tractable situations is given in section 3.2.

The necessary parts of the extensive theory of MCMC are described along with the notation which is used. The GIMH is introduced, understanding the sticking that can occur in the GIMH is facilitated by studying it in the more general terms of the SEMH. A new analysis in section 3.5.3 shows that the variance of the weights distribution can explain the sticking of the GIMH and the bias of the MCWM.

A new approximate algorithm the Kernel Metropolis-Hastings (KMH) is proposed in section 3.6 which is expected to overcome the difficulties encountered in applying the GIMH algorithm in practice which are described in chapter 4. The KMH is demonstrated on a multimodal heavy tailed target distribution.

3.1.1 Some Basic Monte Carlo methods

Monte Carlo methods are central to many of the techniques of modern statistics, in particular Markov chain Monte Carlo (MCMC) methods which have a long history of successful use since their introduction to the statistics community by Hastings (1970) a history is given by Robert and Casella (2011) and relevant references are given below, particularly in the section on MCMC. The successful application to increasingly difficult problems has depended on the parallel development of improvements in the theoretical underpinnings and the exponential increase in the power of computers. However in the case of high dimensional Bayesian posterior distributions the basic methods are not sufficient and require more advanced techniques, either exact or approximate, which are the subject of continuing research. Many of these techniques combine simpler techniques in various ways, in particular the techniques described in section 3.4 combine MCMC and importance sampling both of which are described below.

All of the methods depend on the generation of large numbers of random variables with a chosen distribution, which is invariably based on pseudo-random generators. In the early days of Monte Carlo methods the pseudo-random generators available outside the classified community had major defects and bad generators where widely used, however since the 1980s reliable pseudo-random generators have been widely available. Care in using pseudo-random generators is particularly needed in two situations, in parallel computation and when investigating tail events where the difference between discrete computer arithmetic and the continuous ideal is important. In long MCMC runs events that are of zero probability in a continuous model can occur with a small probability when finite computer arithmetic is used, defensive programming is necessary to ensure valid results. All results in this thesis have used the standard R Mersenne-Twister generator (Matsumoto and Nishimura, 1998). Further discussion of pseudo-random generators and references are available in several books for example (Robert and Casella, 2004).

Target distributions

Underlying all Monte Carlo methods is the aim to produces samples from a target distribution, which we represent by a generic density $\pi(x)$ with an implied dominating measure ν over some unspecified space $\mathcal{X} \subset \mathbb{R}^d \times \mathbb{Z}^m$. Typically in practical situations π will often be a Bayesian posterior distribution, we use π to represent an arbitrary unspecified distribution and follow the common practice of distinguishing different distributions by their argument. These samples can be used to estimate expectations of known functions $h(.)$ of the random variable X, $\mathbb{E}_X(h(X)) = \int_{\mathcal{S}} h(x)\pi(x)\nu(dx)$ where $X \sim \pi(.)$ and $\mathcal{S}$ is the support of the distribution. In realistic examples we are unable to simulate directly from π which may be on a high dimensional space and $\mathcal{S}$ may or may not be connected. In pedagogical examples $\mathcal{S}$ will be known, however when the target is a Bayesian posterior, particularly if it includes hidden variables, π may have discontinuities and $\mathcal{S}$ may not be known. So that we can investigate the performance of techniques and observe general principles, simple pedagogical examples are considered below where $\mathcal{S}$ is known and simulation from π may be possible.

3.2 Importance Sampling

Importance sampling was originally introduced as a variance reducing mechanism (Hammersley and Handscomb, 1964) for estimating low probability events, it has since found much wider applicability. In particular is is used below, in the GIMH, to estimate marginal posterior distributions, the description here is given in general terms and the studies in this section use simple parametric distributions to provide insight into the behaviour of the importance weights when they are intractable as is invariably the case within the GIMH.

The importance sampler uses an approximation $q(x)$ to the target density $\pi(x)$ which is chosen so that we can simulate from it and easily compute $q(x)$, the choice is problem specific and a good choice can be problematic. We refer to this as the *proposal*, it is sometimes referred to as an *importance density* or *instrumental density.* Usually interest is in estimating the integral $\mathbb{E}_X(h(X)) = \int_{\mathcal{S}} h(x)\pi(dx)$ where $X \sim \pi(.)$ and h is some known function and $\mathcal{S}$ is the support of the distribution. So long as we know that $\mathcal{S}_q$ the support of $q(.)$ is larger than $\mathcal{S}$, $\mathcal{S} \subseteq \mathcal{S}_q$, and that both $q()$ and $\pi()$ have densities w.r.t. the same dominating measure $\nu()$ we can write for $A \subseteq \mathcal{S}_q$

$$\mathbb{P}(X \in A) = \int_A \pi(x)\nu(dx) = \int_A \frac{\pi(x)}{q(x)} q(x)\nu(dx) \tag{3.2.1}$$

$$\mathbb{E}_\pi(h(X)) = \int_{\mathcal{S}} h(x)\pi(x)\nu(dx) = \int_{\mathcal{S}_q} h(x)\frac{\pi(x)}{q(x)} q(x)\nu(dx) \tag{3.2.2}$$

and by simulating n i.i.d. observations $x_1 \dots x_n$ from $q()$ we can estimate the expectation by

$$\hat{h} = n^{-1} \sum_{i=1}^{n} h(x_i) \frac{\pi(x_i)}{q(x_i)} \tag{3.2.3}$$

where $x_i \sim q()$. The convergence of $\hat{h} \rightarrow \mathbb{E}_\pi(h(X))$ is guaranteed by the law of large numbers. The random ratios $w_i = \pi(x_i)/q(x_i)$ are called the *importance weights,* their distribution is a function of the distribution of X, $W(X) = \pi(X)/q(X)$, and this governs the behaviour of the estimate. Clearly the expected value of $W(X)$ is 1 and a good choice of $q()$ will be such that the weights are close to 1 in some sense, while maintaining the condition on the support.

The impact of a poor choice of proposal is understood by experts and the impact of infinite variance is mentioned in most introductory texts, sometimes implying that the a.s. convergence with finite variance is sufficient to give an adequate estimator. We investigate the difficulties that can be experienced in more detail.

Even in simple pedagogical examples it is usually difficult to calculate the distribution exactly, however simulation can be used to examine it. The weights typically have extremely skewed distributions, in the case of finite variance the skewness is reduced by the averaging in equation 3.2.3, however in realistic cases it is often difficult to ensure that the variance of W is finite and more importantly it is difficult to control the tail behaviour.

3.2.1 Exponential target distribution

A simple analytically tractable example is where both the target and proposal distributions are negative exponential. The target is $\pi(x) = \lambda e^{-\lambda x}$ and proposal $q(x) = e^{-x}$, the proposal is good (heavier tail) if $\lambda \geq 1$ and poor for small λ. Defining the transformed random variable for the weight $W = \lambda e^{X(1-\lambda)}$ the distribution of W when $X \sim q(.)$ can be calculated and has p.d.f.

$$f_W(w) = \exp(\log(w/\lambda)/(\lambda - 1))/w|(1-\lambda)| \propto w^{1/(\lambda-1)-1}$$

where W takes values in $(0, \lambda]$ if $\lambda > 1$ or $[\lambda, \infty)$ if $\lambda < 1$.[1]

The moments are readily calculated as $\mathbb{E}(W^p) = \lambda^p/(1 + p(\lambda - 1))$ for $\lambda \geq 1$ and $p \geq 0$, for $\lambda < 1$ the moments only exist for $p < 1/(1-\lambda)$ and so we can see that for $\lambda \leq .5$ the variance is infinite. The densities for a range of λ are plotted in figure 3.2.1, remembering that all have mean 1 it should be noted that all the distributions are skewed except for $\lambda = 2$ and have a very appreciable skew except when λ is close to 2. This suggests that in more complex situations where the form and scale of $\pi(.)$ are unknown the weights should usually be assumed to be highly skewed and efforts made to ensure finite variance.

[1] If $\lambda = 1$ then W =1 and $\mathbb{E}(W^p)$ =1

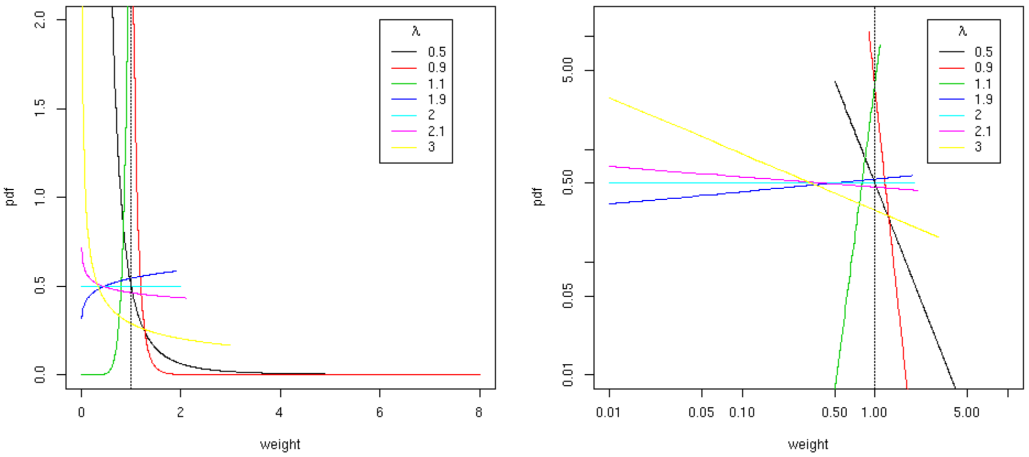


Figure 3.2.1: Density of importance sample weights for exponential target and proposal. (linear scale on left, log scale on right)

3.2.2 Importance sampling in higher dimensions

In higher dimensions the performance of importance sampling will be reduced because the skewness will increase. For example in d dimensions if the target and proposal distributions are both of product form, $\pi(\mathbf{x}) = \prod_{i=1}^{d} f(x_i; \theta)$ and $q(\mathbf{x}) = \prod_{i=1}^{d} f(x_i; \theta_q)$, for some parametric p.d.f. f, the composite weight has the distribution of the product of d weights each with the distribution from the univariate weight. Only in the simplest cases are the products tractable, the case of exponentials considered above with $\lambda = 2$, which results from using a proposal with p.d.f. e^{-x} and target p.d.f. $2e^{-2x}$. This gives the weights a uniform distribution where $W \sim \mathrm{U}(0, 2)$ which often might be regarded as a good proposal. We use this to illustrate that a proposal with good performance in one dimension will often have significantly reduced performance in higher dimensions. The density of an individual product term is $f_W(w) = 2^{-d} \log\left(2^d/w\right)^{d-1}/(d-1)!$ for $w \in [0, 2^d]$ and the moments are $\mathbb{E}(W^p) = (2^p/(p+1))^d$. Although the increase in variance with d is obvious, the serious impact on the importance sampler is less obvious and investigated below.

The distribution of the sample mean and variance from samples of size n can easily be simulated, results of a large simulation for $n = 10^8$ are shown in table 3.2.1, of particular note is the large variability in the sample standard deviation, also of more significance are the columns `ngt1, ngtmu` which are the number of samples > 1 and $> \bar{W}$ (the sample mean) showing the extreme skewness as the dimension

	sample	sample					
d	mean	s.d.	50%	99%	100%	ngt1	ngtmu
2	1.0000	0.88	7.468e−01	3.448e+00	3.999e+00	40340000	40340000
3	0.9999	1.17	5.518e−01	5.171e+00	7.983e+00	34480000	34490000
4	1.0000	1.47	4.069e−01	7.022e+00	1.568e+01	30200000	30200000
6	0.9999	2.15	2.206e−01	1.074e+01	5.768e+01	24020000	24020000
8	0.9999	2.99	1.195e−01	1.400e+01	1.716e+02	19610000	19610000
12	0.9989	5.51	3.505e−02	1.796e+01	1.195e+03	13620000	13630000
16	0.9998	9.88	1.028e−02	1.837e+01	5.663e+03	9754000	9754000
20	1.0020	17.53	3.015e−03	1.612e+01	2.041e+04	7106000	7100000
30	1.0050	79.40	1.401e−04	7.783e+00	4.216e+05	3367000	3359000
40	0.9982	229.60	6.521e−06	2.595e+00	1.346e+06	1656000	1657000
50	0.9389	450.30	3.031e−07	6.865e−01	3.152e+06	832700	859000
60	2.6720	17670.00	1.407e−08	1.532e−01	1.766e+08	423400	257900
70	1.3100	4544.00	6.550e−10	3.025e−02	4.322e+07	219300	192400
80	0.4636	688.80	3.046e−11	5.396e−03	3.782e+06	114000	163200
90	0.3591	643.60	1.415e−12	8.867e−04	4.225e+06	59850	95790
100	0.3096	856.80	6.579e−14	1.352e−04	6.532e+06	31520	54030
120	0.0382	62.05	1.423e−16	2.713e−06	3.157e+05	8688	36840

Table 3.2.1: Simulated product of d U$(0, 2)$ weights, $n = 10^8$, sample mean, s.d. and quantiles . (The mean and standard deviation are also plotted in figure 3.2.2)

increases.

The sample mean and standard deviation are plotted together with a set of simulations each of 10^5 in figure 3.2.2. We know the sample mean has mean 1 and variance $((4/3)^d - 1)/n$, so taking a variance of 1 as a necessary threshold we need $n > (4/3)^d$ and we obtain

d	5	10	20	30	40	50	60	80	100	120
n	4.2	17.8	315	5600	1e5	1.7e6	3.1e7	9.9e+9	3.1e+12	9.8e+14 .

For d below 20 the simulated results on standard deviation indicate that 10^5 samples would be sufficient for most applications, for d between $30 - 50, 10^8$ would be required and for $d \geq 60$ even 10^8 is insufficient.

An alternative approach to understanding the distribution of these weights in high dimensions is to consider the transformed variable $Y = \log(2^d/W) = d\ \log(2) - Z$ then Y has a standard gamma distribution with shape parameter d, and scale parameter 1, which can be used to evaluate quantiles of W, which are in agreement with the sampled values in table Table 3.2.1 on page 71. Also as d increases the distribution of Y approaches a normal distribution and so W approaches a log-normal distribution.

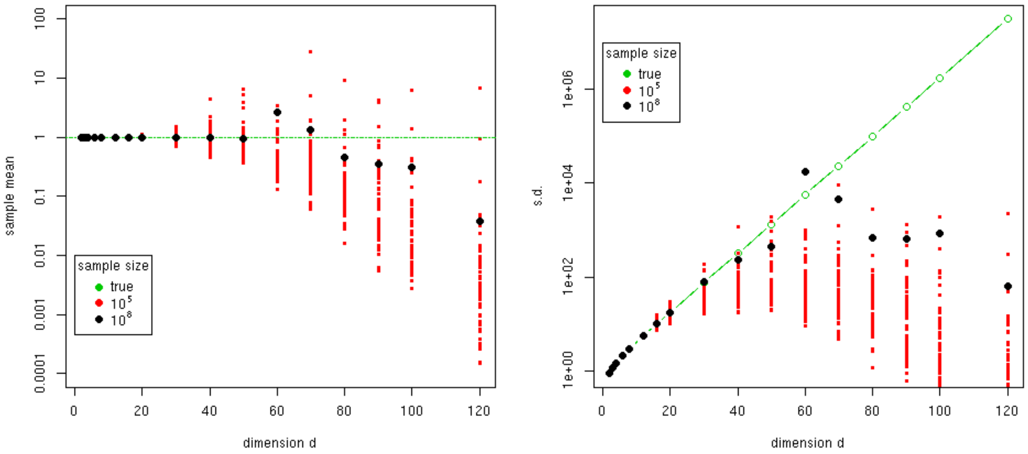


Figure 3.2.2: Simulated product of d U$(0, 2)$ weights, sample mean (left), sample standard deviations (right). Comparison of sample size 10^5 and 10^8 with true value.

3.2.3 Conclusions on Importance Sampling

Although the impact of a poor proposal distribution on importance sampling is well known, this study illustrates the problem and highlights the difficulties that can arise even in well understood situations and with variance known to be finite, in particular in higher dimensions. The increasing error in the sample standard deviations shown in figure 3.2.2 shows that examination of the variance of weights is not sufficient to guarantee acceptable behaviour. Better methods for robustly diagnosing poor performance from analysis of samples would be very useful.

3.3 Markov Chain Monte Carlo

Since the introduction of the first MCMC techniques to the statistics community by Hastings (1970) they have been extended and since (Gelfand and Smith, 1990) they have enjoyed widespread use. The basic concept is easily stated but sometimes difficult to apply successfully: in order to sample from a target density $\pi(x)$ on $\mathcal{X}$ with a dominating measure ν a Markov chain on $\mathcal{X}$ is constructed which has a transition kernel density $\mathcal{K}(x,.)$ with invariant density $\pi(x)$. A long sample $x_1 \dots x_n$ from this will provide correlated samples from $\pi(x)$. The kernel is constructed so that an ergodic theorem applies which justifies using the samples to estimate expectations of known functions $h()$, $\mathbb{E}_\pi(h(X)) = \int_\mathcal{S} h(x)\pi(x)\nu(dx)$ by the estimate $\hat{h} = n^{-1}\sum_{i=1}^{n} h(x_i)$. In general it is necessary to prove that the kernel is irreducible, aperiodic and has the correct invariant distribution. A key result is theorem 1 of Roberts and Smith (1994) reproduced here as theorem 5, which requires the definition of ψ-irreducible.

Definition 4. A Markov chain is ψ-irreducible if there exists a non-zero measure ψ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\psi(A) > 0$ and for all $x \in \mathcal{X}$, there exists a positive integer n such that $\mathcal{K}^n(x,A)$>0.

Theorem 5. *If $\mathcal{K}$ is ψ-irreducible and aperiodic then, for all $x \in \mathcal{S}$,*

1. $|\mathcal{K}_x^n - \pi| \rightarrow 0$ *as* $n \rightarrow \infty$

2. *for real valued,π-integrable h,*

$$n^{-1}\sum_{i=1}^{n} h(X_i) \rightarrow \int h(x)\pi(x)\nu(dx) \quad a.s. \text{ as } n \rightarrow \infty.$$

A sufficient condition is that it satisfies the detailed balance conditions

$$\pi(x)\mathcal{K}(x,x') = \pi(x')\mathcal{K}(x',x) \tag{3.3.1}$$

which also ensure the chain is reversible.

See standard texts such as (Robert and Casella, 2004) for further details.

Construction of a chain that has the correct invariant distribution and converges is simplified by the use of standard basic algorithms. The two main classes of basic algorithms are the Metropolis-Hastings and Gibbs which are described below and form the basis of a wide range of more advanced techniques. It is necessary to choose a value of n that ensures that $\hat{h}$ is sufficiently close to h, heuristic methods are still necessary in most practical situations. Visual examination of plots of components of the chain and their auto-correlation are often sufficient to identify poor choices in the construction of the kernel. Study of rates of convergence of chains in increasingly realistic situations is still an active area of research, an important conclusion is that, as in importance sampling, it is necessary for the proposal to have heavier tails than the target, see for example (Jarner and Roberts, 2007) and the references therein.

3.3.1 Metropolis-Hastings

The basic Metropolis-Hastings MCMC sampler (Hastings, 1970) (MH) is particularly useful in Bayesian statistics as it only requires knowledge of the target density up to an unknown constant of proportionality. A Bayesian posterior of θ given data $\mathbf{y}$ denoted $\pi(\theta|\mathbf{y})$ is given by $\pi(\theta|\mathbf{y}) \propto L(\mathbf{y};\theta)p(\theta) := \pi_u(\theta|\mathbf{y})$ where p is the prior, L the likelihood and π_u the un-normalised density. That is $\pi_u = C\pi$ where C is an unknown constant independent of θ, subsequently we generally use π without the suffix u unless we need to emphasise the difference and also generally do not show the dependence on data and in this chapter use $\pi(x)$ for a generic target which would often in practice be of the type $\pi_u(\theta|\mathbf{y})$.

The Metropolis-Hastings algorithm requires the choice of an initial distribution q_0 (which can be a constant) and a proposal distribution $q(.|x)$ which in general will depend on the current state x,

Algorithm 3.1 Metropolis-Hastings

```
1. Initialise by sampling $X_0$ from $q_0(.)$

2. for $t = 1, 2, \ldots n$ repeat steps 3-5

3. sample a proposed value $X' \sim q(.|x_{t-1})$

4. compute the acceptance ratio $\mathcal{A}^a$ by

        $\mathcal{A}(X'|x_{t-1}) = \frac{\pi(X')q(x_{t-1}|X')}{\pi(x_{t-1})q(X'|x_{t-1})}$        (3.3.2)

5. set $X_t = X'$ with probability $\mathcal{A}(X'|x_{t-1}) \wedge 1$
   else set $X_t = x_{t-1}$
```

[a] note we use $\mathcal{A}$ for the acceptance ratio, not the acceptance probability

the probability of the state remaining unchanged is called the rejection probability

$$\mathfrak{r}(x) = 1 - \int q(y|x) \min(\mathcal{A}(y|x), 1) dy$$

and we use the term acceptance probability for the average acceptance rate from x, $\mathfrak{a}(x) = \int q(y|x) \min(\mathcal{A}(y|x), 1) dy = 1 - \mathfrak{r}(x)$.

The resulting transition kernel can be written as

$$\mathcal{K}(x, y) = q(y|x) \min(\mathcal{A}(y|x), 1) + \delta_x(y)\mathfrak{r}(x).$$

Conditions for the resulting transition kernel $\mathcal{K}$ having the correct invariant distribution are given by Theorem 3 of Roberts and Smith (1994) showing that the convergence properties of $\mathcal{K}$ are inherited from the proposal q the theorem is

Theorem 6.

1. *If q is aperiodic, or $P(X_t = X_{t-1}) > 0$ for some $t \geq 1$, then the Metropolis-Hastings algorithm is aperiodic.*
2. *If q is ψ-irreducible and $q(x, y) = 0$ if and only if $q(y, x) = 0$ then the Metropolis-Hastings algorithm is ψ-irreducible.*

An analysis of rates of convergence is given by Roberts and Tweedie (1996). The choice of proposal is still wide and two particular choices the independence and the random walk are popular because of their simple implementation, they are often adequate for simple problems but can struggle in multi-modal or high dimensional situations. An important consideration in their use remains that of the choice of scaling parameters which is described in section 3.3.3.

Independence Metropolis-Hastings

If the proposal q does not depend on the current value, the algorithm is called the independence Metropolis-Hastings and the acceptance ratio can be written as

$$\mathcal{A}(X'|x_{t-1}) = \frac{\pi(X')/q(X')}{\pi(x_{t-1})/q(x_{t-1})} \tag{3.3.3}$$

which shows a close connection with the independence sampler described in section 3.2. Again the choice of q close to π, but with heavier tails, is essential and the performance can degrade significantly in higher dimensions.

Random walk Metropolis-Hastings

The random walk Metropolis-Hastings (MH) sampler is applicable to many situations, in particular when the support of π, $\mathcal{S}$ is a connected subset of $\mathbb{R}^d$. Use on other spaces for example subsets of $\mathbb{Z}^d$ depends on the target distribution having some smoothness with respect to a metric on $\mathcal{S}$. The simplest form uses symmetric proposals where $q(x'|x) = q(x|x')$ which gives the original Metropolis algorithm. The acceptance ratio now simplifies to

$$\mathcal{A}(X'|x_{t-1}) = \frac{\pi(X')}{\pi(x_{t-1})}. \tag{3.3.4}$$

The simplest symmetric form is the random walk $q(x'|x) = f(x' - x)$ where f is a density on $\mathcal{X}$ w.r.t. the dominating measure, or often $q(x'|x) = f(|x' - x|)$ where f is a density on $\mathbb{R}_+$. Valid, but not necessarily optimal, choices for f include any distribution with support larger than $\mathcal{S}$ or when $\mathcal{S}$ is connected a much wider choice. Common choices in $\mathbb{R}^d$ include multivariate normal, t, or uniform distributions, the choice of scaling parameters is still an issue which can have significant effects on the convergence rates and so on the length of chain that must be simulated, this is discussed further in section 3.3.3.

Boundaries of support of target

When the support is bounded some random walk proposals will be outside the support, the simplest approach is to use them and the resultant value of $\mathcal{A} = 0$, if π has large values close to the boundary, such as a gamma distribution with shape parameter < 1, then the closely related approaches of using a log transform of X or using a log-normal proposal may be appropriate. The log-normal proposal is not symmetric but the ratio $q(x|x')/q(x'|x)$ simplifies to x'/x and so calculation of the proposal density is not needed.

3.3.2 Gibbs sampler

The Gibbs sampler, which also originated in physics and has had a long history in statistics , is used in two situations, the original is where the conditional distributions can be sampled exactly and is often used in high dimensions. The second is where other Markov kernels are combined within a Gibbs framework, typically Metropolis kernels giving the "Metropolis within Gibbs" sampler. The basic algorithm on a space $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ samples each dimension i either regularly in a specified order or in a random order. At each step $X_{t+1} \sim \pi(X_i|X_1, \ldots X_{i-1}, X_{i+1}, \ldots X_d)$, a key result is theorem 2 of Roberts and Smith (1994) which is

Theorem 7. *If ν is n-dimensional Lebesgue measure, $n \geq d$, π is lower semi continuous at 0, $\int \pi(x)dx_i$ is locally bounded for $i = 1, \ldots, k$, and $\mathcal{S}$ is connected, the results of Theorem 5 apply to the Gibbs kernel.*

3.3.3 Scaling and Adaptation of Proposal distributions

The choice of proposal distribution for the MH algorithm as either a heavy tailed independence sampler or a random walk is one aspect of the choice. An equally important problem is that of the choice of scaling for the proposal. As shown in section 3.2 the variance of importance samplers can be large or infinite even for matching parametric families and in higher dimensions the performance degrades exponentially unless the scaling of the proposal is a close match to the target. As most target distributions are Bayesian posteriors the choice of scaling for an independence sampler will be problem specific. When the popular choice of a symmetric random walk proposal is used (RWM) it has been known since the original paper by Metropolis that the choice of scaling has a very significant effect on the rate of convergence of the algorithm. In particular if proposals are all small then most moves will be accepted but movement around the target space $\mathcal{X}$ will be slow, on the other hand if proposals are too large then most proposals will be rejected and the chain will remain for long periods at particular values.

Asymptotic results on the choice of scaling parameters for MCMC proposal kernels have been studied by several authors including Roberts and Rosenthal (2001) and Bédard (2008) in situations amenable to analytic investigation. These studies have provided useful guidance to the choice of scaling, mainly in terms of acceptance rates, 0.234 is close to 'optimal' for many RWM problems. Practically one or more pilot runs are performed with a range of scaling parameters after which a value is chosen for a long run. An alternative approach is to use adaptive methods which automatically adjust the scaling to match the target, care is needed with these methods for both theoretical and practical reasons. A recent summary of the closely linked fields of optimal scaling and adaptive MCMC is given by Rosenthal (2011) with a useful "Frequently asked questions" section which gives a concise summary of many results and much experience. The theoretical problems are now largely understood, however practically there are still issues with the use of adaptive methods on multimodal distributions as the adaption can often lead to only one mode being explored.

Many posteriors that arise in complex models have highly discontinuous multimodal distributions, often with disconnected support. A challenge is to transfer the existing results on scaling and adaptive methods to these distributions, for instance are the same acceptance rates optimal ? An example of such a posterior arises in the partially observed GSE, which is discussed in more detail in chapter 4 along with the more challenging posterior for the Indian buffet epidemic.

3.4 Marginal MCMC

In many situations a hidden data model is either the natural representation or extending the model with augmented data provides a powerful tool for inference. In particular inference for epidemics where the infection times are unobserved provides such an example where usually the primary interest is in the distribution of parameters rather than the conditional distribution of the number of infectives. This is considered in detail in chapter 4 where the GIMH and MCWM algorithms are applied. A generic hidden data model with parameters $\theta \in \Theta$ is that $\mathbf{X} \in \mathcal{X}$ is unobserved or augmented data and $\mathbf{Y} \in \mathcal{Y}$ is observed data with a joint distribution that has a natural factorisation $\pi(\mathbf{y}, \mathbf{x}, \theta) = \pi(\mathbf{y}|\mathbf{x}, \theta)\pi(\mathbf{x}|\theta)p(\theta)$ where p is the prior[2]. When interest is in the marginal posterior $\pi(\theta|\mathbf{y})$, rather than the joint posterior $\pi(\theta, \mathbf{x}|\mathbf{y})$ and the marginal is both intractable and difficult to sample from then the pseudo marginal algorithms of Andrieu and Roberts (2009) are often an appropriate choice.

3.4.1 Full posterior approach

Most previous approaches target $\pi(\theta, \mathbf{x}|\mathbf{y})$ and then marginalise by ignoring $\mathbf{x}$ in the samples obtained from the Markov chain. A standard approach is to use a deterministic scan Gibbs sampler at the top level on $\mathbf{x}, \theta$, often the exact distributions $\pi(\theta|\mathbf{X}, \mathbf{Y})$ and $\pi(\mathbf{X}|\theta, \mathbf{Y})$ are unavailable and so MH steps are used. The resulting algorithm is given in algorithm 3.2.

3.4.2 GIMH

The grouped independence Metropolis-Hastings (GIMH) algorithm was introduced by Beaumont (2003) in a genetics context, and is described by Andrieu and Roberts in terms of the marginal $\pi(\theta)$ of $\pi(\theta, Z)$. An importance sampler estimate $\tilde{\pi}^N(\theta)$ of $\pi(\theta)$ is used within a Metropolis-Hastings step, the justification for this is given in more general terms in section 3.5, in particular in their terms $\tilde{\pi}^N(\theta) = \sum_{i=1}^{N} \pi(\theta, Z_i)/q_\theta(Z_i)$

[2] we use p rather than π to distinguish prior from target

Algorithm 3.2 MH within Gibbs algorithm for hidden data

```
The target is $\pi(\theta,\ \mathbf{X}|\mathbf{y})$ each outer step repeats these steps
    1. MH sample of $\theta|\mathbf{X},\mathbf{y} \propto \pi(\mathbf{X},\mathbf{y}|\theta)p(\theta)$ by
    2. Propose $\theta'$ from $q(\theta'|\theta)$
    3. Accept $\theta'$ with probability $\min(\mathcal{A},1)$ where
           $\mathcal{A} = \frac{\pi(\mathbf{X},\mathbf{Y}|\theta')p(\theta')\,q(\theta|\theta')}{\pi(\mathbf{X},\mathbf{Y}|\theta)p(\theta)\,q(\theta'|\theta)}$
    4. MH sample of $\mathbf{X}|\theta,\mathbf{y} \propto \pi(\mathbf{X},\mathbf{y}|\theta)$ by
    5. Propose $\mathbf{X}'$ from $q(\mathbf{X}'|\mathbf{X})$
    6. Accept $\mathbf{X}'$ with probability $\min(\mathcal{A},1)$ where
           $\mathcal{A} = \frac{\pi(\mathbf{X}',\mathbf{Y}|\theta)\,q(\mathbf{X}|\mathbf{X}')}{\pi(\mathbf{X},\mathbf{Y}|\theta)\,q(\mathbf{X}'|\mathbf{X})}$
```

where $Z_i \sim q_\theta(.)$ in the case we are considering Z is identical to $\mathbf{X}$ and the full posterior is

$$\pi(\theta, Z) = \frac{\pi(\theta,\mathbf{X},\mathbf{Y})}{\pi(\mathbf{Y})} \propto \pi(\mathbf{Y}|\theta,\mathbf{X})\pi(\mathbf{X}|\theta)p(\theta)$$

the constant $\pi(\mathbf{Y})$ cancels in the calculation of the acceptance ratio and the estimate $\tilde{\pi}^N(\theta)$ becomes

$$\tilde{\pi}^N(\theta) = \sum_{i=1}^{n_z} \frac{\pi(\mathbf{y}|\theta,\mathbf{x}_i)\pi(\mathbf{x}_i|\theta)p(\theta)}{q_\theta(\mathbf{x}_i)} \tag{3.4.1}$$

where the $\mathbf{x}_i$ are n_z values i.i.d. $\sim q_\theta(.)$ [3].

A simple (but rarely optimal) choice for $q_\theta(.)$ is $\pi(\mathbf{X}|\theta)$ which is often easy to sample from, in this case the calculation of $\tilde{\pi}^N(\theta)$ simplifies as $q_\theta(.)$ cancels and this also speeds the calculation, especially when $\pi(\mathbf{X}|\theta)$ is expensive to calculate.

In chapter 4 the GIMH is applied to the Indian buffet epidemic.

A common problem with the GIMH algorithm is that the Markov chain can get stuck with a very small probability of moving, this happens when the estimate $\tilde{\pi}^N(\theta)$ is much larger than $\pi(\theta)$ and so also larger than $\pi(\theta')$ for nearly all proposed θ'. Although the eventual escape and convergence to the exact target is guaranteed as the number of steps → ∞ it may require an unacceptable time to do so. The

[3] Their N has been replaced with n_z to avoid confusion with other n's and N's

solution is to improve the estimate of $\pi(\theta)$, the obvious approach of increasing n_z in generating $\tilde{\pi}^N(\theta)$ will help, but if $q_\theta(.)$ is such that the weight distribution has the typical heavy tail n_z, then the results from section 3.2 show that n_z would have to be increased exponentially to achieve the desired improvement. An analysis of the reasons for the sticking of the chain is presented below in section 3.5. The analysis is given in more general terms and to distinguish the original algorithm from the generalised algorithm the name stochastic exact Metropolis-Hastings (SEMH) is introduced.

3.4.3 MCWM

A closely related approximate algorithm was introduced by O'Neill et al. (2000) the Monte Carlo within Metropolis algorithm (MCWM), which is also analysed and generalised by (Andrieu and Roberts, 2009). In the absence of a better proposal $q_\theta(.)$ this provides an alternative which in general does not suffer from sticking but has a bias that in general is not known. The generalisation of MCWM is called the Stochastic Approximate Metropolis-Hastings algorithm (SAMH), and the bias is analysed in section 3.5.4.

Effect of Limited Support on the Proposal

In complex hidden data models it can be difficult to know the support of $\mathbf{X}$ given $\mathbf{y}$ and so proposals will generate values of $\mathbf{x}$ that give zero likelihood, this is not a problem for standard MCMC or the GIMH algorithms as these proposals are rejected, the only effect is to reduce the acceptance rate. However in some of the examples considered in chapter 4 this happened sufficiently often that the probability of all n_z importance samples being zero and hence $W = 0$ in one or more steps of a long run was significant. MCWM when both $W = 0$ and $W' = 0$ requires specifying the behavior as accepting with probability $p_0 \in [0, 1]$, the subsequent examples all used $p_0 = 0$.

3.4.4 MCWM bias

Simulations have been performed to illustrate the bias in the MCWM algorithm as the dimension d increases, the target chosen has the same dimension d for θ and Z with a known marginal, $\pi(\theta)$ multivariate normal $\sim N(0_d, 1_d)$ and $\pi(Z|\theta) \sim N(\theta, 1_d)$. In order to compare the d-dimensional simulated distributions with the true values we compare the distance from the origin, which is χ^2, and plot the sample median against the expectation. The points shown in figure 3.4.1 are for increasing dimension and 3 values of n_z=10,20,40.

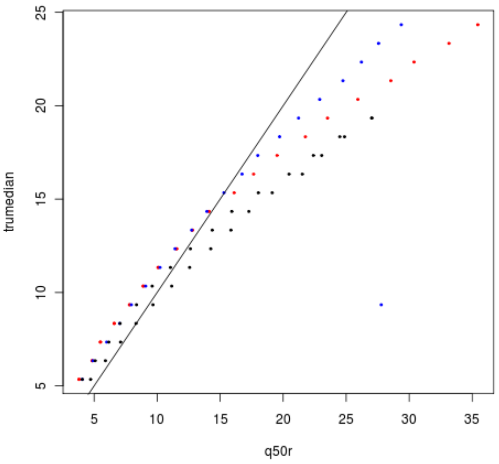


Figure 3.4.1: Bias of MCWM multivariate normal, sample median vs true median $n_z = 10$ (black), 20 (red), 40 (blue)

3.5 Stochastic Exact Metropolis-Hastings Algorithm

A recently discovered aspect of the MH sampler is that if the calculation of the target density $\pi(x)$ is replaced by an estimate $\tilde{\pi}(x)$ and used in the MH algorithm, subject to some conditions a valid algorithm results which still has an exact invariant density π. The first algorithm to use this technique was the grouped independence Metropolis-Hastings (GIMH) algorithm Beaumont (2003) which was analysed further and generalised by Andrieu and Roberts (2009), who give detailed convergence results. A simplified presentation of their results in a more general form is given by Wilkinson (2011) on his blog, which inspired this description and study. This description separates the analysis of the GIMH into components, the Stochastic exact Metropolis-Hastings described here, which uses an importance sampler as described above and the integration of these into the GIMH which is described in section 3.4. In other situations alternative estimates $\tilde{\pi}(x)$ could be used in place of the importance sampler but these are not considered here.

The motivating situation is the GIMH described below, where $\tilde{\pi}$ is an importance sampler for π which is an intractable marginal distribution. More generally we consider for any $x \in \mathcal{X}$ a stochastic estimate $\tilde{\pi}(x)$ of $\pi(x)$ and define

$W(x) = \tilde{\pi}(x)/\pi(x)$, which is a random variable for each x, W may be discrete or continuous and have a finite or infinite support. We also require that $\mathbb{E}(W(x)) = c$ where $c > 0$ is a constant independent of x. The Stochastic exact Metropolis-Hastings algorithm simply replaces π with $\tilde{\pi}$ in the acceptance ratio (equation 3.3.2) of the standard MH algorithm giving

$$\mathcal{A}(X'|x_{t-1}) \quad = \quad \frac{\tilde{\pi}(X')q(x_{t-1}|X')}{\tilde{\pi}(x_{t-1})q(X'|x_{t-1})}. \tag{3.5.1}$$

The key to understanding the algorithm is the observation by Beaumont that this defines a Markov chain of (X, W) on $\mathcal{X} \times \mathcal{W}$, $\mathcal{W} \subset \mathbb{R}$ where W is not directly observed, this is encapsulated in:

Lemma 3. *The Metropolis-Hastings algorithm using equation 3.5.1 has an invariant distribution $\pi(x, w) = \pi(x)wf_W(w|x)$ on $\mathcal{X} \times \mathcal{W}$, $\mathcal{W} \subset \mathbb{R}$, where $W = \tilde{\pi}(X)/\pi(X)$ is not directly observed and $f_W(.|x)$ denotes the density of $\tilde{\pi}(x)/\pi(x)$ w.r.t. an appropriate measure on $\mathcal{W}$.*

Proof. This follows from rewriting $\tilde{\pi}(X') = \pi(X')W'$ so adding the dependence on W, equation 3.5.1 is replaced by

$$\mathcal{A}(X', W'|x_{t-1}, w) = \frac{\pi(X')W'q(x_{t-1}|X')}{\pi(x_{t-1})wq(X'|x_{t-1})} \tag{3.5.2}$$

and we can write this as

$$\mathcal{A}(X', W'|x_{t-1}, w) = \frac{\pi(X')W'f_W(W'|X') \;\; f_W(w|x_{t-1})q(x_{t-1}|X')}{\pi(x_{t-1})wf_W(w|x_{t-1}) \quad f_W(W'|X')q(X'|x_{t-1})} \tag{3.5.3}$$

which is the acceptance ratio for a MH Markov chain on $\mathcal{X} \times \mathcal{W}$ with target $\pi(X)Wf_W(W|X)$ and proposal density $f_W(W'|X')q(X'|x_{t-1})$, so the result follows from the standard MH result. □

Corollary 8. *The marginal distribution of X from the Markov chain on $\mathcal{X} \times \mathcal{W}$ is $\pi(X)\mathbb{E}(W|X)$ and as $\mathbb{E}(W|X) = c$, independent of X, the marginal distribution is $\pi(.)$.*

Proposition 2. *If $\tilde{\pi}(x)$ is a point-wise estimator of $\pi(x)$ such that $\mathbb{E}\left(\tilde{\pi}(x)/\pi(x)\right) = c$ where $c > 0$ is independent of x then the Metropolis-Hastings algorithm, using equation 3.5.2 as the acceptance ratio, has a stationary distribution $\pi(x)$ when q() satisfies the conditions for the standard algorithm.*

Proof. The conditions on $q()$ ensures that the basic chain on $\mathcal{X}$ is irreducible, W is chosen independently from $f_W(.|x)$ for each x and so the chain on $\mathcal{X} \times \mathcal{W}$ is irreducible. A similar argument shows it is aperiodic. The result follows from lemma 3 and corollary 8. □

When used within the GIMH algorithm the analysis will require the probability of remaining in state (w,x) or accepting a move from it, which is just an extension of the notation above to acknowledge the expanded state. We have $\mathfrak{r}(w,x)$ and $\mathfrak{a}(x,w)$

$$\mathfrak{r}(x,w) \quad = \quad 1 - \int_{\mathcal{X}} \int_{\mathcal{W}} \min(\mathcal{A}(x', w'|x, w), 1) f_W(w'|x') q(x'|x) dw' dx' \quad (3.5.4)$$

.

3.5.1 Conditional weight distribution

In analysing the performance of the algorithm it is necessary to consider the conditional distribution of W for a fixed x, as the invariant distribution is $\propto \pi(x_{t-1}) W f_W(W|x)$ which we know is a density because of the condition $\mathbb{E}(W) = c$. We note that the factor W ensures that this density has a heavier tail than W, which is as pointed out in section 3.2 is often already heavy tailed if $\tilde{\pi}$ is an importance sampler. Considering the importance sample weight distribution derived from the exponential distribution in section 3.2.1 as $f_W(w) =\propto w^{1/(\lambda-1)-1}$ on either $w \in (0, \lambda]$ or $w \in [\lambda, \infty)$ if $\lambda < 1$, so the invariant distribution is of similar form $\propto w^{1/(\lambda-1)}$. It is important to note from this example the general result that for the pth moment of $W f_W(W|x)$ to exist it is necessary for the (p+1) moment of $f_W()$ to exist.

Another example where this is readily studied is for W log-normal and so $f_W(w) = \frac{1}{w\sigma\sqrt{2\pi}} \exp(-\frac{(\log w - \mu)^2}{2\sigma^2}), \quad w > 0$ where μ, σ are possibly dependent on x but subject to $\mathbb{E}(W) = \exp(\mu + \sigma^2/2) = c$, the invariant distribution is $\propto \exp(-\frac{(\log w - \mu)^2}{2\sigma^2})$ which by noting $(\log w - \mu)^2 = (\log w - \mu - \sigma^2)^2 - 2\sigma^2 \log w + \text{constant}$ can be seen to again be log-normal with parameters $\mu + \sigma^2, \sigma$.

Examples of conditional weights

An alternative approach to deriving the same equations is to consider the Markov chain on $\mathcal{W}$ with proposal $f_W(.)$ and acceptance ratio $\mathcal{A} = w'/w$ which has the invariant density $\propto w f_W(w)$, this also provides a mechanism for understanding the behaviour. Sample density estimates from simulations, each of length 10^5, for a variety of weight distributions were compared with the exact invariant density and are shown in figure 3.5.1 those on the right are of the log weights which have a density $e^{2x} f_W(e^x)$.

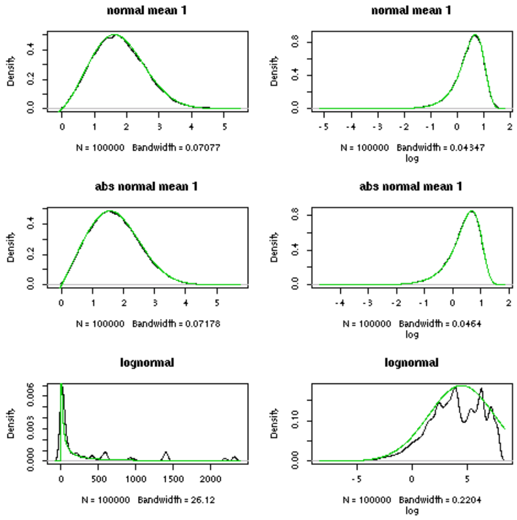


Figure 3.5.1: Simulations of weights in SEMH

3.5.2 Examples of Stochastic exact Metropolis-Hastings

Wilkinson gives simple pedagogical examples which demonstrate good performance with runs of length 10^4, however remembering that in the GIMH W will be from an importance sampler (section 3.2) and so very likely to have a highly skewed distribution. We investigate his example with a more realistic noise distribution, in particular we use his target $N(0,1)$ and a uniform $U[-.5,.5]$ proposal, but replace the noise with a log-normal distribution with parameters $\mu = -4.5, \sigma = 3$, which is more typical of the weights from an importance sampler, in particular it corresponds to a 9-dimensional problem with exponential target described in section 3.2.2. This gives the typical behaviour of GIMH getting stuck, see the left half of figure 3.5.2, the longer run on the right shows that in spite of the long stick periods, and the very high auto correlation it appears to be converging slowly.

3.5.3 Sticking of chain

All MH algorithms remain in the same state for periods, the distribution of lengths of these runs is geometric with the rejection probability $\mathfrak{r}(x)$ and so the expected number of consecutive repeats of x is $\mathfrak{r}(x)/(1-\mathfrak{r}(x))$. In the Stochastic exact Metropolis-Hastings as the state space is extended to include W which is unknown, on entry

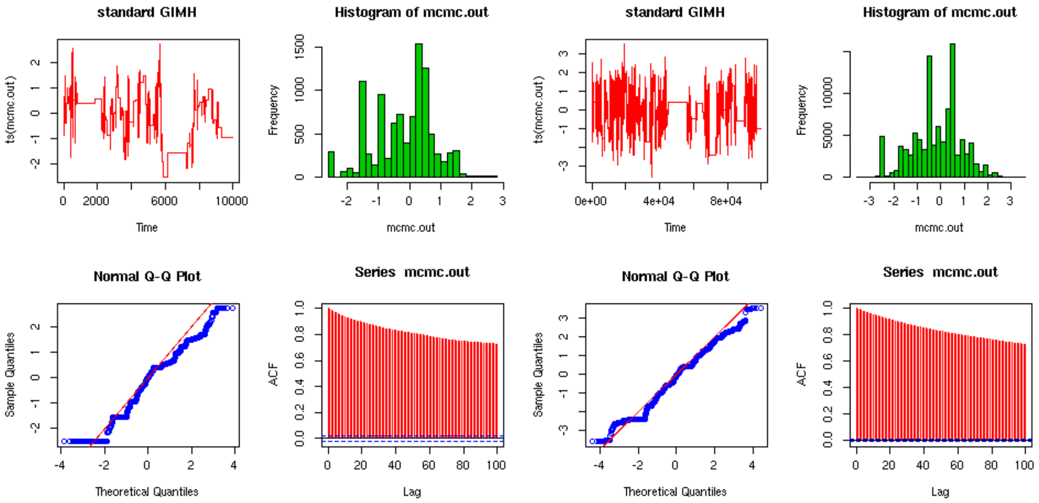


Figure 3.5.2: Stochastic exact Metropolis-Hastings example showing typical GIMH “sticking”

to a state (x, w) the rejection probability is sampled from $\mathfrak{r}(x, W)$. Although each W is sampled from $f_W(.|x)$ the accepted states at equilibrium have the conditional weight distribution $wf_W(w|x)/\mathbb{E}(W)$. The description is simplified by assuming for the remainder of section 3.5 that w.l.o.g. $\mathbb{E}(W) = 1$. First we introduce a term acceptance bound which we use below in the analysis of the SEMH , it is applicable to the analysis of any MH chain.

Definition 5. acceptance bound $\zeta(x) = \int \frac{\pi(y)}{\pi(x)} q(x|y)dy$ which is an upper bound on $\mathfrak{a}(x)$.

Lemma 4. $\mathfrak{a}(x) \leq \zeta(x)$

Proof. follows from the definitions

$$\mathfrak{a}(x) = \int q(y|x) \min(\mathcal{A}(y|x), 1)dy \leq \int q(y|x)\mathcal{A}(y|x)dy = \int \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)} q(y|x)dy.$$

□

We note that for the independence MH sampler $\zeta(x) = q(x)/\pi(x)$ which relates ζ to the well known requirement for proposals to have heavier tail.

To understand the behaviour, initially we consider the limiting case of concentrating the proposal on the current value, so replacing $q(x'|x)$ with $\delta_x(x')$ in equations 3.5.2 and 3.5.4 and taking expectations we get

$$\mathfrak{r}_\delta(x,w) = 1 - \int_{\mathcal{W}} \min(\frac{w'}{w},1) f_W(w'|x)dw' \tag{3.5.5}$$

and as $\mathcal{A} = w'/w$ and recalling that the invariant distribution of W has density $wf_W(w|x)$ (a density as $\mathbb{E}(W) = 1$)

$$\begin{aligned}\mathbb{E}(\mathfrak{r}_\delta(x,W)) &= \int w f_W(w|x)\mathfrak{r}_\delta(x,w)dw \\ &= 1 - \int_{\mathcal{W}}\int_{\mathcal{W}} w\min(\frac{w'}{w},1) f_W(w'|x) f_W(w|x)dw'dw \quad (3.5.6)\end{aligned}$$

this can be evaluated for some weight distributions, see section 3.5.3. However it is the tail behaviour that has a significant effect on the performance of the algorithm for realistic lengths of chain. If the expected rejection probability $\mathbb{E}(\mathfrak{r}(x,W))$ is close to one for a value of x that is sampled then a long sequence of repeated values will result with high probability, so informally for an acceptable chain it is necessary that $\mathfrak{r}(x,W)$ is only close to 1 for a set of x that has a very low probability of being sampled, this result is formalised in Theorem 8 of Andrieu and Roberts (2009) to give conditions for geometric convergence.

Bounds on expected time in a state

We can obtain bounds on the expected time of remaining in a state x after entry, denoting the number of time steps that the chain remains at x after entry by $N_b(x)$.

Proposition 3. $\mathbb{E}(N_b(x)) \geq \mathbb{E}(W_x^2) - 1$

Proof. We have □

$$\mathbb{E}(N_b(x)) = \int w f_W(w|x)\frac{\mathfrak{r}(x,w)}{1-\mathfrak{r}(x,w)}dw = \int w f_W(w|x)\frac{1}{1-\mathfrak{r}(x,w)}dw - 1$$

which as $1 - \mathfrak{r}(x,w) \leq \int_{\mathcal{W}} \frac{y}{w} f_W(y|x)dy = 1/w$ gives the result.

Remark 1. It is necessary for W_x to have finite variance for all x with $\pi(x) > 0$ to ensure the chain does not get stuck.

Comparison of the bound with exact calculation for selected weight distributions (section 3.5.3) shows that the bound is close for large variance.

Acceptance rates for selected weight distributions

The mean acceptance probability on entry to a state x, w is for a $\delta_x(.)$ proposal

$$a(w) = \int_{\mathcal{W}} \min(\frac{y}{w}, 1) f_W(y) dy = \int_0^w \frac{y}{w} f_W(y) dy + 1 - F_W(w) \tag{3.5.7}$$

for some distributions of weights this can be calculated analytically. In particular for the exponential IS weights and log-normal considered in section 3.2.

1. For the exponential ratio

$$F_W(w) = \begin{cases} (w/\lambda)^\xi & \text{on } [0, \lambda] \ \lambda > 1 \\ 1 - (w/\lambda)^\xi & \text{on } [\lambda, \infty] \ \lambda < 1 \end{cases}$$

where $\xi = 1/(\lambda - 1)$ straightforward integration yields the acceptance rate

$$a(w) = \begin{cases} 1 - (w/\lambda)^\xi/(1+\xi) & \lambda > 1 \\ 1/w + (w/\lambda)^\xi/(1+\xi) & \lambda < 1 \end{cases}$$

noting that the $(1 + \xi)$ term is of opposite sign in the two cases.

2. For the log-normal $f_W(w) = \frac{1}{w\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln w - \mu)^2}{2\sigma^2}\right)$

$$a(w) = w^{-1} \int_0^w \frac{1}{\sigma\sqrt{2\pi}} \left(-\frac{(\ln w - \mu)^2}{2\sigma^2}\right) dy + 1 - \Phi((\log(w) - \mu)/\sigma)$$

substituting $x = (\log(y) - \mu)/\sigma$ in the integral, it becomes

$$(2\pi)^{-1/2} \int_{-\infty}^{(\log(w)-\mu)/\sigma} \exp(\mu + \sigma x - \frac{x^2}{2}) dx$$

completing the square and noting that for $\mathbb{E}(W) = 1$ $\exp(\mu + \sigma^2/2) = 1$ gives

$$a(w) = w^{-1}\Phi(-\sigma + (\log(w) - \mu)/\sigma) + 1 - \Phi((\log(w) - \mu)/\sigma)$$

3. For the d-dimensional exponential ratio

$$f_W(w) = (-\log(w/2^d))^{d-1}/\Gamma(d) \quad w \in [0, 2^d]$$

transforming with $Y = \log(2^d/W)$ then Y has a standard gamma distribution with shape parameter d, denoting its c.d.f. by $F_\Gamma(.)$

$$a(w) = w^{-1}\int_{d\log(2)-\log(w)}^{\infty} y^{d-1}\exp(-2y + d\log(2))/\Gamma(d)dy + F_\Gamma(d\log(2) - \log(w))$$

$$a(w) = \frac{1}{w}(1 - F_\Gamma(2(d\log(2) - \log(w))) + F_\Gamma(d\log(2) - \log(w))$$

Comparison of empirical acceptance rates from simulations with the exact formulae compare well up to the largest weight in the samples.

Calculating expected "stick length" numerically

In order to compare the exact value of the expected stick length with the lower bound we need to evaluate $\int_0^\infty w f_W(w)\frac{1}{a(w)}dw$ where $f_W(.)$ and $a(w)$ are analytically known, numerical integration of this directly using `quadpack` via the R `integrate` routine[4] works well when the distribution of W is concentrated near one but can give numerical problems for realistic IS distributions. A log transform yields a stable integral $\int_{-\infty}^{\infty} e^{2x} f_W(e^x)/a(e^x)dx$, which has been used to calculate the exact values for $\mathbb{E}(N_b(x))$ numerically for the three weight distributions considered above for a range of parameters. The example of a log normal weight distribution is shown in figure 3.5.3, the results for the other distributions are very similar, showing a close approximation of $\mathbb{E}(N_b(x))$ by the variance for all cases where sticking may be of concern.

3.5.4 Stochastic approximate Metropolis-Hastings

A closely related approximate algorithm was introduced by O'Neill et al. (2000) the Monte Carlo within Metropolis algorithm (MCWM), which is also analysed and generalised by (Andrieu and Roberts, 2009). In terms of the Stochastic exact Metropolis-Hastings at each stage this resamples a new weight $W \sim f_W(.|x_{t_1})$ for the current state as well as $W' \sim f_W(.|X')$ for the proposed state, and then uses them both to calculate an acceptance ratio

[4] Based on QUADPACK routines `dqags` and `dqagi` by R. Piessens and E. deDoncker-Kapenga, see Piessens et al. (1983).

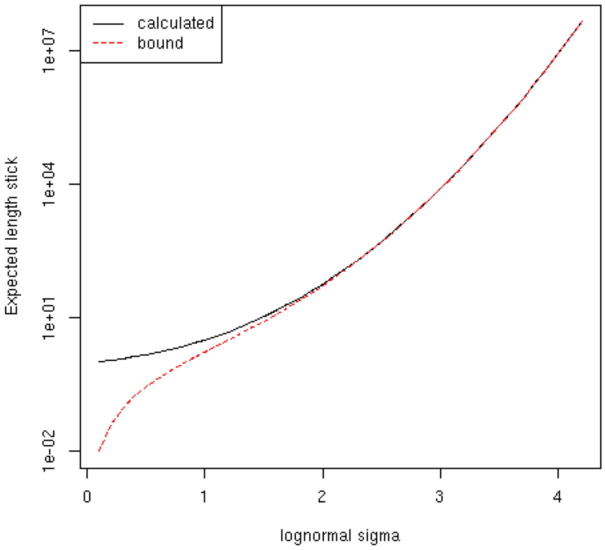


Figure 3.5.3: Comparison of exact calculation and lower bound for a log-normal weight distribution.

$$\mathcal{A}(X'|x_{t-1}) = \frac{W'}{W}\frac{\pi(X')q(x_{t-1}|X')}{\pi(x_{t-1})q(X'|x_{t-1})} \tag{3.5.8}$$

which is the same as equation 3.5.2. However this generates a Markov Chain on $\mathcal{X}$ with an unknown invariant distribution, whose bias is unknown both in size and type. We call this the stochastic approximate Metropolis-Hastings algorithm, reserving MCWM for the case where W results from an importance sampler. The exact invariant distribution can be calculated when both $\mathcal{X}$ and $\mathcal{W}$ are finite and discrete by constructing the transition matrix, some examples are given below. In the particular case considered by O'Neill et al. they are able to approximate the bias and introduce a bias correction. As motivation for this study of the possible biases the example from section 3.5.2 is repeated using this approximate algorithm, showing a distinct bias but apparently acceptable otherwise, see figure 3.5.4. This example illustrates that standard MCMC diagnostics will not reveal the presence or absence of significant bias.

The increasing bias in a multivariate normal case, as the dimension increases, of the standard MCWM, is described below in section 3.4.4.

We examine the bias in some particular cases and make some general conclusions, first we examine the kernel density estimates for simulations from a variety of

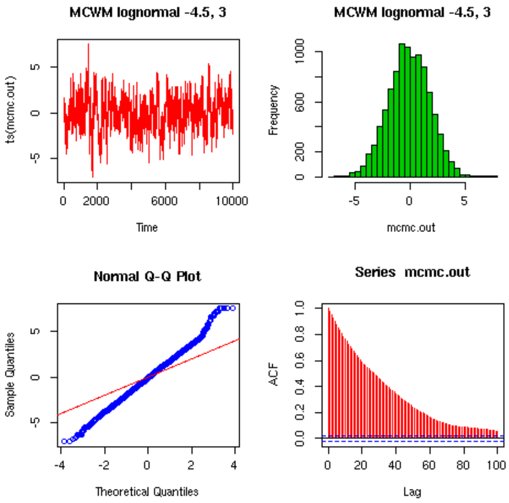


Figure 3.5.4: Bias of MCWM variant of SEMH

weight distributions, all using a N(0,1) target with a U(-.5,.5) proposal, for comparison results are also shown for GIMH. The results for weight distributions which are independent of x are shown in the left half of figure 3.5.5 where log-normal weights are used, with $\sigma = 1, 2, 3, 4$ and variance 1.72, 53.6, 8.10×10^3, 8.89×10^6 as the variance and skewness increases so does the bias. More interesting is the right hand set where the variance decreases with x (labelled z in the legend) and so with the target density $\pi(x)$. Each of these three examples results in the mass of the invariant density being pushed away from the mode of the target distribution, resulting in a bi-modal density when $f_W(w|x)$ is log-normal with $\sigma = 2/(1+x^2)$.

Acceptance rates for stochastic approximate Metropolis-Hastings

When the weight distribution is independent of x for a $\delta_x(.)$ proposal the equivalent to equation 3.5.3 above for the average acceptance rate is

$$\mathfrak{a}_\delta = \int_{\mathcal{W}} \int_{\mathcal{W}} \min(\frac{y}{w}, 1) f_W(y) f_W(w) dw dy$$

this integral is symmetric about the line $y = w$ and so can be written as

$$\mathfrak{a}_\delta = 2 \int_0^\infty \int_0^w \frac{y}{w} f_W(y) f_W(w) dy dw$$

the inner integral can be evaluated for the three examples we get the following:

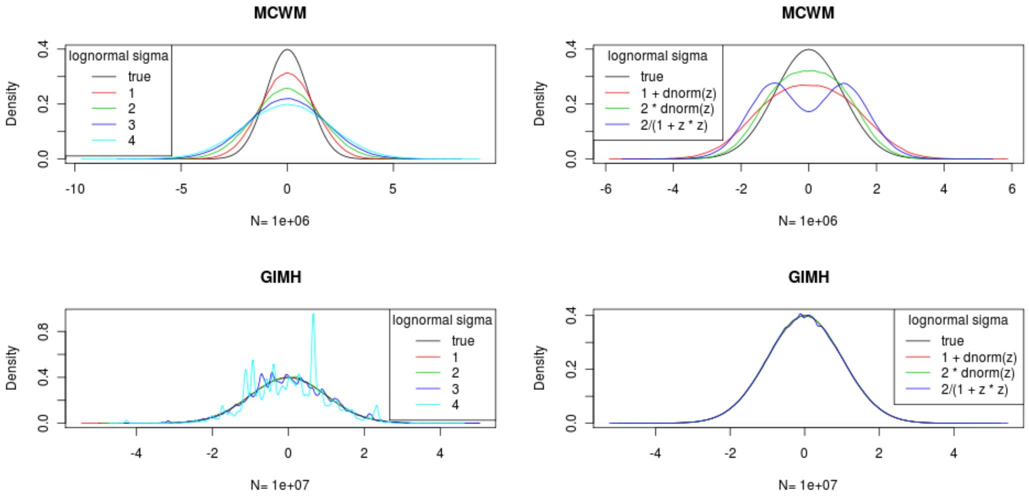


Figure 3.5.5: Simulated MCWM and GIMH, left half weights independent of x, right half variance increases as density decreases.

log-normal

$$\mathfrak{a}_\delta = 2\int_0^\infty w^{-1}\Phi(-\sigma + (\log(w)-\mu)/\sigma)\phi((\log(w)-\mu)/\sigma)dw$$

which can be numerically integrated.

d-dimensional exponential ratio

$$a_\delta = 2\int_0^{2^d} \frac{1}{w}(1 - F_\Gamma(2(d\log(2) - \log(w)))f_W(w)dw$$

3.5.5 Conclusions of Analysis of SAMH and SEMH

The novel analysis presented here of the SEMH shows that the analysis of GIMH and MCWM are simplified by considering them in a more general framework. The typical examples of weight distributions studied here show behaviours in the resulting SEMH algorithms that are expected to transfer qualitatively to more complex situations such as the GIMH and MCWM. The least unsurprising observation is that the spikiness and sticking of the chain gets worse as the variance of the weights increases, a tight lower bound is derived. The bias of the SAMH increases as the variance of the weights increases. When the distribution of the weights depends on x when the variance is higher in the tails the SEMH algorithm struggle to reach the tails, the MCWM variant shows more pronounced light tails. When the variance is

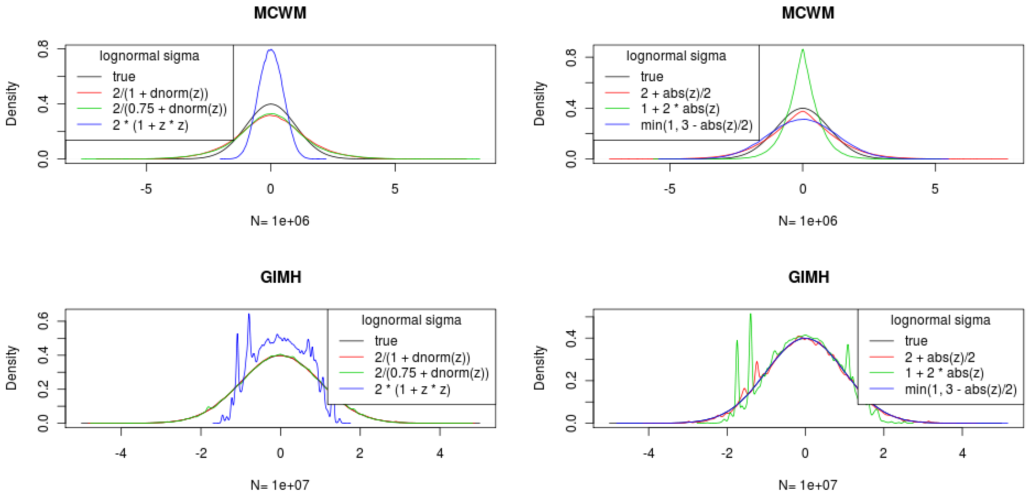


Figure 3.5.6: Simulated MCWM and GIMH, left half variance decreases with density, right half variance increases with density.

higher, near the mode the MCWM variant can display pronounced bias, such as a bimodal result for a true unimodal target.

3.6 Kernel Density Metropolis-Hastings Algorithm

When θ has a small dimension and the marginal posterior $\pi(\theta|\mathbf{y})$ is believed to be smooth a new algorithm is proposed which overcomes the sticking of the GIMH and the bias of MCWM by utilizing the assumed smoothness. We note in passing that when the distribution of $\mathbf{Y}$ is considered then the exact marginal $\pi(\theta|\mathbf{Y})$ is a random function.

We have again assumed that we have available unbiased but noisy point estimates $\tilde{\pi}(\theta)$ of $\pi(\theta)$, where $\tilde{\pi}$ is often a posterior distribution of a high dimensional model and want to use MCMC to investigate it, the estimates could be obtained from an importance sampler as used in the GIMH described above or via other methods such as a particle filter.

The proposed algorithm is based on a sequence $\widetilde{\pi(.)}_j$ of kernel density estimates of $\pi(\theta)$, these are estimates of the whole density in contrast to the underlying point estimates. A standard Metropolis-Hastings is run over θ using $\widetilde{\pi(.)}_j$ to calculate the acceptance ratio, the algorithm is

1. initialise θ

2. initialise $\widetilde{\pi(\theta)}_0$

3. for $j = 1 \dots N_{mcmc}$

   (a) propose θ^* from $q(\theta^*|\theta)$

   (b) obtain $\tilde{\pi}(\theta^*)$

   (c) calculate the new density estimate $d_j = \widetilde{\pi(\theta)}$

   (d) accept θ^* with probability $\min(\mathcal{A}, 1)$ where

$$\mathcal{A} = \frac{\widetilde{\pi(\theta^*)}_j q(\theta|\theta^*)}{\widetilde{\pi(\theta)}_j q(\theta^*|\theta)}$$

Conceptually at (c) we compute it for all θ, in practice we only need it at the points θ and θ^*. Note that we make use of all the estimates $\tilde{\pi}(\theta^*)$ in computing $\widetilde{\pi(\theta)}$. We use a standard kernel estimate

$$\widetilde{\pi(\theta)}_n = \sum_{i=1}^{n} \tilde{\pi}(\theta_i) K(\frac{\theta - \theta_i}{h_n}) / \sum_{i=1}^{n} K(\frac{\theta - \theta_i}{h_n}) \tag{3.6.1}$$

where h_n is a predefined non-increasing sequence of bandwidths and $K(.)$ is a symmetric kernel. Computationally the use of a kernel with bounded support gives several options for efficiently computing the KDE sequentially. When the target is a Bayesian posterior the initial estimate $\widetilde{\pi(\theta)}_0$ can be taken from the prior on θ.

This algorithm generates a sequence of values of θ which because of the dependence on past values is no longer Markov. It is hoped that the sequence will converge to $\pi(\theta)$ subject to some conditions on the target and proposal distributions and the sequence h_n. The initial experiments described below have used a constant value, in this case the best that can be hoped for is that $\widetilde{\pi(.)}_j$ converges to the convolution of the target $\pi(\theta)$ and the kernel with bandwidth h.

3.6.1 Kernel MH Algorithm - Naive implementation

A naive implementation which is computationally inefficient has been used to investigate the behaviour on the example used in section 3.5.2 and a more challenging 2-d example. At each iteration n the KDE is recomputed for the two values θ, θ^* from the stored values θ_j $\tilde{\pi}(\theta_j)$ $j = 1 \dots n$ which is $O(n^2)$. A Gaussian kernel with a range of constant bandwidths (bw=1,.1,.01,.001) gives the results shown below (which can be compared with the GIMH results in figure 3.5.2), bw=1 is over smoothed giving biased results, on this short run on a simple toy bw=.1 may be the best. The KDE was initialised from 100 observations from “a prior” of N(2,2), this initialisation is still visible in these short runs for all bandwidths < 1.

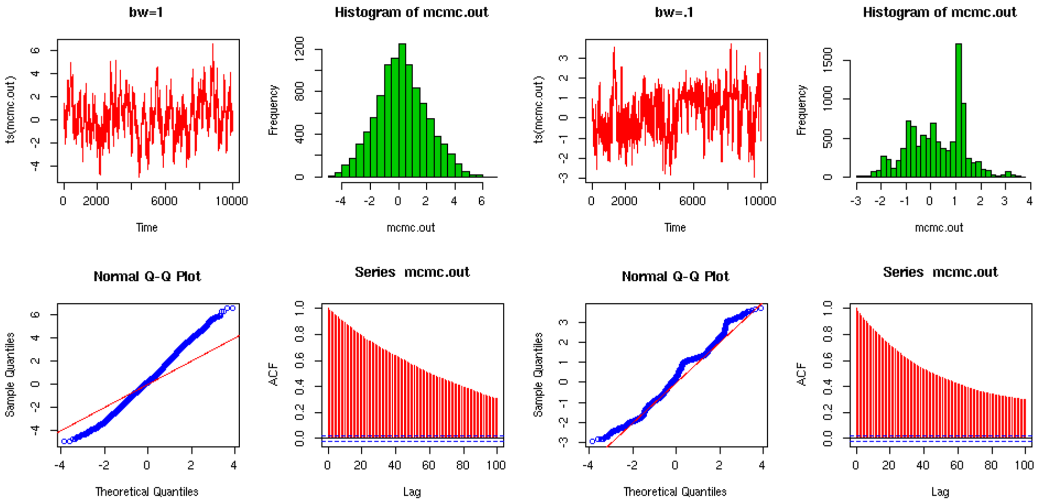


Figure 3.6.1: Kernel MH example, bw=1, .1

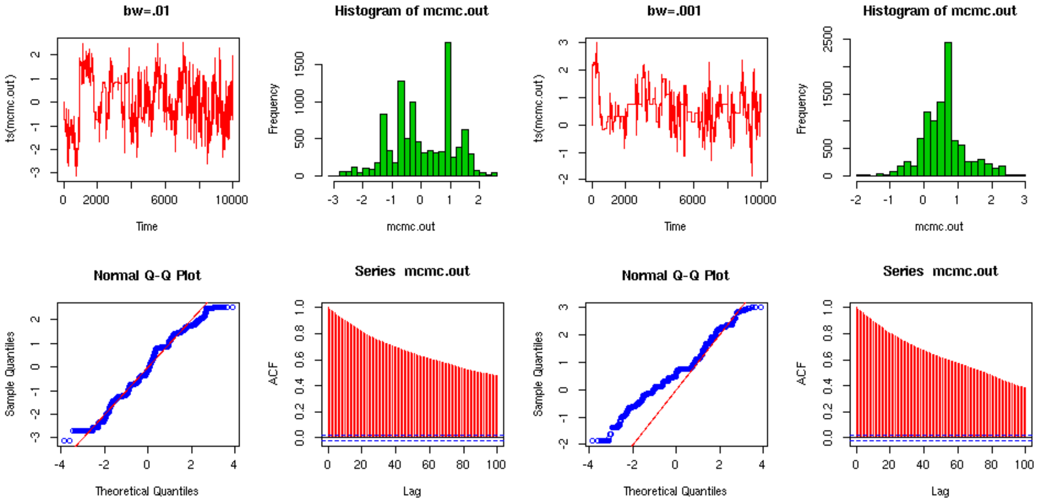


Figure 3.6.2: Kernel MH example, bw=.01, .001

Himmelblau Example Distribution

The Kernel Metropolis-Hastings (KMH) algorithm has been investigated in higher dimensions, 2-d and 5-d and appears to work well, further programming to improve efficiency is needed before any more extensive runs. A 5-d $N(\mu, I_5)$ target is used, the run time is still $O(n^2)$ the KMH appears to work well where the SEMH would get stuck, running independent parallel chains circumvents the worst effects of the $O(n^2)$.

In 2-d a challenging multimodal example with heavy tails based on the Himmelblau function[5] $H(x,y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$ has been used. The target density is $\pi(x,y) \propto 1/(1 + H(x,y))$ and contours of its logarithm are shown in figure alongside a perspective view.

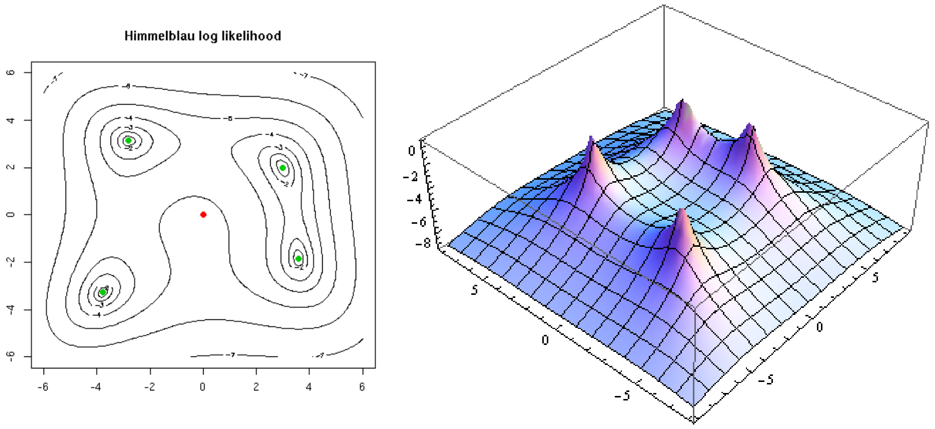


Figure 3.6.3: Himmelblau example target distribution, the log-likelihood is shown as contours and a perspective view.The green dots indicate the position of local maxima, the red dot a local minima.

The marginal distributions are intractable and so an "exact" MH run of length 10^8 was used to obtain them along with the table below. Although the positions of the 4 modes are known exactly, the position of them on the two marginals is not, they are close to the projections of the peaks. Comparisons have been made using these as exact probabilities (shown below as %).

	(−Inf,−10]	(−10,−5]	(−5,0]	(0,5]	(5,10]	(10, Inf]
(−Inf,−10]	0.12	0.08	0.08	0.07	0.06	0.10
(−10,−5]	0.07	0.39	1.05	0.69	0.26	0.06
(−5,0]	0.07	0.74	16.83	21.26	0.76	0.07
(0,5]	0.07	0.52	24.28	29.01	0.56	0.07
(5,10]	0.06	0.25	0.98	0.72	0.19	0.06
(10, Inf]	0.09	0.07	0.07	0.07	0.06	0.09

Results in 2-d

KMH runs have been compared with the SEMH and SAMH algorithms for a range of parameters. The "noise" is log-normal with parameter σ one of 2,3,4,5. For $\sigma = 2$

[5] Himmelblau's function

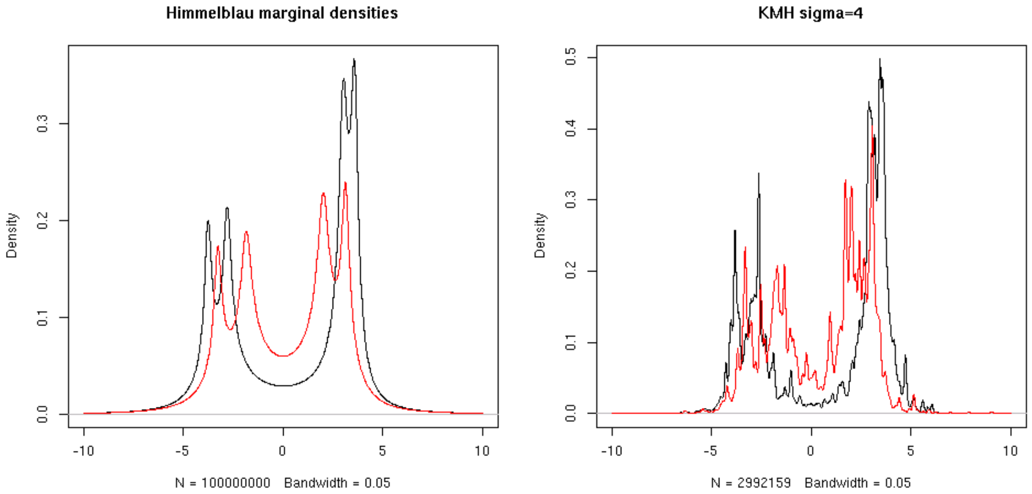


Figure 3.6.4: Himmelblau example, 1-d marginal densities of the 2-d target. The left hand plot shows the true densities obtained by an MCMC run of length 10^8. The right hand plot shows an estimate obtained from a KMH run of $\approx 3 \times 10^6$.

the SEMH algorithm worked well, a longer run would be necessary in practice, the SAMH was heavily biased not identifying the modes and putting too much weight in the tails, the bias of SAMH is expected to increase with σ so was not considered for higher values of σ. For $\sigma = 3$ the SEMH algorithm showed significant sticking but was still acceptable, (see left hand plot in 3.6.5). The KMH for $\sigma = 3$ at 3 bandwidths produced slightly better results, as measured by χ^2 identification of the 4 modes and tails, for a lower number of samples, but with the current implementation required more cpu time. The SEMH with $\sigma = 4$ got badly stuck, longer runs are unlikely to improve this. The KMH for $\sigma = 4$ at 3 bandwidths produced significantly better results, (see right hand plot in 3.6.5), although longer runs are necessary the 4 modes are correctly identified. Even with $\sigma = 5$ useful results are obtained, tuning of its parameters and/or a longer run is necessary to fully sample all 4 modes.

Programming details

Different approaches have been tried in the 2-d and 5-d examples they must be integrated. Currently logarithms of densities are calculated as the smoothing is linear, computation is dominated by the exp(.) function, this should be changed. In 2-d an index of which observations are in each of a grid of squares side h, is maintained so that the kernel is not evaluated when known to be zero. In 5-d a

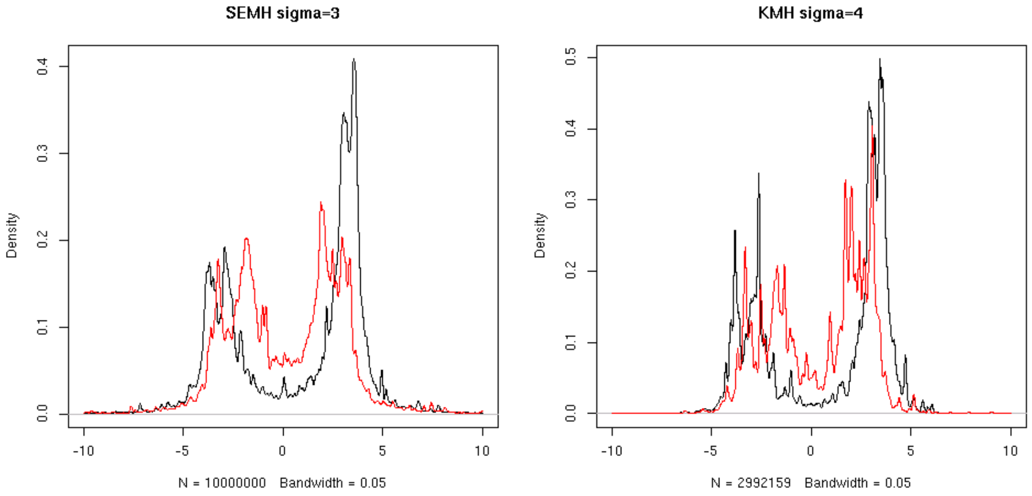


Figure 3.6.5: KMH SEMH comparisons

decreasing kernel bandwidth h is used.

KMH on Indian Buffet Posteriors

Attempts have been made to use the KMH on Indian Buffet Posteriors, these are not described in detail, or described in the next chapter as they have been unsuccesful. The reason appears to be the far greater variance of the log likelihood estimates, which can be equivalent to a log normal parameter σ of 100 or more. The result is that the samples become concentrated in an area centered on one large value but matching the proposal distribution.

3.7 Concluding Remarks on MCMC

Although the impact of a poor proposal distribution on importance sampling is well known, the study in section 3.2 illustrates the problem and highlights the difficulties that can arise even in well understood situations where the variance is known to be finite. The deleterious effect of increasing dimension has been demonstrated for dimensions as low as 50. The increasing error in the sample standard deviations shown in figure 3.2.2 shows that examination of the variance of weights is not sufficient to guarantee acceptable behaviour. Better methods for robustly diagnosing poor performance from analysis of samples would be very useful.

The SEMH algorithm can provide good estimates of the target distribution for moderate values of the variance of the estimator $\hat{\pi}(x)$ of the target density. The bias of the SAMH has been investigated and it has been shown that the bias can be significant, it provides a useful algorithm for exploratory analysis in low variance situations but requires great care in other uses. In GIMH the variance is often unknown and often infinite, effort should be concentrated on improved proposal distributions to reduce this variance. When a better proposal can not be found and the GIMH is sticking badly then the KMH provides a useful approximate algorithm whose bias appears to be much smaller than that of MCWM. However until some theoretical underpinning is available it does not merit the programming effort needed to implement it efficiently.

Chapter 4

Epidemic Inference

4.1 Introduction

Inference for the parameters of epidemic models has been studied for many years but still provides challenges and lags the development of models. Classic approaches are described in the monograph by (Becker, 1989), and more recent approaches in the book by (Andersson and Britton, 2000).

This chapter presents results for inference in the homogeneous GSE using the exact marginal likelihood for both continuously and regularly observed epidemics. Attention then switches to the much more challenging problem of inference in heterogeneous models, focusing on the Indian buffet epidemic.

Data on epidemics are almost always incomplete in several ways and in practical situations the choice of model and availability of data cannot be cleanly separated. The times of infection are invariably unavailable, except in rare laboratory experiments (such as Charleston et al., 2011). An important distinction for inference is whether the epidemic is still in progress and prediction is the primary aim or complete and a retrospective study is being performed, the relevance of this distinction is shown below by the bi-modality of the Bayesian posterior distribution of the infection rate.

An additional way in which data are incomplete is the resolution of recording times of events, usually data on epidemics are only available on a daily basis, or sometimes less frequently. In a slowly progressing disease such as AIDS this will not be an issue but in any disease where the infectious period is only a few days the distinction is important. When data are only available at regular times e.g. daily, two approaches to model choice and subsequent inference are possible, either a discrete time model as described in section 2.4 or a continuous time model with discrete observations. The choice between these classes of model will often be guided by epidemiological considerations, for instance it could be argued that a discrete time model is more appropriate as there are clearly differences in infection processes during a day. For example it is often the case that adults and children mainly mix with the same age group from 9am-5pm, within the family 7am-9am, and 5pm-12pm and only with a partner at night. These differences in contact rates might be influential if an age stratified model is in use. Exact inference for the homogeneous GSE is demonstrated below for both approaches.

A range of observed or unobserved covariates for individuals, in particular age or location, can have significant effects on some or all of contact patterns, infectiousness, susceptibility and durations of phases in an epidemic. When covariate data are available the number of parameters increases and care must be taken to ensure that the extra parameters in the model are identifiable. Good inference with co-variates builds on a good understanding of the underlying model and so co-variates are not considered here.

4.2 Inference for the General Stochastic Epidemic (GSE)

Given complete data of all infection and removal times T_j^I and T_j^R the likelihood of the GSE is readily obtained from the definition in equation 2.2.1 and the standard non-stationary Poisson process likelihood, from which classical or Bayesian inference may proceed. Before giving the equation two points on the data should be noted which are important when the complete data is inferred as part of an MCMC approach. The complete data include the time of infection, from outside the population and the scope of the stochastic model, of the initial infective T_1^I. In some circumstances such as when considering simulated data or experimental data this initial infection time may be known, t_1^I, and could be taken as the time origin, however usually only times of removal are available and the natural time origin is the first removal, in which case $T_1^I < 0$ is a random time relative to 0. The second point is that data are often available as counts of infectives and removals and the individuals are unlabeled, that is the times T_j^I and T_j^R cannot be paired and each set of times is ordered. The full data likelihood for the GSE with parameters (λ, ρ) in a population size n_p observed over the period from t_1^I to T_{obs} with m infections at $t_j^I \, j = 1 \ldots m$, including the initial one, and n removals at $t_i^R \, i = 1 \ldots n$ is

$$L=\prod_{i=1}^{n}\rho I\left(t_i^R-\right)\prod_{j=2}^{m}\lambda S\left(t_j^I-\right)I\left(t_j^I-\right)\exp\left\{-\int_{t_1^I}^{T_{\text{obs}}}\left(\lambda S\left(t\right)I\left(t\right)+\rho I\left(t\right)\right)dt\right\} \tag{4.2.1}$$

where $n\leq m\leq n_p$ and t_j^I and t_i^R are ordered. $S\left(t\right)$ and $I\left(t\right)$ are the numbers of susceptibles and infectives at t and limits from the left are indicated by $S\left(t-\right)$. The counts are obtained from the times as $S\left(t\right)=n_p-\sum_{j=1}^{m}\mathbf{1}\left[t\geq t_j^I\right]$ and $I\left(t\right)=n_p-S\left(t\right)-\sum_{i=1}^{n}\mathbf{1}\left[t\geq t_i^R\right]$.

To ensure that $S\left(t\right)I\left(t\right)>0$ the event times must satisfy $t_{j+1}^I<t_j^R$ for $j<n$. When data are available as paired times, which is necessary when considering non-exponential distributions of removal times, the likelihood must be modified. See Jewell and Roberts (2012) for details of the general case, if the epidemic is known to be finished $I\left(T_{\text{obs}}\right)=0$ and $m=n$ a simpler form can be used as given by Neal and Roberts (2005),

$$L\propto\rho^n\prod_{j=2}^{m}\lambda S\left(t_j^I-\right)I\left(t_j^I-\right)\exp\left\{-\int_{t_1^I}^{T_{\text{obs}}}\left(\lambda S\left(t\right)I\left(t\right)\right)dt-\sum_{j=1}^{m}\rho\left(t_j^R-t_j^I\right)\right\} \tag{4.2.2}$$

where $t_j^I<t_j^R$ for $1\leq j\leq n$ and $I\left(t\right)>0$ for $t<\max_j(t_j^I)$.

As the GSE is a superposition of two non-stationary Poisson processes, which are independent conditioned on the state, the likelihood separates and so simple MLE are available.

$$\hat{\lambda}=\frac{m-1}{\int_{t_1^I}^{T_{\text{obs}}}S\left(t\right)I\left(t\right)dt} \tag{4.2.3}$$

$$\hat{\rho}=\frac{n}{\int_{t_1^I}^{T_{\text{obs}}}I\left(t\right)dt} \tag{4.2.4}$$

The Bayesian conjugate prior for a Poisson distribution is a gamma distribution and so if the prior for (λ,ρ) is taken to be independent gamma distributions the Bayesian posterior is also immediately available.

However data at this level of detail are very rarely available and so neither is this straightforward approach to inference. Data are usually only available for the removal times and so inference for this situtation has been extensively studied, the two most important approaches are MCMC and martingales, both of which are described below.

4.2.1 Martingale Estimators for the GSE

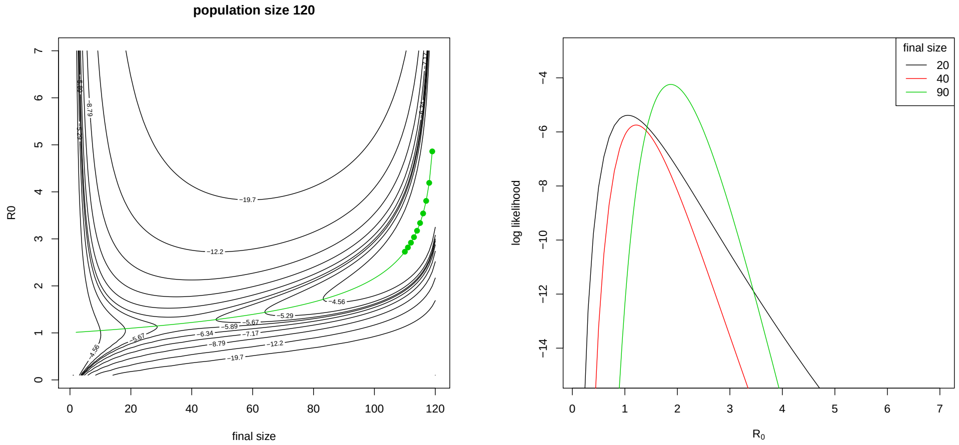
Martingales can provide a powerful approach to point estimates with missing data and asymptotic results can be used to provide confidence intervals. If $M(t;\theta)$ is a martingale observed on $[0,T]$ setting $M(T;\theta)=0$ can provide an equation in θ which can be used as an estimator. The monograph by Becker (1989) shows in chapter 7 that the martingale $M(t) = C\left(t\right) - \int_0^t \lambda S\left(\tau\right) I\left(\tau\right) d\tau$ can be used to derive the same estimate for the infection rate as the MLE given in equation 4.2.3, and gives its standard error using the martingale variation process. Perhaps more usefully he shows that a martingale estimate for $\mathcal{R}_0$ based only on the final size ν can be derived as $\hat{\mathcal{R}_0} = \left(n_p/\nu\right) \sum_{i=1}^{\nu} \left(n_p - i\right)^{-1}$ and again gives the standard error. In a paper Becker and Hasofer (1997) show how the martingale $M(t) = S(t)(1+\mathcal{R}_0/n_p)^{R(t)}$, which was described in section 2.3.5, can be used to derive an estimator for the removal rate ρ only based on the removal times, however it requires knowledge of either T_1^I the time of the first infection or the number of infectives at the first removal, which are generally not known.

4.2.2 Inference for $\mathcal{R}_0$ the Basic Reproduction Number in the GSE

When analysis is being conducted on a completed epidemic, the martingale estimator above shows that direct inference for $\mathcal{R}_0$ is possible using only the final size. The asymptotic values for the standard error can be inaccurate for small n_p or small $\mathcal{R}_0$. The embedded Markov chain (EMC) described in chapter 2 can be used to provide exact calculations for the probabilities of the final size for a given parameter value and hence used in its estimation. It is straightforward to calculate numerical values of the probabilities of each final size $0, 1, \dots n_p$ at a suitable chosen set of $\mathcal{R}_0$ values, so giving the likelihood. A contour plot of the log likelihood for $n_p = 120$ is shown in figure 4.2.1. Numerical integration, using linear interpolation between the calculated values of the log likelihood, can be used to obtain the marginal likelihood for $\mathcal{R}_0$ given the final size as shown in the right hand plot.

Impact of the choice of prior on the posterior of $\mathcal{R}_0$

Multiplication of the likelihood on the grid by a prior provides a posterior density which can be used to give a Bayesian estimate for $\mathcal{R}_0$, some care is needed in the choice of prior. In particular when the final size $= n_p$ a prior that is at least moderately informative is needed for the posterior $\mathcal{R}_0$ to exist. An example is shown in figure 4.2.2 for a population of size 50 with a prior $\sim \Gamma(3,1)$ (mean and variance 3), for less informative priors the posterior density for final size 50 can not



(a) Contours of log likelihood

(b) Examples of log likelihood

Figure 4.2.1: GSE log likelihood for $\mathcal{R}_0$ given final size. Sub-figure b shows vertical cross sections of sub-figure a.

be integrated.

As the likelihood separates into terms dependent on the two rates (λ, ρ), the natural and often taken approach is two use independent conjugate priors, which are gamma distributions. However care must be taken, as it was shown in chapter 3 that a ratio of gammas can imply a very heavy tailed prior on $\mathcal{R}_0$, which can imply a heavy tailed posterior. The implied prior on $\mathcal{R}_0$ must be considered when performing MCMC for (λ, ρ) as if the posterior for $\mathcal{R}_0$ is heavy tailed it will adversely affect the mixing of MCMC. The same effect has also been identified in Clancy and O'Neill (2008) where they show that the heavy tailed posterior can result in large values for $\mathbb{E}(\mathcal{R}_0)$ and also say "Such findings illustrate the need for caution when using $\mathcal{R}_0$ alone as a summary measure of an epidemic".

4.2.3 MCMC Inference for the GSE

Two papers Gibson and Renshaw (1998) and O'Neill and Roberts (1999) introduced the use of MCMC for epidemics where only the removal times are observed, the basic algorithms have since been extended to more complex algorithms and improvements such as the use of non-centering (Neal and Roberts, 2005) have been introduced. The basic approach is to provide a prior on (λ, ρ, T_1^I) and given the removal times $t_j^R\, j = 1 \dots n$ use MCMC to generate samples from $(\lambda, \rho, T_j^I\, j = 1 \dots m)$ from which the marginal distribution of (λ, ρ) is obtained. Two approaches are possible, either paired or ordered infection and removal times. The use of paired infection and

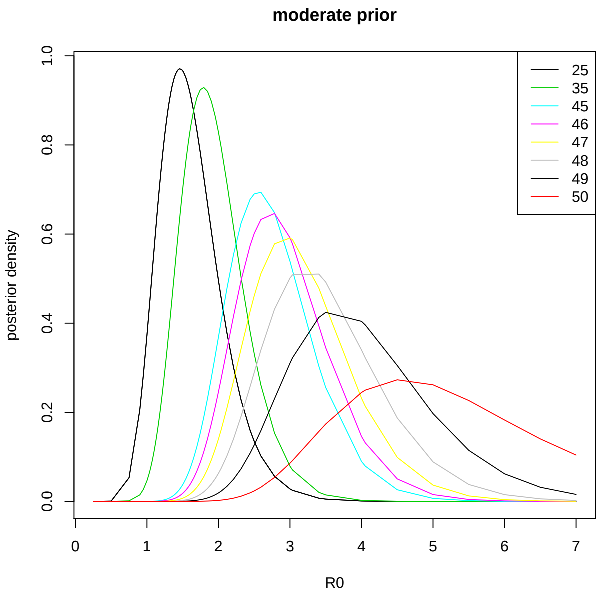


Figure 4.2.2: Posterior densities for $\mathcal{R}_0$, in the GSE in a population of 50

removal times has the advantage of handling non-exponential distributions. The alternative which facilitates handling epidemics which are still in progress is to use the ordered infection times as originally proposed by Gibson and Renshaw and demonstrated by O'Neill and Roberts (1999), these approaches have been combined and extended by Jewell and Roberts (2012). When the epidemic is known to be complete, $n = m$ and the posterior is on a subset of $\mathbb{R}^{n+2}$, the conditions on the support of the likelihood, given in equation 4.2.1, create a posterior with a complicated support and other discontinuities. These discontinuities also explain why other algorithms that handle missing data such as the EM (Dempster et al., 1977) cannot be used sucessfully. When the epidemic is still in progress $n \leq m$ and reverse jump MCMC is needed to sample from the posterior.

Although a variety of MCMC algorithms are often used successfuly on epidemic data and have been applied to large problems, for example Jewell et al. (2008), the mixing of the MCMC can be poor and some data appear to give posteriors that are more difficult to sample.

An approach to understanding these problems using the exact marginal distribution is considered in the next section, which also provides an alternative algorithm.

4.3 Inference for the GSE using the Exact Marginal Distribution

The marginal distribution of the removal times from the GSE is not readily available in an analytic form, however the Markov representation of the GSE $X_t = (S(t), I(t))^T$ and the availability of software to calculate the matrix exponential allows numerical calculation of the marginal likelihood. The use of the matrix exponential to calculate probabilities in SIR epidemics, has been known as a theoretical result for many years, but computational resources limited its use. The calculation of the exact marginal likelihood using it is believed to be original. The use in inference and the identification of the bimodal nature of the likelihood is original.

The calculation of the likelihood for the GSE Markov process would be very simple if the state X_t was observed at the removal times, it is not, all that is known is that $S(t)+I(t) = n_p-j$ for $t \in [t_j^R, t_{j+1}^R)$. However by constructing a set of modified chains the joint distribution of state and time between removals is obtained. In order to calculate the distribution between t_{j-1}^R and t_j^R a modified transition matrix is constructed in which all the states with j removals, where $S(t)+I(t) = n_p - j$ are absorbing. First some notation, extending that in section 2.3.2 is introduced: the full state space $\mathcal{X} \subset \mathbb{Z}^2$, of size $n_s = (n_p+1)(n_p+2)/2$ is partitioned by the number of removals into $\mathcal{X} = \bigcup_{j=0}^{n_p} \mathcal{X}_j$, where $\mathcal{X}_j = \left\{(s,i) \in \mathbb{Z}^2 : \, s+i = n_p - j, \, s \geq 0, \, i \geq 0\right\}$. The possible transitions are either infections which remain in the same subset $\mathcal{X}_j$ or removals which move from $\mathcal{X}_j$ to $\mathcal{X}_{j+1}$. The transitions of X_t between t_j^R and t_{j+1}^R are governed by the subset of the transition rate matrix for the entire process $\mathbf{Q}_\theta$ on the states $\mathcal{X}_j \cup \mathcal{X}_{j+1}$. The modified transition matrix $\mathbf{Q}_j$ where $\mathcal{X}_{j+1}$ is absorbing is

$$\mathbf{Q}_j = \begin{pmatrix} \mathbf{Q}_{\mathcal{X}_j,\mathcal{X}_j} & \mathbf{Q}_{\mathcal{X}_j,\mathcal{X}_{j+1}} \\ 0 & 0 \end{pmatrix} \tag{4.3.1}$$

this is a valid transition rate matrix as the only transitions out of $\mathcal{X}_j$ are to $\mathcal{X}_{j+1}$.

The joint density of time between removals j and $j+1$ and probability of the state at removal $j+1$, conditioned on the state at removal j, is denoted

$$f_j(t, x_{j+1}|x_j) = \lim_{dt \to 0} \mathbb{P}\left(X_{s+t} = x_{j+1}, \, X_{s+t-dt} \in \mathcal{X}_j | X_s = x_j\right)/dt$$

on the Markov process governed by $\mathbf{Q}_j$. This can be calculated using lemma 2 as

$$f_j(t, x_{j+1}|x_j) = \left[\mathbf{Q}_j \exp(t\mathbf{Q}_j)\right]_{x_j x_{j+1}} \tag{4.3.2}$$

and a recursive calculation gives the marginal likelihood. The distribution of the number of infectives immediately after the first removal was given in proposition 1, and so for subsequent removals the joint density of the removal times $1 \dots j$ and the probability of the state at removal j, denoted $g_j\left(t_{1:j}, x_j\right)$ is calculated as

$$g_{j+1}\left(t_{1:j+1}, x_{j+1}\right) = \sum_{i \in \mathcal{X}_j} g_j\left(t_{1:j}, i\right) f_j\left(t_{j+1} - t_j, x_{j+1} | i\right) \tag{4.3.3}$$

where $g_1(t, l) = p_l$ is independent of t and p_l is from proposition 1, dropping the zero term as we know there is at least one infective

$$p_l = \frac{1}{1 + (n_p - l - 1)\mathcal{R}_0 / n_p} \prod_{i=1}^{l-1} (1 - p_i)$$

(the product for $l = 1$ is taken as 1).

After the last observed removal at t_n two situations are considered, the epidemic is known to be complete or it is observed until time $T_{\text{obs}} > t_n$ without any further removals. In the first case the final state is known and is $x_c = (n_p - n, 0)$, so the density of the last transition is $f_{n-1}\left(t, x_c | x_{n-1}\right)$.

Lemma 5. *The marginal likelihood of the n removal times $t_1 \dots t_n$ of a completed GSE is*

$$L = \sum_{i \in \mathcal{X}_{n-1}} g_{n-1}\left(t_{1:n-1}, i\right) f_{n-1}\left(t_n - t_{n-1}, x_c | i\right) \tag{4.3.4}$$

Proof. by construction using the recursion in equation 4.3.3 □

Lemma 6. *The marginal likelihood of the n removal times $t_1 \dots t_n$ of GSE still in progress, observed until t_n is*

$$L = \sum_{i \in \mathcal{X}_n} g_n\left(t_{1:n}, i\right) \tag{4.3.5}$$

Proof. directly from equation 4.3.3 □

The likelihood can now be used for inference, either by numerical maximization for a point estimate or in a straightforward MCMC for the parameters (λ, ρ). A motivation for obtaining this exact marginal likelihood was to understand the difficulties sometimes encountered in MCMC for epidemic models and so contour plots of the log likelihood have been calculated, an example is shown in figure 4.3.1 superimposed on a shaded scatter plot of the results of an MCMC run using the algorithm

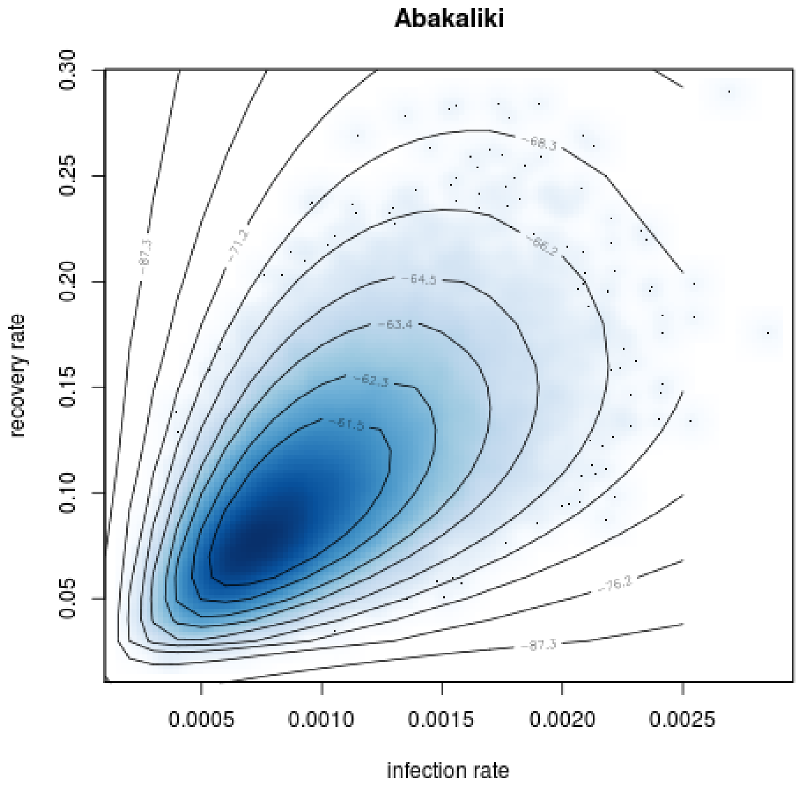


Figure 4.3.1: Abakiliki posteriors

and data from O'Neill and Roberts (1999) for 30 smallpox cases in a population of 120 from Abakaliki, which confirms that both approaches produce the same results. These data were also used as an example by Fearnhead and Meligkotsidou (2004) who use an alternative way of calculating the exact likelihood and then multiply the likelihood by a prior for θ and normalize by numerical integration.

4.3.1 Calculating the Matrix Exponential

The algorithm presented above depends fundamentally on the ability to calculate the matrix exponential repeatedly for different parameter values. This can be limited by the memory and cpu time required for the calculations. The memory problems are alleviated by the use of standard sparse matrix techniques, such as those implemented in the `R Matrix`[1] package. Several ideas for accurate approximation have been considered, in particular techniques based on the Fréchet derivative have been

[1] http://Matrix.R-forge.R-project.org/

tried and appear to have considerable potential. However further work is needed to obtain bounds on the size of the approximation errors and so they have not been used for the results presented in this thesis and are outlined in appendix A together with other techniques used in the computations.

Although the complete state matrix is large, the matrix exponential is only calculated on subsets defined in equation 4.3.1 which are $O(n_p)$. So larger populations can be handled by this algorithm than can be by algorithms which are reliant on the exponential of the full Markov rate matrix, such as those described in section 4.4.

4.3.2 Bi-modality of Posterior Distribution for the In Progress GSE

As an epidemic progresses and more data in the form of removal times becomes available the posterior distribution of the parameters evolves and becomes more informative. The exact calculation of the marginal likelihood that was derived above provides a mechanism by which the evolution of the posterior has been studied. In practical terms there is usually more interest in inference for an epidemic that is in progress than for one that is known to be complete. The results show that for simulated epidemics the likelihood for the “in progress” case is bi-modal, with one mode approaching the mode for the complete epidemic and another corresponding to a very high infection rate, with many cases still infected.

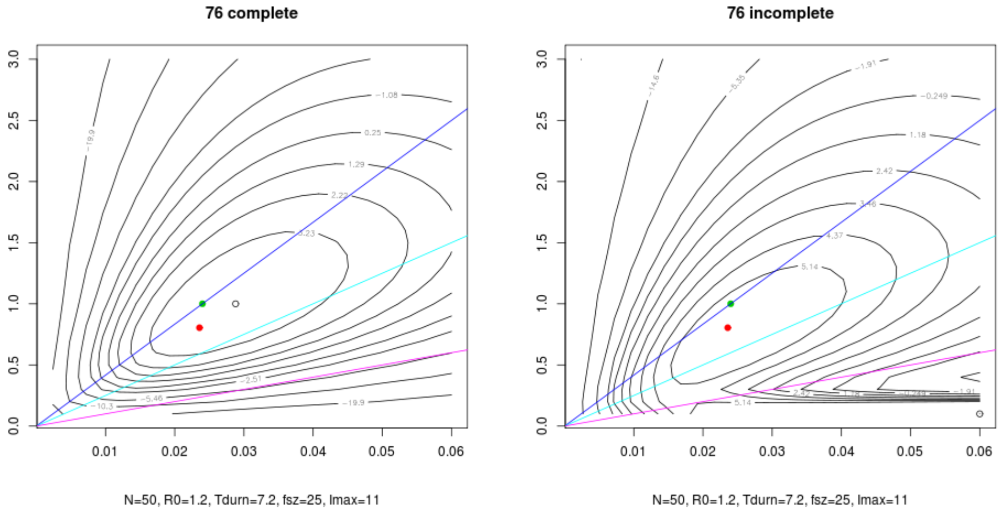


Figure 4.3.2: GSE marginal likelihood for simulated epidemic with $n_p = 50$, $\mathcal{R}_0 = 1.2$, left hand plot complete epidemic size=25, right hand plot epidemic in progress, x-axis is the infection rate and the y-axis is the recovery rate. The green dot indicates the true parameters, and the red dot the MLE for the completed epidemic, the straight lines indicate $\mathcal{R}_0 = 1.2$, 2, 5.

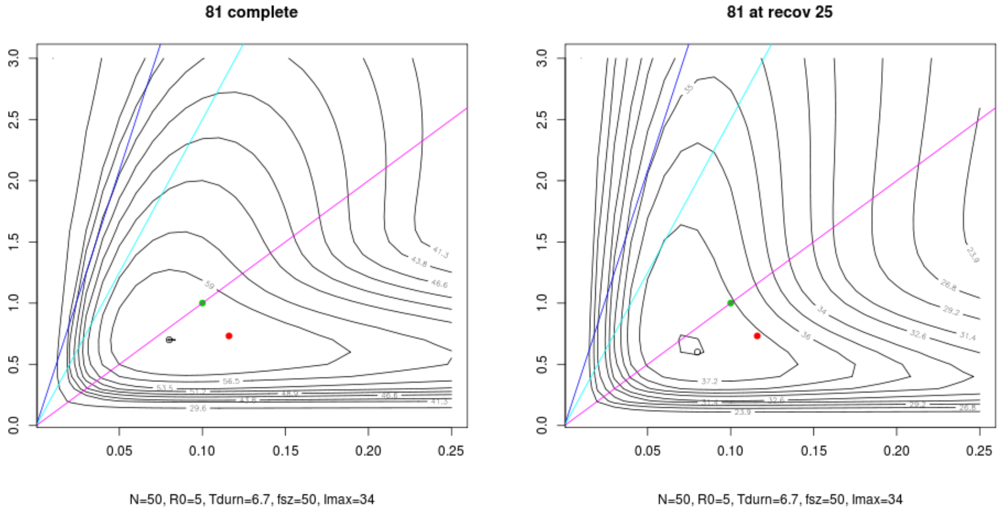


Figure 4.3.3: GSE marginal likelihood complete and in progress

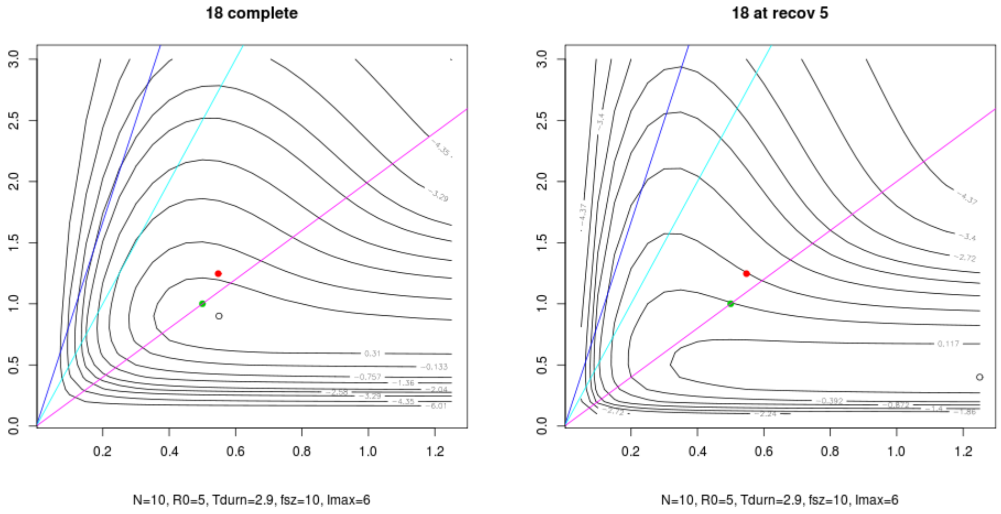


Figure 4.3.4: GSE marginal likelihood complete and in progress

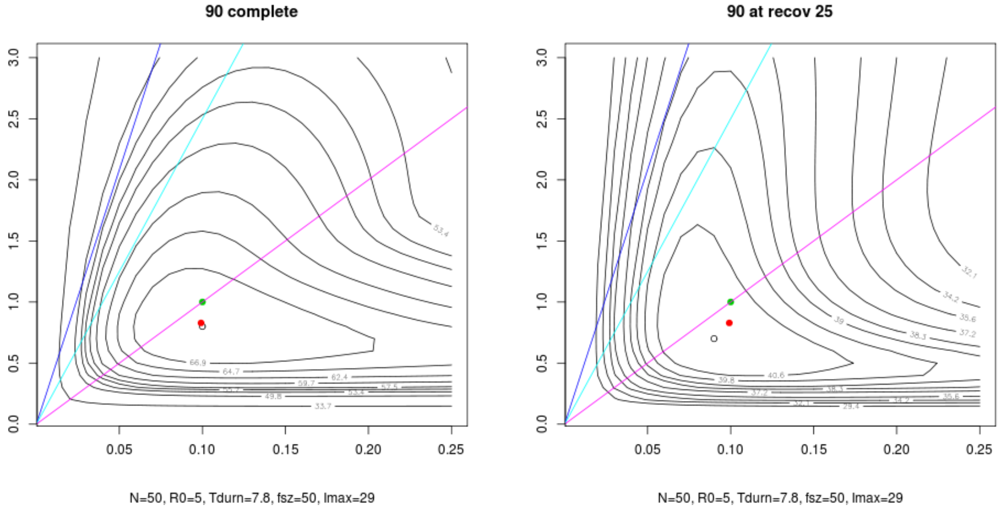


Figure 4.3.5: GSE marginal likelihood complete and in progress long tail example

Some examples are shown in figures 4.3.2-4.3.5, of contours of the exact marginal log likelihood for simulated data on a population of 50, in each figure the green dot indicates the true parameters, and the red dot the MLE for the completed epidemic, the straight lines indicate $\mathcal{R}_0 = 1.2,\ 2,\ 5$, the x-axis is the infection rate and the y-axis is the recovery rate. Figure 4.3.2 shows the difference between a complete and in progress epidemic and has “nice” elliptical contours matching the simulated $\mathcal{R}_0$ value of 1.2, an MCMC algorithm on these data might be expected to perform well. The example in figure 4.3.3 is more surprising, the final size is 50, all are infected and the marginal estimate of the recovery rate is reasonable, however the posterior for the infection rate covers a wide range of values. Figures 4.3.4 and 4.3.5 also correspond to the whole population being infected, with the first from a small population showing even less information in the posterior on the infection rate.

4.4 Inference for Regularly Observed Epidemics

Exact Bayesian inference for epidemics in which all the removal times are observed exactly has been considered in section 4.3, however usually data on epidemics are only available on a daily basis, or sometimes less frequently. Two approaches to model choice and subsequent inference are possible, either a discrete time model as described in section 2.4 or a continuous time model with discrete observations. The choice between these classes of model will often be guided by epidemiological considerations, which are not considered here. It is often easier to obtain analytic results from a continuous time model than from a realistic discrete time model however when numerical computations are used this advantage is reduced considerably.

The Markov representation of the GSE as described in section 2.3 combined with the matrix exponential and the binomial model developed in section 2.4 provide two very similar discrete time, discrete state space hidden Markov models (HMM). Inference for the parameters of an HMM has an extensive literature, in addition to the even larger literature studying estimates of the hidden states with known parameters.

This section considers a general finite state space HMM. We follow the common practice of denoting ranges of vectors or random variables as $X_{1:a} = X_1, X_2 \ldots, X_a$ and using $X_{-t} = (X_{1:t-1}, X_{t+1,T})$ where T is the size of X and known from the context. We frame the description in general terms and consider the case where there are n_s states, and the state at t is X_t we have T observations $Y_{1:T}$ and the transition matrix is $P = (p_{ij})\ i, , j \in \{1 \ldots n_s\}$ where $p_{ij} = \mathbb{P}\left(X_{t+1} = j | X_t = i\right)$ with an initial state distribution $\mathbb{P}\left(X_1 = i\right) = \nu_i$, the observation distribution is $\mathbb{P}\left(Y_t = j | X_t = i\right) = g_i\left(j\right)$ and is independent of the other X.

In many applications of HMM the transition matrix is known exactly, and interest is in efficient algorithms to calculate the exact marginal posterior distributions $P(X_t|Y_{1:n})$ which can be calculated using the forward-backward algorithm a standard technique see e.g. Rabiner (1989) also the MAP path or the Viterbi path can be calculated[2]. These calculations and the inference algorithms below rely on recursive computations of quantities, often known as alpha and beta. In the case of finite state and time, which we are considering, they are the probabilities defined as

$$\alpha_t(i) = \mathbb{P}\left(Y_{1:t},\ X_t = i\right) \text{ and } \beta_t(i) = \mathbb{P}\left(Y_{t+1:T} | X_t = i\right) \tag{4.4.1}$$

in more general cases the definitions are similar, Cappé et al. (2005) gives details, he also describes the scaling necessary to prevent underflow in their calculation. These are calculated as $\alpha_1(i) = \nu_i g_i(y_1)$ for $1 \leq i \leq n_s$ and a forward recursion

$$\alpha_{t+1}(j) = \sum_i \alpha_t(i) p_{ij} g_j(y_{t+1}) \text{ for } 1 \leq t < T$$

and $\beta_T(i) = 1$ for $1 \leq i \leq n_s$ and a backward recursion

$$\beta_t(i) = \sum_j \beta_{t+1}(j) p_{ij} g_j(y_{t+1}) \text{ for } 1 \leq t < T$$

the calculation of both recursions is $O(n_s^2 T)$.

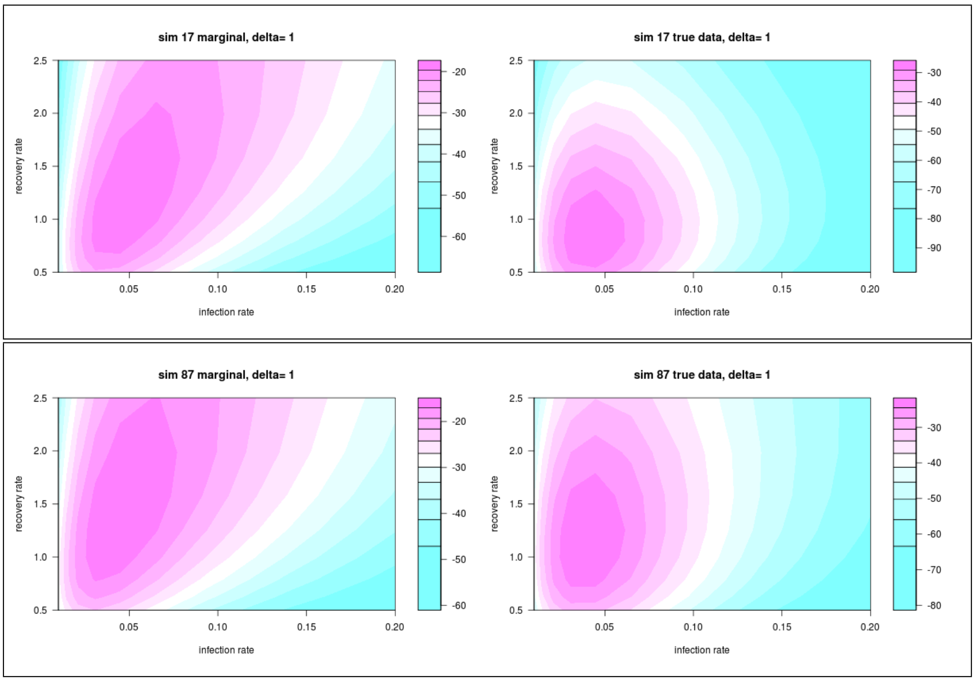


Figure 4.4.1: Exact marginal and full data log likelihood, for regularly observed GSE. 2 simulations numbered 17 and 87 both with $n_p = 50$. In the marginal likelihood the epidemic is assumed complete.

[2] The MAP is not necessarily a valid path the Viterbi is.

Many authors have considered inference in HMM, the book Cappé et al. (2005) provides a comprehensive treatment, also of note is Fearnhead (2011). The particular case of Markov jump processes, of which the SIR epidemic is an example, has been considered by Bladt and Sorensen (2005) who compare the EM and a standard MCMC algorithm. Golightly and Wilkinson (2011) describe a Pseudo Marginal Metopolis Hastings (PMMH) algorithm for a rabge of Markov jump processes. Often in the HMM literature the state space is assumed to be small and the parameter space is large and the noise in the observations is significant, whereas in the models we are considering the state space is large, the parameter space is small and the observations are partial without noise.

This section demonstrates that the exact calculation of the transition matrix using the matrix exponential can be used in standard HMM algorithms to provide exact inference for small population sizes.

We frame the algorithms in general terms and consider the case where there are n_s states, and the state at t is X_t we have T observations $Y_{1:T}$ and the transition matrix is $P = (p_{ij})\, i, j \in \{1 \dots n_s\}$ where $p_{ij} = \mathbb{P}\left(X_{t+1} = j | X_t = i\right)$ the dependence on θ is often dropped below to ease the notation.

Care is needed when using the term likelihood as it can mean $\mathbb{P}\left(X_{1:T}, Y_{1:T}|\theta\right)$ or $\mathbb{P}\left(Y_{1:T}|\theta\right)$, the full data likelihood is readily calculated as

$$\mathbb{P}\left(X_{1:T} = x_{1:T}|\theta\right)\mathbb{P}\left(Y_{1:T} = y_{1:T}|x_{1:T}, \theta\right) = \mathbb{P}\left(X_1 = x_1|\theta\right)\prod_{t=2}^{T} p_{x_{t-1}x_t}\prod_{t=1}^{T} g_{x_t}(y_t)$$

while the marginal or observed data likelihood is obtained from the forward recursions as

$$\mathbb{P}(y_{1:n}) = \mathbb{P}(y_1)\prod_{t=2}^{T}\mathbb{P}(y_t|y_{1:t-1}) \tag{4.4.2}$$

Consideration of the boundary conditions is important and problem specific, usually the first state is assumed to be drawn from a specified distribution and the final state is unconstrained. The SIR epidemic differs, often the epidemic is assumed complete corresponding to a known final state X_T. Although the epidemic model starts with a known number of infectives, which is taken to be 1, the time of the first infection is unknown and therefore so is the number of infectives at the first observation.

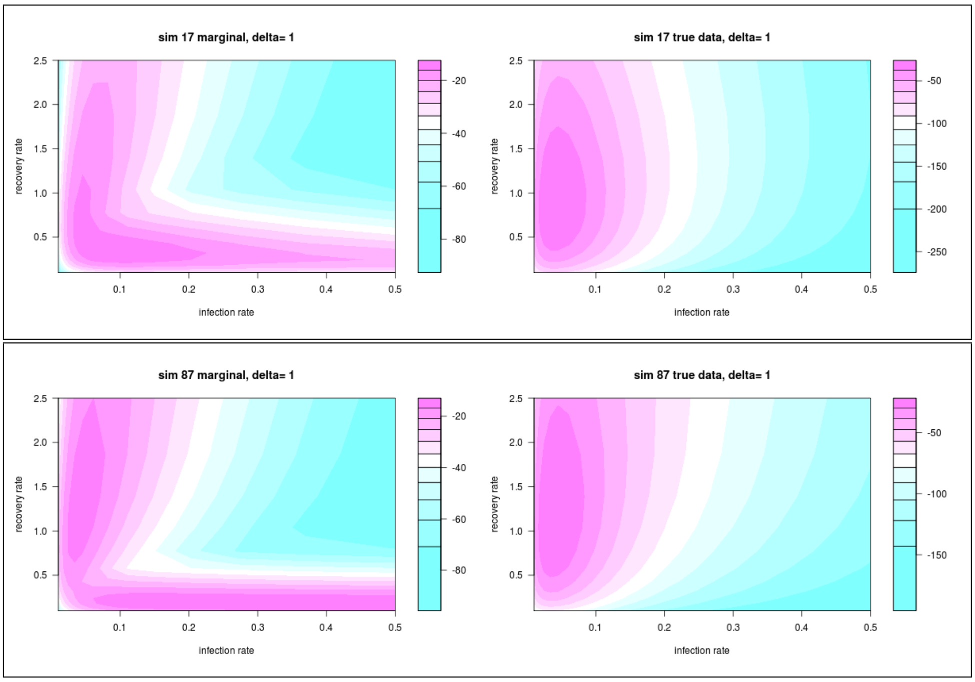


Figure 4.4.2: Exact marginal and full data log likelihood, for regularly observed GSE. 2 simulations numbered 17 and 87 both with $n_p = 50$. In the marginal likelihood it is not asssumed that the epidemic is complete.

Given values of the parameters θ the exact transition matrix can be calculated using the matrix exponential as shown previously. The likelihoods for example data can then be calculated using the forward recursions and can be maximised numerically to calculate a MLE. Contour plots of likelihoods for simulated epidemics are shown in figure 4.4.1 showing the greater dispersion of the marginal likelihood over the full data likelihood. Thes plots are for $\Delta t = 1$ similar plots (not shown) were obtained for $\Delta t = 0.1$. When the epidemic is complete the likelihood is uni-modal, however when this assumption is not made a bi-modal distribution is possible. The same data are analysed without the assumption of complete data and contours are shown in figure 4.4.2. The two modes corresponds to the true model and a "false" model with a high infection rate with a low recovery rate giving a large number of infectives at the last observation. These calculations can be used in a Bayesian framework by multiplying the likelihood by a prior for θ and normalizing by numerical integration, this approach was taken by Fearnhead and Meligkotsidou (2004) using a continuous time observation model, and an alternative way of calculating the exact likelihood.

MCMC Algorithms for Regularly Observed Epidemics A variety of MCMC algorithms have been developed for inference in HMM, a recent review is given in Fearnhead (2011). Investigation of the relative performance of the algorithms described there and other GIMH algorithms using the exact transition matrix for the GSE would be useful topic for future study. These algorithms are described in appendix C.

4.4.1 Concluding Remarks on Inference for Regularly Observed Epidemics

Although the matrix of transition rates for the GSE is very sparse the transition probability matrix is dense, which limits the size of problem that can be handled in this way. However the majority of elements are extremely small and replacing those less than some threshold with zero is a pragmatic approach. Some investigation would be needed to determine suitable values of the threshold.

The calculation of the matrix exponential is the limiting factor in the use of these MCMC algorithms and as mentioned previously further consideration of it's calculation is given in appendix A. As these algorithms all use the full state transition matrix the limit on population sizes is much smaller than that for the exact marginal algorithm for continuous observation given in section 4.3.

Inference for the binomial model described in section 2.4.1 is straightforward, with the same choice of algorithms as for the regularly observed GSE. The calculation of the transition matrix is much quicker than using the matrix exponential but suffers from the same memory limitations unless thresholding of probabilities is used.

4.5 Inference on Heterogeneous Epidemic Models

Inference for epidemic models that incorporate some form of heterogeneity has been widely studied, in order to make progress models that are believed to incorporate the most important components of variability but maintain simplicity have been studied. The two areas that have received the widest attention are household models and multitype epidemics, which are briefly reviewed.

Inference for Household Epidemic Models Household models provide a natural breakdown of the population where infection rates are expected to differ, they are also at a level at which data is frequently available. Inference can be based purely on the numbers infected in each household without knowledge of the times of infection and removal. The monograph by Becker (1989) shows how comparatively simple methods can be used. A limitation of this approach is that it is only applicable after an epidemic is complete, however it can be important for providing information relevant to vaccination strategies.

The same stochastic model can be applicable to farm animals kept in groups where the transmission within a group is much higher than that between groups an example of such an analysis is given by Hohle et al., 2005.

Inference for Multitype Epidemic Models The general multitype epidemic described in section 2.5.3 with m types has up to m^2+1 parameters $\psi_{i,j}\, i, , j \in 1 \ldots m$ and ρ, and usually they will not all be identifiable. An example of statistical inference on a simple version of this model is that of Becker (1989, chapter 5) who analyses an epidemic of a respiratory disease on Tristan da Cunha and shows that from final size information alone it is possible to identify a higher rate of transmission in school children. These models are important as outbreak control measures are often based around structures within populations (e.g. school closures).

A frequently studied model has both susceptibility and infectiousness varying between groups so that $\psi_{i,j} = c_i d_j$ where c_i is the infectiousness and d_j the susceptibility of individuals in groups i and j. Inference for this model has been studied by Britton (1998) who derives maximum likelihood estimators for the fully observed process and martingale estimators for the partially observed process.

An approach to the more general case has been studied in a series of papers by Demiris and O'Neill (2005a,b) who develop MCMC algorithms for inference from final-outcome data.

Inference for Other Heterogeneous Epidemic Models Inference for more complex models invariably uses MCMC approaches and are typified by the computationally intensive results obtained in Jewell et al. (2009b) which are illustrated by an analysis of the 2001 UK Foot and Mouth epidemic, and modelling the potential risk from a possible future Avian Influenza epidemic to the UK Poultry industry.

4.5.1 Inference on Bipartite Graph Epidemic Models

The bipartite graph epidemic BipE $(\mathbf{A}, \boldsymbol{\lambda}, \rho, g_i)$ defined in section 2.6.3 is in general not amenable to analytic inference, however MCMC techniques are in principle straightforward and are investigated. A variety of combinations of known and unknown parameters can be considered, many of which may be unidentifiable, also restrictions on the parameters could be considered. The case where $\mathbf{A}$ and g_i are known and $\boldsymbol{\lambda}$ and ρ are unknown is considered.

Two variants of the full data likelihood for the GSE were given above in equations 4.2.1 and 4.2.2, as the infection rate of BipE $(\mathbf{A}, \boldsymbol{\lambda}, \rho, g_i)$ depends on which individuals are infected when, the likelihood can only be sensibly considered for labeled data where the infection and removal times are associated to a row of $\mathbf{A}$ and so are available as paired times, which is also necessary when considering non-exponential distributions of removal times. The likelihood is an extension of that for the GSE, the time dependent infection rate in the GSE is $\lambda S(t) I(t)$ which is replaced by a sum of individual infection rates. The instantaneous rate of infections on a susceptible individual j was given in equations 2.6.1 and 2.6.2 which are repeated here, for each susceptible individual $j \in \mathcal{S}(t)$ the rate of infections at time t is

$$\eta_j(t) = \sum_{k \in V} a_{jk} \lambda_k I_k(t) \tag{4.5.1}$$

where

$$I_k(t) = \sum_{j \in U} a_{jk} \mathbf{1}[X_{j,t} = \mathsf{I}] = \sum_{j \in \mathcal{I}(t)} a_{jk} \tag{4.5.2}$$

and the likelihood where the n infections are at t_j^I and m removals at t_j^R and $t_j^I < t_j^R$ for $1 \leq j \leq n$ and $I(t) > 0$ for $t < \max_j(t_j^I)$ is

$$\begin{aligned} L &= \prod_{j \in \mathcal{B}} \eta_j\left(t_j^I\right) \exp\left\{-\int_{t_{\text{init}}^I}^{T_{\text{obs}}} \sum_{j=1}^{n_p} \eta_j(t)\, dt\right\} \times \\ & \prod_{j \in \mathcal{R}(T_{\text{obs}})} \rho \exp\left\{t_j^I - t_j^R\right\} \prod_{j \in \mathcal{I}(T_{\text{obs}})} \exp\left\{t_j^I - T_{\text{obs}}\right\} \end{aligned} \tag{4.5.3}$$

where $\mathcal{R}(T_{\text{obs}})$ is the set of individuals who are infected and recover by T_{obs}, $\mathcal{I}(T_{\text{obs}})$ the set of individuals still infectious at T_{obs}, $\mathcal{B} = \mathcal{R}(T_{\text{obs}}) \cup \mathcal{I}(T_{\text{obs}})$ is the set who have been infected by T_{obs} and $t_{\text{init}}^I = \min_j(t_j^I)$. To incorporate a general distribution for the recovery period the term inside the second product is replaced with its p.d.f. and the third by its c.d.f..

MCMC for Completely Observed Bipartite Graph Epidemics

When all the times are available and the parameters ρ and $\boldsymbol{\lambda}$ are unknown a standard approach to inference on a bipartite graph epidemic is to use MCMC, a range of possible algorithms exist, the appropriate choice will depend on the size and structure of $\mathbf{A}$. As an example a standard random walk Metropolis-Hastings algorithm was used on an example where $\mathbf{A}$ was 1000×21 and $\lambda_k = .5/n_k$, the resulting epidemic has has 626 infections and the number of infectives is plotted against time in figure 4.5.1.

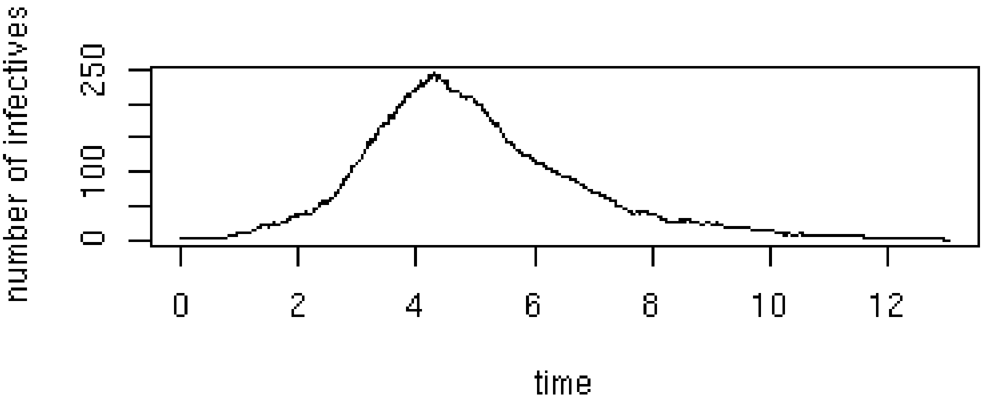


Figure 4.5.1: Bipartite graph epidemic example of 626 infections, a simulation with $\mathbf{A}$ 1000×21 and $\lambda_k = .5/n_k$

An example of the MCMC output, using a moderately informative prior, a set of independent exponential distributions mean 1, is shown in figure 4.5.2. The mixing was adequate without any special effort and the posteriors for the first two columns are compatible with the true values, 0.00058 0.00150, however the posterior for column 21, shown in the bottom plot is very close to the prior. This is because the non zero entries of column 21 of $\mathbf{A}$ were not infected and so provide no information on λ_{21}, which had a true value of 0.17000.

The main purpose of investigating such algorithms was as preparation for, and later to help in understanding difficulties, encountered with MCMC for the Indian Buffet Epidemic, however when combined with imputation of infection times they would provide a general means of investigating inference for other models with a bipartite graph representation.

4.6 Inference for the Indian Buffet Epidemic (IBufE)

The Indian Buffet Epidemic, which was introduced and defined in section 2.7.2 provides a flexible model which includes a wide range of heterogeneity in the contact process. When little is known about the structure of the contact process an al-

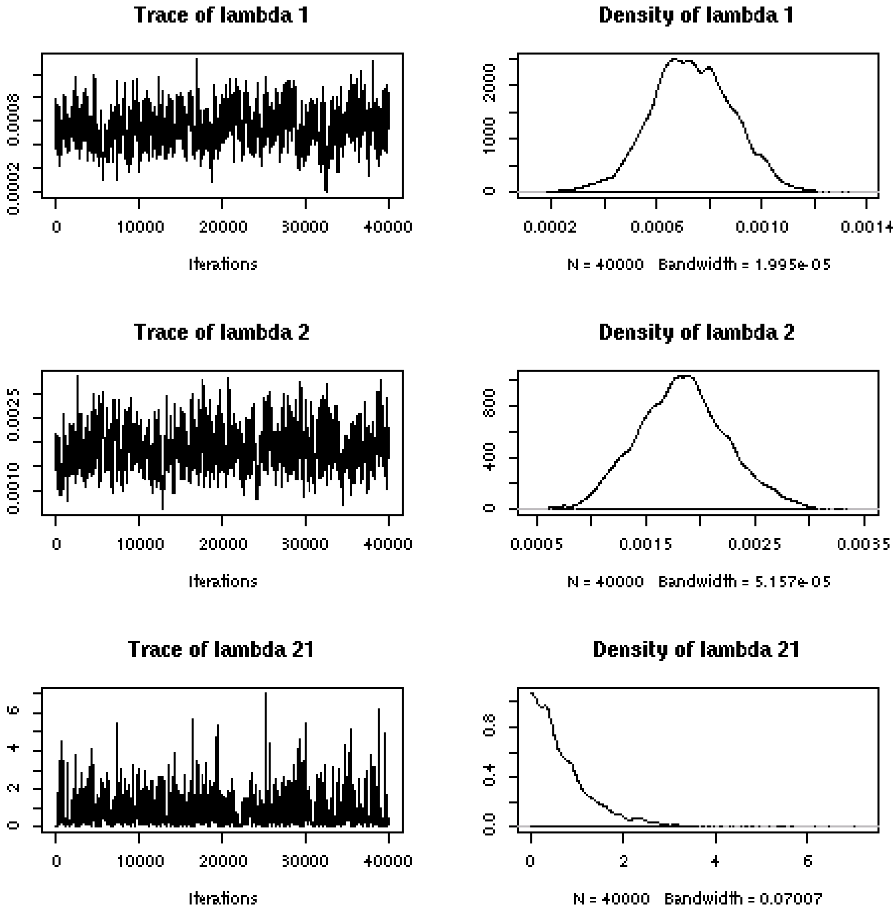


Figure 4.5.2: MCMC results for a bipartite graph epidemic

ternative to specific models such as a household model is to consider the structure as unknown and estimate some aspect of it. This has been done in papers such as Britton and O'Neill (2002) where the parameter p of an Erdös-Renyi graph is estimated. In a related way, here we consider marginal inference for the parameters of the Indian Buffet Epidemic not specific estimates of the underlying matrix $\mathbf{Z}$.

Inference for IBufE (θ, n_p, ξ, g_i) where both the parameters of the IBP (α, β) and the epidemic parameters (λ, ρ) are unknown and the population n_p, the infection rate scaling function $\xi(n, \lambda)$ and the initial infective distribution g_i are known, is in principal straightforward using MCMC, however in practice it provides several challenges. The remainder of this section describes the approaches taken and the progress made.

All of the approaches have considered the case where both infection times and removal times are available and are based on augmenting the data with the contact matrix $\mathbf{Z} \sim \text{IBP}\,(\alpha, \beta, n_p)$. The likelihood of the observed data consisting of n infections at t_j^I and m removals at t_j^R under $\text{IBufE}\,(\theta, n_p, \xi, g_i)$ when augmented with $\mathbf{Z}$ is the product of the likelihood of $\mathbf{Z}$ given by the appropriate choice from the three equations 2.7.1,2.7.2 or 2.7.3 and that of the epidemic given by equation 4.5.3. If the distribution of the initial infective g_i is other than uniform or deterministic then it must also be incorporated into the likelihood.

The calculation of the likelihood of the IBP can be relatively time consuming, as can sorting a matrix into `lof` form. The main terms involve only the sums of columns, m_k, but the term that differs between the sequential and `lof` forms is more complex. When generated by the sequential IBP process the term $\prod_{i=1}^{N} K_1^{(i)}!$ is available immediately. However if generated from an alternative method it is necessary to identify any repeated columns, and calculate $K_h!$, when $n_p > 50$ with high probability the repeats have $m_k = 1, 2$ or $m_k = n_p - 1, n_p$ but there is a non zero probability of other repeats as well. This can be calculated efficiently by noting that for columns to be the same they must have the same column sum m_k. Identifying the position of the first and last 1 in each column with the same m_k further identifies columns that cannot be repeats.

4.6.1 Random Walk Metropolis-Hastings for the Indian Buffet Epidemic

Some initial investigations looked at using a random walk for $\mathbf{Z}$ over $\mathcal{Z}_{\text{K}}$ (all $2^{n_p K}$ binary $n_p \times K$ matrices) for a fixed K and known $\beta = 1$. As the data include infection and removal times and the likelihood factorises, ρ can be independently estimated using maximum likelihood or a Bayesian analysis. Three random walk Metropolis-Hastings steps within a Gibbs framework are used, each samples from the conditional posterior distributions of each parameter given complete data and the other parameters.

Each step of the MCMC algorithm executes the following three substeps :

1. sample $\lambda \sim$MH using a random walk with Gaussian proposal

2. sample $\alpha \sim$MH using a random walk with Gaussian proposal

3. sample $Z \sim$MH on Z, proposal methods are described below, the acceptance probability is calculated using equation 2.7.1.

Where MH indicates a standard Metropolis-Hasting proposal and acceptance step as described in chapter 3. The first heuristic proposal for Z moves was at each step to do one of three moves "flip","swap" or "permute” chosen at random: where “flip” is pick i and k uniformly and set $z_{ik} = 1 - z_{ik}$, “swap” $z_{ik} \longleftrightarrow z_{i'k'}$, and “permute” was a permutation of rows within one column; this performed poorly with very low acceptance rates. A second proposal method was developed:

1. At each step: K i.i.d. column flip probabilities ψ_k are sampled from a beta distribution with parameters K and $0.8/K$.
2. Within each column, each bit is “flipped” independently with probability ψ_k.

These parameters were chosen so that the expected number of flips in each column is close to 1 but there is a small chance of a large number of flips, so that there is a small probability of a large step and a large probability of small step. This algorithm still performed poorly except on small problems and performance was very variable across data sets.

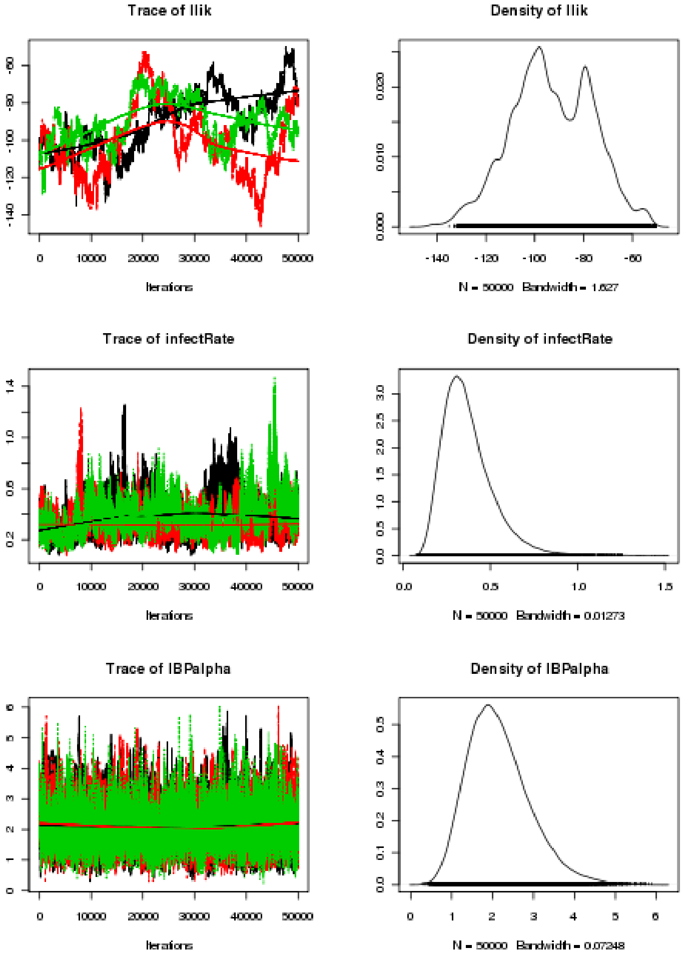


Figure 4.6.1: Example results for a heuristic random walk Metropolis-Hastings algorithm on an Indian buffet epidemic with population 28.

A typical example of diagnostic plots from three parallel chains on an example simulation with parameters $n_p = 28$, $K = 8$, $\alpha = 2$, $\lambda = 0.1$ is shown in figure 4.6.1. This appears to show reasonable convergence of α and λ, however the log-liklihood is showing large variations which were caused by jumping between modes with significantly different IBP matrices Z. This algorithm scaled very poorly with population size n_p and much longer runs and better tuning of proposals would have been required to achieve convergence on this example. This algorithm was not considered further.

4.6.2 Non Centered Parameterisation for the Indian Buffet Process

One of the problems identified in the first algorithm was the strong correlation between the IBP parameter α and the matrix $\mathbf{Z}$. On many other difficult MCMC problems the "non-centered" approach of Papaspiliopoulos et al. (2007) has proved useful, for instance in Neal and Roberts (2005), and so non centered algorithms where investigated.

The approach is to augment the problem with a uniformly distributed matrix U and a vector V from which the matrix $\mathbf{Z}$ is defined for given IBP parameters by a deterministic function $h(U, V, \alpha, \beta)$, so aiming to reduce the correlation between the imputed data and the parameters of interest.
Two mappings from the random U, V and the IBP parameters α and β to $\mathbf{Z} = h(U, V, \alpha, \beta)$ have been considered, the length of V is different for the two mappings. One based on the the finite K representation which produces $\mathbf{Z} \in \mathscr{Z}_{\mathrm{K}}$ and an alternative based on the sequential representation of the IBP which generates $\mathbf{Z} \in \mathscr{Z}_{\mathrm{seq}}$. The simpler mapping using the finite K representation was found to be more efficient, a large value of K is chosen, e.g. $2n_p$ and a uniform$(0,1)$ random $n_p \times K$ matrix U and a uniform$(0,1)$ random K vector $V = (V_1, V_2 \ldots V_K)$ are used. V is mapped to ψ using the inverse of the c.d.f. of the beta distribution

$$\psi_j = F^{-1}(v_j; \alpha\beta/K, \beta)$$

where $F(.; \alpha, \beta)$ is the c.d.f. of the beta distribution; $\mathbf{Z}$ is then simply given $Z_{i,j} = \mathbf{1}(U_{i,j} < \psi_j)$.

The sequential mapping $\mathbf{Z} = h(U, V, \alpha, \beta)$ is described according to the metaphor for the sequential representation of the IBP. A uniform$(0,1)$ random $n_p \times K_w$ matrix U and a uniform$(0,1)$ random n_p vector $V = (V_1, V_2 \ldots V_{n_p})$ are used. Conceptually U has an infinite number of columns, only a finite number K_w are used. V is mapped to the number of 'extra' dishes for each row and U is mapped to the entries of Z which are dependent on the previous customers. The mapping from U and V to Z, $h(U, V, \alpha, \beta)$ is defined by the sequential algorithm 4.1.

Algorithm 4.1 Sequential non-centered generation of IBP

```
1. for $i = 1 \ldots n_p$ calculate $N_i$ the number of extra dishes for
   customer $i$ as $N_i = F^{-1}(v_i; \alpha\beta/(\beta + i - 1))$ where $F(.;\alpha)$is
   the c.d.f. for Poisson mean $\alpha$.
2. set $K_+ = \sum_{i=1}^{n_p} N_i$
3. if $K_+ > K_w$
   (a) increase size of $U$ to $n_p \times K_+$ and set $K_w = K_+$
   (b) set new columns of $U$ to uniform$(0, 1)$
4. set $Z_{i,j} = 0$ for $i = 1 \ldots n_p$ and $j = 1 \ldots K_+$
5. set $Z_{1,j} = 1$ for $j = 1..N_1$
6. set vector $m = Z_{1,.}$
7. repeat the following steps for each $i$ in $2 \ldots N$ do
   (a) for each $j$ set $Z_{i,j} = 1$ if $U_{ij} > m_j/(\beta + i - 1)$
   (b) set $Z_{i,j} = 1$ for $j$ in the $N_i$ extra columns
   (c) for each $j$ set $m_j = m_j + Z_{i,j}$
```

Although this sequential algorithm appeared to have some advantages over the finite K algorithm, it in fact disrupts the key property of a non-centered representation. A small change in α gives a discontinuous change in the number of 'extra dishes' which causes different columns of U to be used and a small change in β gives different values of the cumulative counts m both resulting in potentially large changes in the components of the likelihood.

Various approaches to the top level MCMC algorithm are possible, the approach chosen was to produce samples of $p(\theta, U, V|T_I, T_R)$ where $\theta = (\alpha, \beta, \lambda)$ using a mix of M.H and Gibbs steps within a Gibbs MCMC by the following steps

Repeat the following steps

1. M.H. step on V, a random walk proposal is used which selects 10% of values and on these uses i.i.d. $N(0, \sigma_V)$
2. M.H. step on U, a random walk proposal is used which selects 10% of rows and 10% of columns and on these uses i.i.d. $N(0, \sigma_U)$
3. M.H. step on infection rate λ, a random walk proposal is used $\sim N(0, \sigma_\lambda)$
4. MH step on the IBP parameters α and β, a random walk proposal is used $\sim N(0, \sigma_\alpha)$ and$\sim N(0, \sigma_\beta)$

Concluding Remarks on Non-centering for the IBP

Initialisation of the non centered variables U and V requires further consideration, if they are generated uniformly a large number of attempts are often required to obtain a value of $\mathbf{Z}$ within the support of the posterior. An alternative is force a feasible initial value, the simplest approach is to set one column of U to all 1's. Using this or other feasible initialisations gives a Markov chain that has started close to a mode and invariably struggles to find other modes.

Although this algorithm performed much better than the previously described heuristic random walk algorithm, there were still problems, one of which was identified but not solved. The close relation between the infection rate and the IBP causes problems, a similar effect was noted by Britton and O'Neill (2002) and Neal and Roberts (2005) in their analyses of epidemics on Erdös-Renyi graphs. They found that $p\lambda$ was very close to constant in the posterior, where p is the edge probability in the graph and λ is the infection rate[3]. An additional problem is that the likelihood of the epidemic conditional on $\mathbf{Z}$ is the same as that obtained by repeating every column and halving the infection rate, denoting the matrix of size $n_p \times 2K$ with repeated columns as $\mathbf{Z}|\mathbf{Z}$ we have $\text{BipE}\left(\mathbf{Z}, \boldsymbol{\lambda}, \rho, g_i\right) = \text{BipE}\left(\mathbf{Z}|\mathbf{Z}, \boldsymbol{\lambda}/2, \rho, g_i\right)$ considered as distributions of the infection and removal times. The problem arises because the likelihood of $\mathbf{Z} \sim \text{IBP}\left(\alpha, \beta, n_p\right)$ is similar to $\mathbf{Z}|\mathbf{Z} \sim \text{IBP}\left(2\alpha, \beta, n_p\right)$ and so the posterior has many modes of similar heights, this strong multimodality makes the design of any random walk MCMC algorithm difficult. A reparameterisation using $\mathcal{R}_{\text{IB}}$ (see section 2.7.3) might alleviate this problem. Although this exact duplication could be identified within an algorithm, the multimodality arises also

[3] they use β for the infection rate, λ is used to avoid confusion with the IBP parameter β

from very similar values of the likelihood when columns only differ for non-infected individuals or by a single position. Also although in the bipartite graph epidemic repeated columns are redundant, in the Indian buffet epidemic they are a mechanism where higher infection rates within one group can be modelled.

4.6.3 IBP Proposal Distribution in the Support of the Posterior Distribution for the Indian Buffet Epidemic

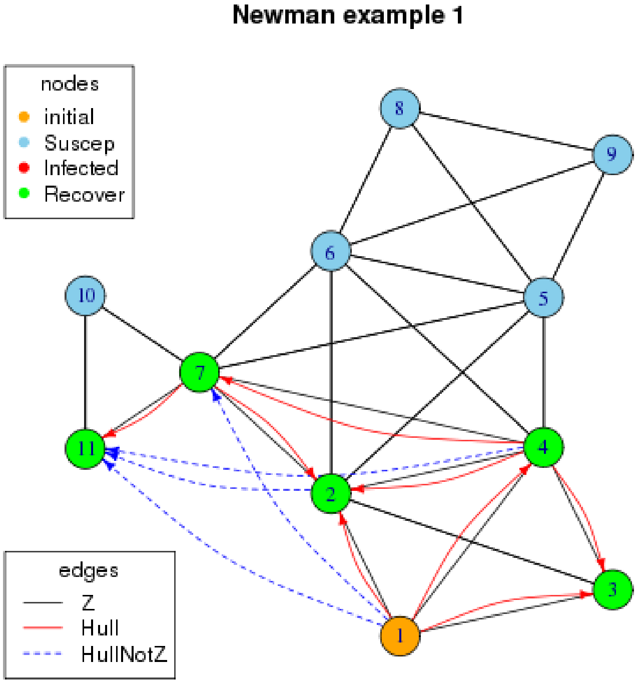
The difficulties described above in developing MCMC algorithms for inference for the Indian Buffet Epidemic provided insight that has enabled better algorithms to be developed. The key lesson from the different approaches which have been studied is that a proposal distribution for $\mathbf{Z}$ that is closer to the posterior is needed to achieve a reliable algorithm, in particular restricting the proposal to be close to the support while still being able to calculate its density is needed. Given a set of epidemic data times $\mathbf{T}$ a large fraction of $\mathbf{Z} \in \mathcal{Z}_{\text{lof}}$ are infeasible for $\mathbf{T} \sim \text{BipE}(\mathbf{Z}, \boldsymbol{\lambda}, \rho, g_i)$. That is the likelihood is zero or equivalently they are outside the support of the posterior of $\mathbf{Z}|\mathbf{T}, \boldsymbol{\lambda}, \rho$, which is denoted $\mathcal{Z}_{\text{sup}}(\mathbf{T})$. The initial infective in $\mathbf{T}$ is assumed to be within the support of g_i. The infection hull of a set of infection and removal times, is the set of sets of possible infectors and is denoted $\mathcal{H}(\mathbf{T})$, this is called the "set of suspects" in Britton and O'Neill (2002).

The infection hull $\mathcal{H}(\mathbf{T})$ can be used with many epidemic models, in particular any bipartite graph epidemic, to rapidly check if the structure is consistent with the observed or imputed times. To clarify the relation between the infection hull and the network an example simulation of an epidemic, run on an example bipartite network of 11 nodes (from figure 2.6.1) is shown in 4.6.2. The infection hull can be considered as a directed graph, with edges where infection is possible. The intersection of this graph and the graph on which the epidemic is runnning gives the possible, infection paths.

A difficulty with designing a proposal restricted to $\mathcal{Z}_{\text{sup}}(\mathbf{T})$ is that that the ratio of the size of the support $|\mathcal{Z}_{\text{sup}}(\mathbf{T})|$ to that of the sample space $|\mathcal{Z}_{\text{lof}}|$ is often very small and the probability changes significantly with the IBP parameters and the final size of the epidemic. The smallest probabilities are for small α, large β and large final size.

4.6.4 Efficient Independence Sampler for Indian Buffet Epidemics

The first algorithm using the infection hull $\mathcal{H}(\mathbf{T})$ combined a rapid check if the structure is consistent with the observed or imputed times with the sequential generation



vertex	Infection time	Removal time	Hull
1	0.000	2.822	
4	0.723	1.606	1,4,7
7	0.753	0.959	1,4
2	0.845	0.872	1
11	0.866	0.993	1,4
3	1.229	1.251	1,2,4,7

Figure 4.6.2: Example of a simulated bipartite epidemic showing the Infection Hull.

of an IBP to provide an independence sampler which generates $\mathbf{Z} \sim \mathrm{IBP}\left(\alpha, \beta, n_p\right)$, but is aware of $\mathcal{H}\left(\mathbf{T}\right)$ and on detecting that $\mathbf{Z} \notin \mathcal{Z}_{\text{sup}}\left(\mathbf{T}\right)$ abandons the generation and returns a flag. The MCMC algorithm immediately rejects the proposal without having to calculate the likelihood. This sampler is computationally efficient, in avoiding unnecessary calculations, but suffers from the usual problems of exponentially decreasing acceptance rates as the dimension of the problem, which is $\propto n_p$, increases.

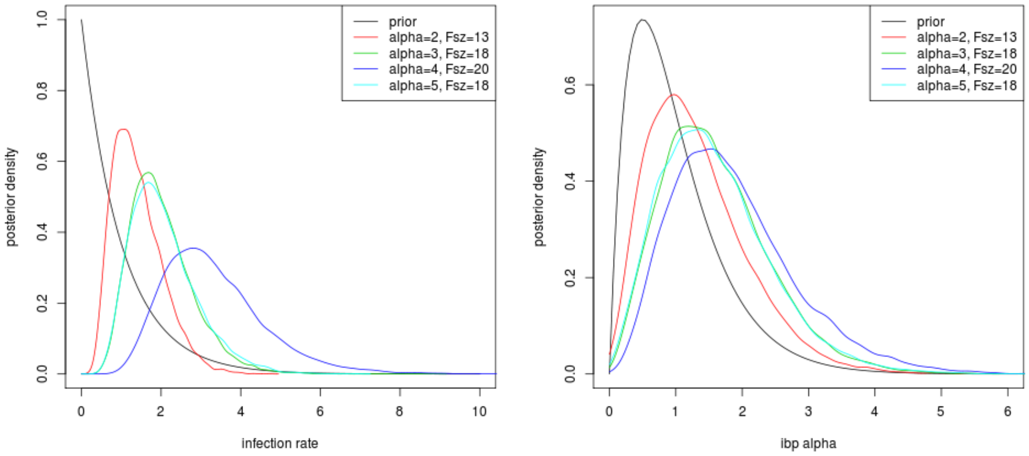


Figure 4.6.3: Examples of posteriors for IBufE parameters, from 4 simulated epidemics, using the efficient independence sampler.

The algorithm performs well on populations up to 25 and adequately up to 50, unfortunately for such small populations most epidemics are indistinguishable from a GSE and so the algorithm is of limited use. The resulting posterior distributions for the IBP parameters are close to the prior, and the posterior for the infection rate is wider than that obtained assuming a GSE and is strongly dominated by the observed final size, examples of posteriors are shown in figure 4.6.3 for runs on 4 example epidemics..

4.6.5 Algorithm for Sequential IBP Proposal Distribution in the Support of an Epidemic

The standard sequential algorithm for the generation of an IBP can be combined with $\mathcal{H}\left(\mathbf{T}\right)$ to provide a proposal distribution for $\mathbf{Z} \in \mathcal{Z}_{\text{sup}}\left(\mathbf{T}\right)$ which it is hoped is close to $\mathbf{Z} \sim \mathrm{IBP}\left(\alpha, \beta, n_p\right) \left|\mathbf{Z} \in \mathcal{Z}_{\text{sup}}\left(\mathbf{T}\right)\right.$. The difference is that the epidemic is considered in order of infection times and at each stage at least one bit is added to ensure that $\mathbf{Z} \in \mathcal{Z}_{\text{sup}}(\mathbf{T})$. The probability of $\mathbf{Z} \in \mathcal{Z}_{\text{sup}}(\mathbf{T})$ is also calculated sequentially, it is shown in algorithm 4.2. The resulting Z and $q(Z)$ can be used in a Markov Chain Independence sampler.

Algorithm 4.2 Sequential generation of IBP proposal in the support of an epidemic.

```
1. for $i = 1$ sample $N_1$ from $N_1 \sim \text{ZTPoisson}(\alpha)$, the zero truncated Poisson distribution.
2. for $i = 2 \dots n_p$ sample $N_i$ the number of extra columns from $N_i \sim \text{Poisson}(\alpha\beta/(\beta + i - 1))$
3. set $q_p$ to the sum of the log probabilities of $N_i$
4. for $j = 1 \dots N_1$ set $Z_{1,j} = 1$ and $m_j = 1$
5. set $K = N_1$
6. repeat the following steps for each $i$ in $2 \dots n_p$ do
    (a) set $H_i$ to $\mathcal{H}(i)$ the set of possible infectors of $i$
    (b) for each $j$ in $H_i$ set $p_j = m_j/(\beta+i-1)$ the IBP sequential algorithm probability
    (c) for each $j$ in $H_i$ set $Z_{ij} = 1$ with probability $p_j$ and update $q_p$
    (d) choose a $j$ from $H_i$ with probabilities $\propto p_j$ set $Z_{ij} = 1$ and update $q_p$
    (e) for each $j$ in $H_i$ set $m_j = m_j + Z_{ij}$
    (f) for $j = 1 \dots N_i$ set $Z_{1,j+K} = 1$ and $m_j = 1$
    (g) set $K = K + N_i$
7. return $Z$ and $q_p$
```

4.6.6 GIMH algorithm for Indian Buffet Epidemics

The grouped independence Metropolis-Hastings (GIMH) was described in chapter 3 and appears ideally suited to inference for the Indian buffet epidemic where interest is in the marginal posterior distribution of the parameters $\theta = (\alpha, \beta, \lambda, \rho)$ not in the posterior distribution of $\mathbf{Z}$. The algorithm combines the efficient independence sampler described in section 4.6.4 with the GIMH algorithm. The GIMH algorithm uses

Algorithm 4.3 GIMH algorithm for the Indian buffet epidemic

```
1. Sample initial value for $\theta = (\alpha, \beta, \lambda, \rho)$ from prior.
2. For $i = 1 \ldots n_z$ i.i.d. sample $\mathbf{Z}_i \sim \text{IBP}\,(\alpha, \beta, n_p)$, calculate $\tilde{\pi}^N(\theta)$ using equation 4.6.1.
3. Repeat the following steps a large number of times
    (a) sample $\theta' = q(\theta'|\theta)$
    (b) For $i = 1 \ldots n_z$ i.i.d. sample $\mathbf{Z}_i \sim \text{IBP}\,(\alpha', \beta', n_p)$, calculate $\tilde{\pi}^N(\theta')$
    (c) Accept $\theta'$ with probability $\min(\mathcal{A}, 1)$ where
            $\mathcal{A} = \dfrac{\tilde{\pi}^N(\theta')p(\theta')\,q(\theta|\theta')}{\tilde{\pi}^N(\theta)p(\theta)\,q(\theta'|\theta)}$
```

an estimate of the marginal likelihood $\tilde{\pi}^N(\theta)$ to compute the acceptance ratio for a Metropolis-Hastings algorithm for θ. Although the infection times are available and so the posterior for the removal rate ρ could be separated, if independent conjugate priors are used, the general case is considered with ρ within the GIMH framework. Using the IBP as the proposal for $\mathbf{Z}_i$ means that $\tilde{\pi}^N(\theta)$ given in equation 3.4.1 simplifies to

$$\tilde{\pi}^N(\theta) = \sum_{i=1}^{n_z} L\left(\mathbf{T}|\theta, \mathbf{Z}_i\right) p(\theta) \tag{4.6.1}$$

where the $\mathbf{Z}_i$ are n_z values i.i.d. $\sim \text{IBP}\,(\alpha, \beta, n_p)$, L is the conditional likelihood given by equation 4.5.3 and $p(\theta)$ is the prior for θ. The resulting algorithm is given in algorithm 4.3.

The results for this algorithm were variable, GIMH runs were performed on a set of simulated epidemics with population sizes of $n_p = 20, 50, 100, 150, 250$. Adequate mixing was obtained more frequently with the smaller populations but some examples sometimes worked well but occasionally, with a different seed for the random number generator, failed to mix adequately, other examples failed most of the time. For $n_p = 20$ mixing was generally adequate but the posterior for the IBP parameters was indistinguishable from the prior, an example is shown in figure 4.6.4. For $n_p = 50$ mixing was sometimes adequate on some examples but very poor on others, for $n_p = 100$ it was poor and for larger population sizes mixing was hopeless. The main problem was identified as the sticking of the GIMH algorithm which happens when a sampled value $\tilde{\pi}^N(\theta)$ is much bigger than the true value $\pi(\theta)$, the result is that the current state of the Markov chain is then maintained for a long time often

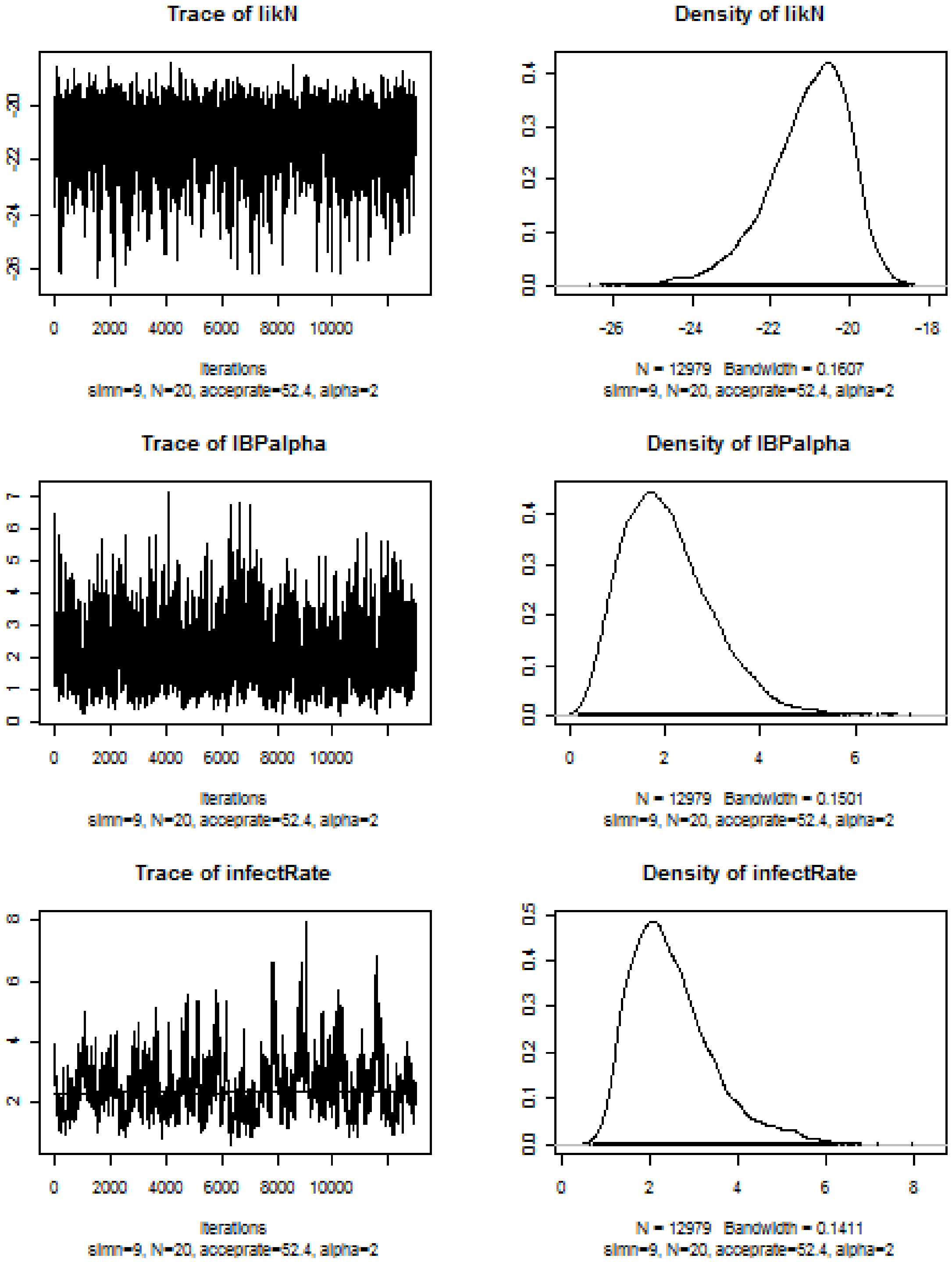


Figure 4.6.4: Example output from GIMH algorithm on an IBufE example with population size $n_p = 20$

10^3 samples or more. As a check on the algorithm and programming $\tilde{\pi}^N(\theta)$ has been evaluated at a grid of θ values for large values of $n_z = 400,000$, an example is shown in figure 4.6.5.

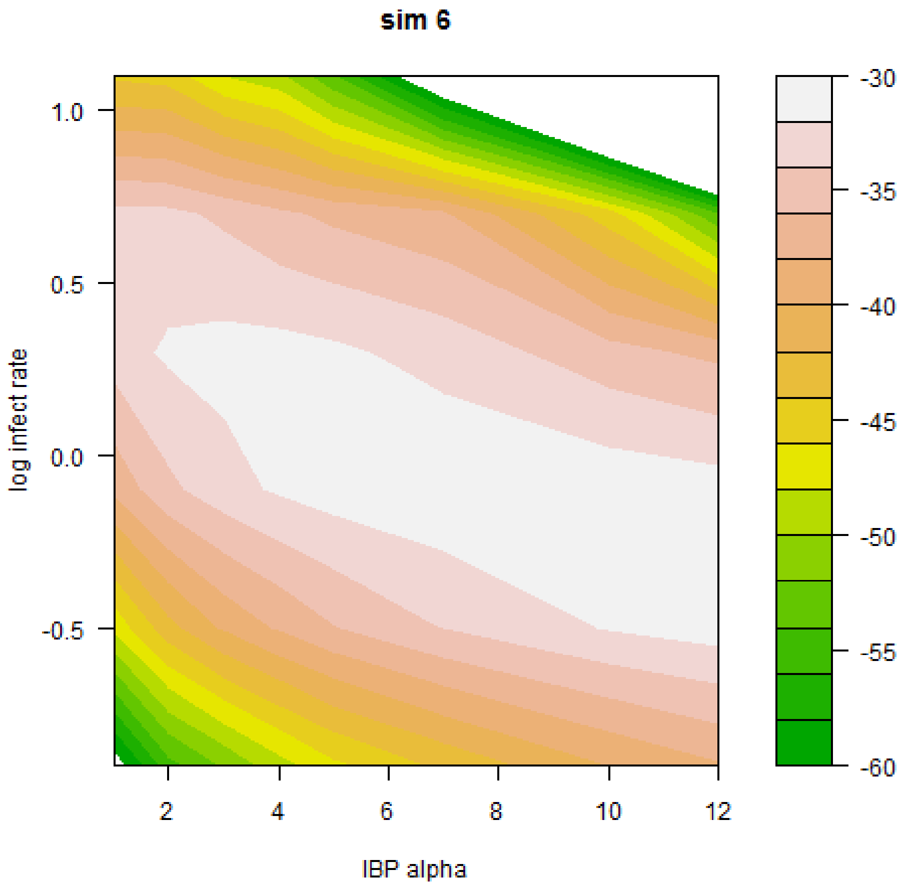


Figure 4.6.5: Contours of log likelihood for example IBufE $n_p = 20, \alpha = 3$, obtained using $\tilde{\pi}^N(\theta)$ with $N = 400000$ from the GIMH algorithm.

A new variant was investigated which is called the interleaved GIMH algorithm (IGIMH) which alternates a simple GIMH step with an i.i.d. independence proposal $q_\theta(.)$ for Z, with a local move on Z which targets the same distribution. This new algorithm is designed to overcome the sticking problem while remaining within the pseudo marginal framework and so inheriting the results in Andrieu and Roberts (2009). The performance was very similar to the GIMH, still getting badly stuck.

4.6.7 MCWM Algorithm for Indian Buffet Epidemics

The Monte Carlo within Metropolis algorithm (MCWM) was described in chapter 3 and is also ideally suited to inference for the Indian buffet epidemic and only requires a small change to the program used for the GIMH. Unfortunately it is known to be biased and as shown in chapter 3 the size of the bias can be large but is generally unknown. The MCWM does not usually suffer from getting stuck as badly as the GIMH as $\tilde{\pi}^N(\theta)$ is re-sampled for the existing state as well as the proposed state at each iteration. An inappropriate choice of the scaling for the proposal distribution $q(\theta'|\theta)$ can still result in some sticking.

In early results on very small examples this algorithm performed better than any of the other algorithms considered, at that stage, comparisons of the results shown in figure 4.6.6 with results from the GIMH and calculations of the marginal likelihood at a grid (shown in figure 4.6.5) does not reveal any obvious bias. Slow performance limited study on larger epidemics.

4.6.8 MCWM on Hagelloch Data

A real example dataset chosen to try and demonstrate the ability of the IBufE to detect heterogenity is the 1861 Hagelloch measles epidemic, which has 188 cases in a population of 197 children, (previously analysed by Neal and Roberts (2004) and subsequently by Britton et al. (2011) and Groendyke et al. (2011)). These data are known to involve spatial and classroom clustering and so it was thought they might show some heterogenity in the contact process when analysed without using the spatial locations. Computational improvements to the existing MCWM algorithm were necessary to work adequately on larger data sets, such as this. The improvements made to the basic MCWM algorithm result in the MCWM Hull reject algorithm 4.4.

Attempts were made to run the MCWM algorithm on the data, 188 paired infection and removal times, from the Hagelloch epidemic (obtained from the R epinet package). Various combinations of proposal scaling and priors were tried however none was fully acceptable. Two examples are shown in figure 4.6.7, the first attempt used, a moderately informative prior the second a more informative prior. The priors used are all offset gamma distributions, some with an offset from zero and are shown in table 4.6.1.

Algorithm 4.4 MCWM Hull reject algorithm for the Indian buffet epidemic

```
1. Sample initial value for $\theta = (\alpha, \beta, \lambda, \rho)$ from prior.
2. For $i = 1 \ldots n_z$ i.i.d. sample $\mathbf{Z}_i \sim \text{IBP}\,(\alpha, \beta, n_p)$,
    (a) check if $\mathbf{Z}_i \in \mathcal{Z}_{\text{sup}}\,(\mathbf{T})$ if not set $\mathbf{Z}_i$ to NULL
    (b) calculate $\tilde{\pi}^N(\theta)$ using equation 4.6.1.
3. Repeat the following steps a large number of times
    (a) sample $\theta' = q(\theta'|\theta)$
    (b) For $i = 1 \ldots n_z$ i.i.d. sample $\mathbf{Z}_i \sim \text{IBP}\,(\alpha', \beta', n_p)$,
        i. check if $\mathbf{Z}_i \in \mathcal{Z}_{\text{sup}}\,(\mathbf{T})$ if not set $\mathbf{Z}_i$ to NULL
    (c) calculate $\tilde{\pi}^N(\theta')$ using that first set of $\mathbf{Z}_i$ using 0 for the probability when
        $\mathbf{Z}_i$ is NULL
    (d) For $i = 1 \ldots n_z$ i.i.d. sample $\mathbf{Z}_i \sim \text{IBP}\,(\alpha, \beta, n_p)$
        i. check if $\mathbf{Z}_i \in \mathcal{Z}_{\text{sup}}\,(\mathbf{T})$ if not set $\mathbf{Z}_i$ to NULL
    (e) calculate $\tilde{\pi}^N(\theta)$ using the second set of $\mathbf{Z}_i$ using 0 for the probability
        when $\mathbf{Z}_i$ is NULL
    (f) Accept $\theta'$ with probability $\min(\mathcal{A}, 1)$ where
            $$\mathcal{A} = \frac{\tilde{\pi}^N(\theta')p(\theta')\,q(\theta|\theta')}{\tilde{\pi}^N(\theta)p(\theta)\,q(\theta'|\theta)}$$
```

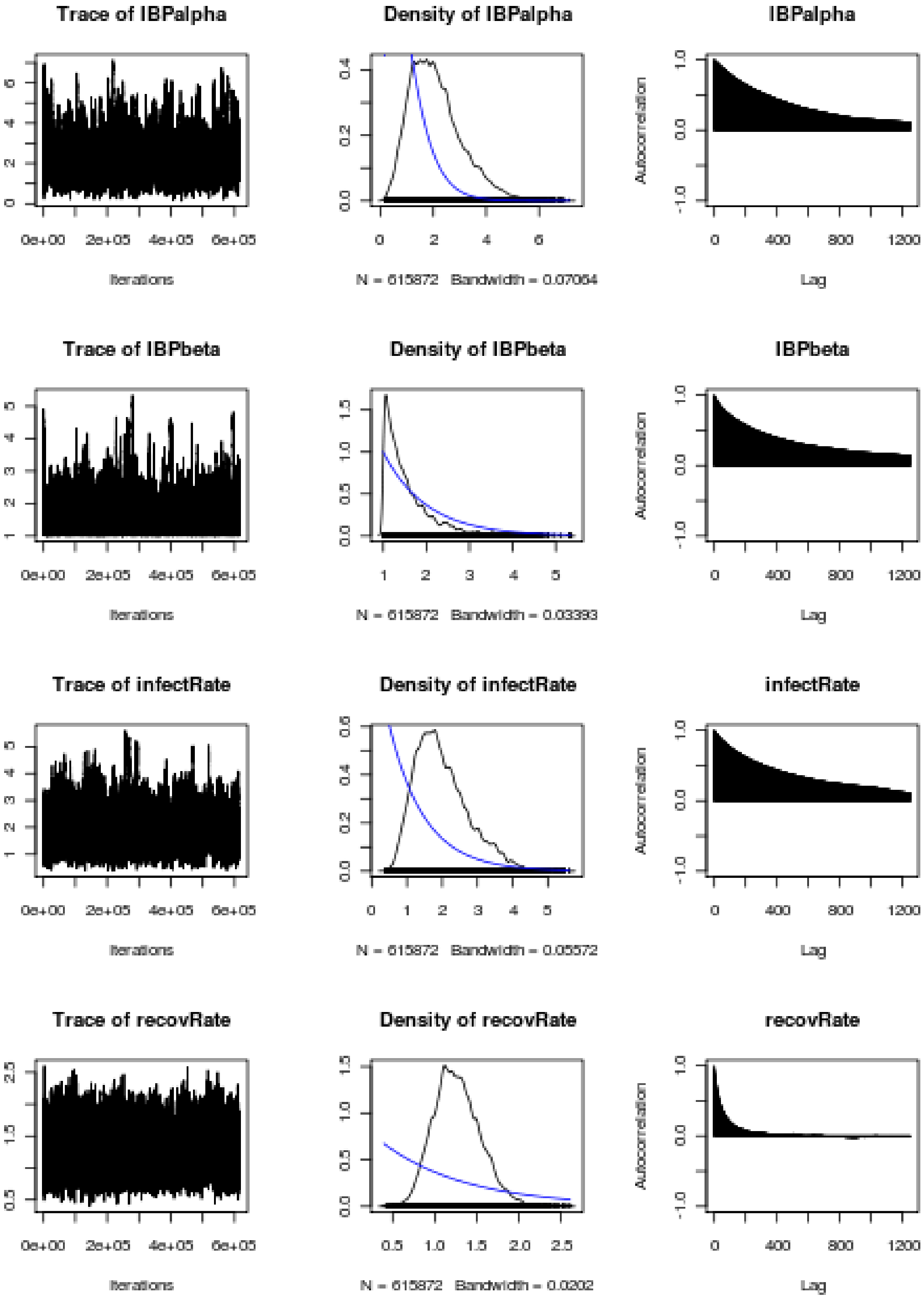


Figure 4.6.6: Example of the MCWM algorithm on a simulated IBufE, N=20.

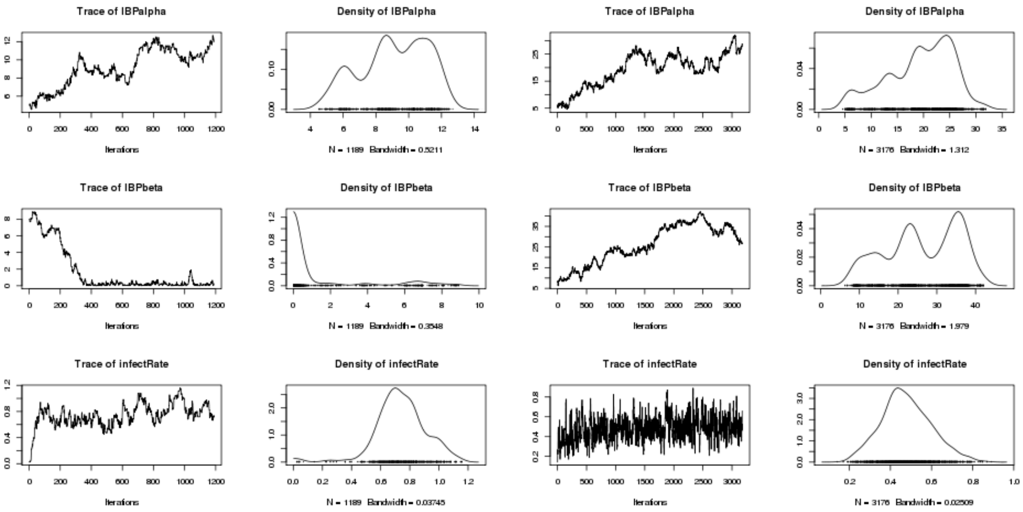


Figure 4.6.7: MCWM output from Hagelloch epidemic data. The priors used are shown in table 4.6.1.

		α		β			λ		ρ	
figures	shape	rate	offset	shape	rate	offset	shape	rate	shape	rate
4.6.6	2	2	0	1	1	1	1	1	1	1
4.6.7 l, 4.6.8	2	2	0	1	1	1	1	1	1	1
4.6.7 r, 4.6.9	2	0.1	1	2	0.1	1	1	10^{-3}	1	10^{-3}

Table 4.6.1: Priors used in MCWM runs on IBufE

4.6.9 Results for the MCWM and Hull Independence Sampler

Hull Independence Sampler on Hagelloch Data

The new algorithm appears to work well, trace plots showed good mixing, two sets of runs with different priors are shown here.

In the first a set of 5 parallel chains of length 5×10^5 had acceptance rates of 0.437, 0.621, 0.494, 0.436, 0.897, plots are shown in figure4.6.8. The performance and results are strongly influenced by the choice of prior and proposal distribution. An informative prior was used for α and β, while an uniformative prior was used for the infection rate parameter $\xi(n, \lambda)$. The independence proposal distribution used was also informative. In this first set there is little difference between the proposal and posterior for α and β.

In the second set a different prior for α and β was used which kept them

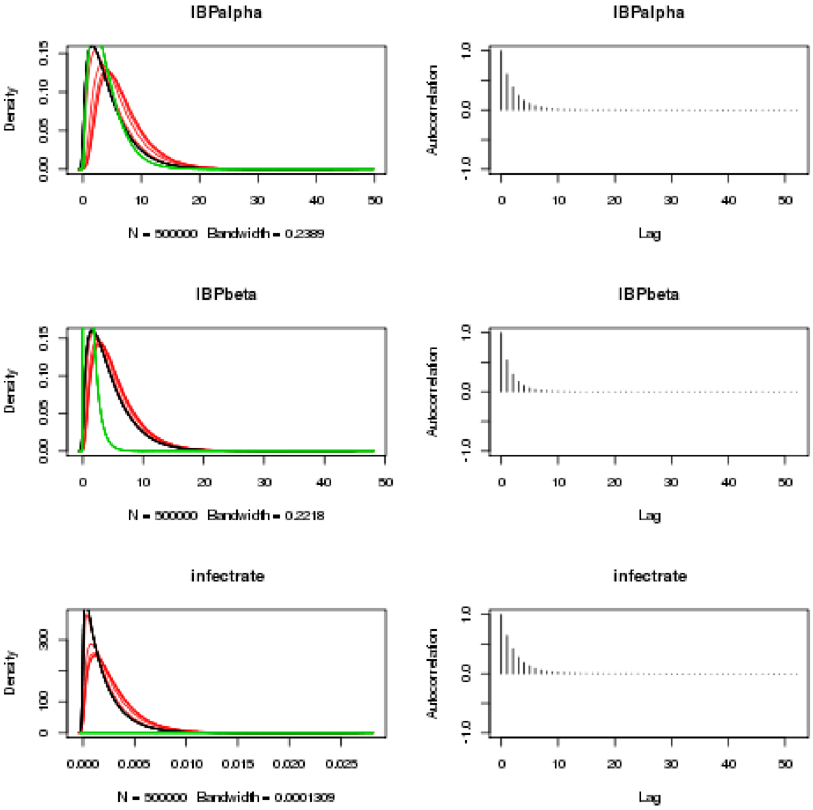


Figure 4.6.8: Density and autocorrelation plots for the Hull Independence Sampler on Hagelloch Data. The results of the density estimates are shown in red, the prior in green and the proposal in black.

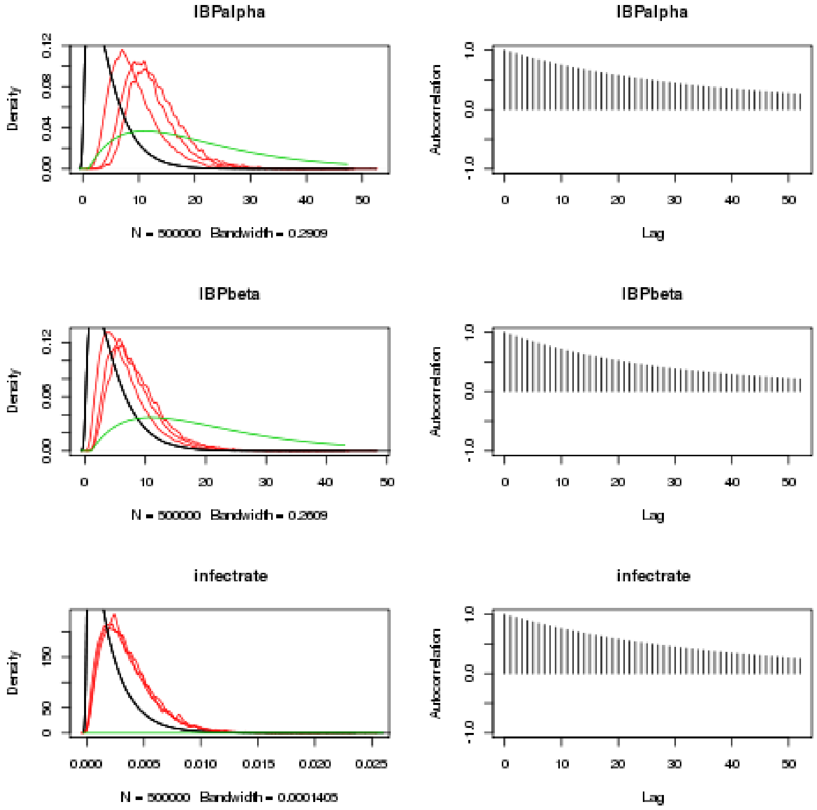


Figure 4.6.9: Density and autocorrelation plots for the Hull Independence Sampler on Hagelloch Data. The results of the density estimates are shown in red, the prior in green and the proposal in black.

from the low values where the contact structure is indistinguishable from homogeneous mixing. In this set 3 parallel chains of length 5×10^5 had acceptance rates of 0.030, 0.165, 0.055, plots are shown in figure4.6.9. The posteriors show some difference from both the prior and proposal distributions.

4.7 Concluding Remarks on Inference for Epidemic Models

Exact inference for the GSE, where the data consist of removal times has been demonstrated using a new algorithm based on caclulating the exact marginal likelihood using the matrix exponential. The exact marginal likelihood has been used to demonstrate that for an in progress epidemic the posterior can be bi-modal.

Although MCMC algorithms for linear models based on the IBP have been succesfully demonstrated inference for non-linear models such as the Indian buffet epidemic is significantly harder. The principal reason is the small ratio of the size of the support to the size of the IBP sample space.

Some understanding of the difficulties in MCMC inference for the Indian Buffet Epidemic has been achieved and algorithms have been developed that permit inference on complete data. Alternative sequential algorithms were also considered, it was hoped to combine the sequential process with the evolution of the epidemic. An algorithm based on infection trees inspired by the algorithm presented in Britton and O'Neill (2002) was also investigated, however although the Indian buffet process is exchangeable, if an ordering based on the epidemic such as infection times is used this is dependent on the structure and destroys this property.

The key lesson from the different approaches which have been studied is that a proposal distribution for $\mathbf{Z}$ that is closer to the posterior is needed to achieve a reliable algorithm, in particular restricting the proposal to be close to the support while still being able to calculate its density is needed. A n algorithm that does this has been implemented, a performance difficulty remains because the ratio of the size of the support $|\mathscr{Z}_{\text{sup}}(\mathbf{T})|$ to that of the sample space $|\mathscr{Z}_{\text{lof}}|$ is small and changes with the IBP parameters.

An alternative algorithm, still using the independence proposal, is the kernel Metropolis-Hasting algorithm described in chapter 3, which appeared promising on toy examples, but appears to suffer from similar sticking as the GIMH and bias as does the MCWM. The bias is thought to be affected by the wide changes in variance of the likelihood estimates as the parameters vary. When a reliable algorithm is identified for the situation considered here, where both infection times and removal times are available, it is planned to consider the more realistic situation of only having removal times by combining it with existing algorithms for imputing the infection times.

Chapter 5

Conclusions

Inference and model choice for partially observed epidemics provides a variety of challenges. This thesis has studied some related aspects of models for epidemics, their inference and some underpinning aspects of the GIMH algorithm, the key advances are summarised in the following paragraphs.

Exact Calculation in the General Stochastic Epidemic The use of the matrix exponential to facilitate exact calculations in the GSE has been demonstrated in providing the basis for inference in continuous and regularly observed epidemics. The use of the exact marginal likelihoood for inference is demonstrated in section 4.3 and the exact matrix of transition probabilities is used for HMM inference in section 4.4.

Bipartite Graph Epidemics The bipartite graph epidemic has been defined and shown to be a flexible framework which encompasses many existing models. It also provides a way in which a deeper understanding of the relation between existing models could be obtained.

Indian Buffet Epidemics The Indian buffet epidemic has been introduced as a non-parametric approach to modeling unknown heterogeneous contact structures in epidemics. Inference for the Indian buffet epidemic is a challenging problem, the algorithms which have been studied do not yet scale to the size of problem where significant differences from the GSE are apparent.

Importance Sampling and its Impact on GIMH and MCWM Evidence confirming and demonstrating the importance of understanding the tail behaviour of proposals in importance sampling has been presented in section 3.2. The adverse impact of heavy tailed proposals on the GIMH and MCWM algorithms has been shown.

Kernel Metropolis Hastings Algorithm A new algorithm, the KMH, has been proposed to provide an approximate algorithm for low dimensional marginal inference in situations where the GIMH algorithm fails because of sticking. The KMH has been demonstrated on a challenging 2-d problem. Further work is envisaged in two areas: a more efficient implementation in programming terms and more detailed understanding of the reasons for sticking and the size of the approximation and its impact on the posterior.

Appendix A

Matrix Exponentials and their Calculation

A.1 Matrix Exponentials

The matrix exponential can be defined for any square complex matrix A as

$$e^A = \sum_{j=0}^{\infty} A^j/j!$$

proofs of convergence and many other properties can be found in Higham (2008). In this thesis interest is only in finite square matrices with real entries. The most basic properties are that for any matrices A and B of the same dimension, integer n and scalar c :

$e^{cA} = e^c e^A$

$e^{nA} = (e^A)^n$.

Many but not all properties are inherited from the scalar exponential a significant difference is that

$e^{A+B} \neq e^A e^B$ unless A and B commute.

A.2 Calculating Matrix Exponentials

We have used calculations of e^{tQ} which in the context of inference have the potential for considerable speed up, with little loss of accuracy, using approximations. Calculating matrix exponentials is a well studied problem, see Moler and Van Loan (2003), and reliable software is available in R which has been used. The `expm` function in the Matrix package works on sparse matrices, the expm package contains newer (faster and more accurate) algorithms for `expm`() and includes `logm` and `sqrtm` but only works on standard matrices.

The condition number of the matrix exponential function can also be calculated to study the numerical accuracy of the caclulation, the results obtained have not shown any evidence of numerical instability and as interpretation of condition numbers is not straightforward a detailed analysis of numerical accuracy has not been performed. Further study would be needed to relate the magnitude of errors in the calculation of e^{tQ} with errors in a likelihood calculation that uses many calculations.

The acyclic property of the transition matrix for the SIR epidemic results in Q being upper triangular[1] and sparse, Stewart (1991) says that the references Severo (1969a) and Maire et al. (1987) contain more efficient ways of calculating e^Q when it is acyclic which could speed up all the algorithms but would not affect the relative performance significantly. An alternative approach which ensures the results are distributions is developed by van de Liefvoort and Heindl (2005). Bladt and Sorensen (2005) suggest using $\mathbf{B} = \mathbf{I} + \psi^{-1}\mathbf{Q}$ where $\psi \geq \max_i(-q_{ii})$ and the identity $\mathbf{Q}t = -\psi t\mathbf{I} + \psi t\mathbf{B}$ to calculate $\exp(\mathbf{Q}t)$ from

$$\exp(\mathbf{Q}t) = \sum_{j=0}^{\infty} e^{-\psi t} \frac{(\psi t)^n}{n!} \mathbf{B}^n$$

however as the row sums of $\mathbf{B}$ are all 1, truncating the series at n will give row sums < 1, an open question is how to adjust this and other algorithms to ensure the resulting matrix is stochastic.

A.2.1 Approximate Matrix Exponential Calculation

The greatest effect on the speed of inference algorithms which use the matrix exponential will come from using good approximations to $\exp(Q_{\theta'})$ calculated from $\exp(Q_\theta)$ when $|\theta - \theta'|$ is small. Because of the Markov property, or equivalently because tQ and sQ commute, we know $\exp((t+s)Q) = \exp(tQ)\exp(sQ) \quad \forall s, t \geq 0$. So when the change in θ is effectively a small change in timescale, $\theta' = \theta(1+\delta)$, we have $Q_{\theta'} = (1+\delta)Q_\theta$ and so

$$\exp(Q_{\theta'}) = \exp(Q_\theta)\exp(\delta Q_\theta) \approx \exp(Q_\theta)(I + \delta Q_\theta). \tag{A.2.1}$$

Changes in $\exp(Q_\theta)$ from changes in θ in the orthogonal direction, (a change in the R_0 for the epidemic) can by approximated using results of Al-Mohy and

[1] with the chosen lexicographic ordering of states

Higham (2008) who describe how to compute the Fréchet derivative which can be used to approximate $\exp(Q_{\theta'})$. The general approximation result they give is:

$$\begin{aligned}\exp(A(t+\theta h)) &= \exp\left(A+\theta\sum_{i=1}^{p} h_i \frac{\partial A}{\partial t}\right)\\ &= \exp\left(A\right)+\theta L\left(A,\sum_{i=1}^{p} h_i \frac{\partial A}{\partial t_i}\right)+o(\theta)\end{aligned}$$

where $L(A,B)$ denotes the Fréchet derivative of $\exp(A)$ in the direction B. In the case we are considering the matrix is linear in the parameters so the partial derivatives simplify.

Calculation of the Fréchet derivative of $\exp(A)$ is implemented in the R package `expm` and this has been used to calculate approximations to the matrix exponentials used in the GSE and compare them with the exact value. Initial results show the approximation appears good on Q of dimension 7381×7381, (which arise from the population of 120 used in the calculations for figure 4.3.1), for changes of up-to 20% in θ from (.01,.1), which would cover a large part of the posterior distribution as calculated by O'Neill and Roberts (1999). Further work is needed to quantify the range of $|\theta-\theta'|$ over which the approximations are reasonable and the size of the error induced in the posterior distributions.

A.3 Semi-symbolic computation

A new approach called semi-symbolic has been used for all the calculations involving matrices in this thesis. The approach has potential for wider use in any stochastic process where Q is a linear function of a low dimensional parameter θ, that is $Q_\theta = \sum_i Q_i\theta_i$ where Q_i are known fixed matrices. The uses here have all involved a matrix Q which is of the form $\lambda Q_I + \rho Q_R$ where Q_I and Q_R are sparse integer matrices. With a lexicographic ordering of states Q_I and Q_R are upper triangular and nil-potent. From the representation as 2 sparse integer matrices we can generate either

- sparse numeric matrices
- symbolic input for Maple, Mathematica
- manipulate directly (for example for aggregation of the states of a Markov chain)

only the first use has been reported above. This technique also permits delaying the substitution of numerical values for λ and ρ until needed which facilitates efficient computations.

Appendix B

Indian Buffet Process - Properties and Examples

B.1 Introduction

Distributions of many aspects of the Indian buffet process are given in Griffiths and Ghahramani (2011), various other derived distributions are useful in understanding the process and presented here.

B.2 Probability of Repeated Columns

The probability of repeated columns can be obtained from another representation of the IBP called a "history collection" in Griffiths and Ghahramani (2005). Each of the $2^N - 1$ possible columns has between 1 and N bits, which we denote their number by m. The number of each possible column has a Poisson distribution with rate $\gamma = \alpha \mathrm{B}(m, N-m+1)$ so the probability P_m of all being ≤ 1 is $\{(1+\gamma)\exp(-\gamma)\}^{\binom{N}{m}}$.

For small N we can calculate these directly, but rounding errors are significant for $N>$ 50 so taking logs and noting that $\gamma = \frac{\alpha}{m\binom{N}{m}}$, expanding $\log P_m$ gives

$$\log P_m = \frac{\alpha}{m} \sum_{j=1}^{\infty} \frac{(-\gamma)^j}{j+1} \tag{B.2.1}$$

The probability is concentrated on the ends of the range and using the exact expression for $m = 1$ and the first 3 terms of the expansion appears to give numerically accurate values. The probabilities for $N = 10$ and $\alpha = 2, 4$ are shown in figure B.2.1.

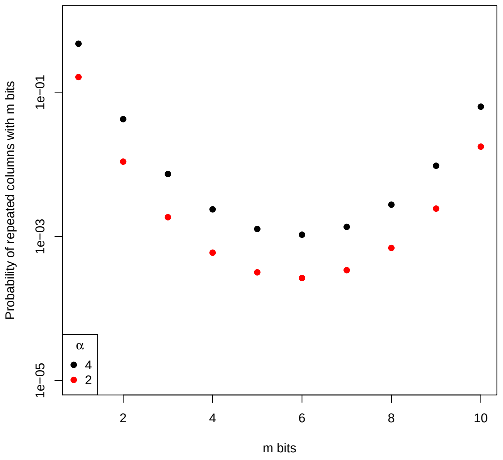


Figure B.2.1: Probabilities of any repeated columns with m bits in the IBP for $N = 10$ and $\alpha = 2, 4$

To examine the variation of these probabilities as N varies the probabilities for $m = 1, 2, N-1, N$ are plotted in figure B.2.2 together with some composite probabilities, the probability that there are any repeated columns is labeled "all", and is nearly coincident with the probability for $m = 1$ which is the dominany value. The line labeled "all>1" is the more relevant for the Indian buffet epidemic and is the probability that there is a repeated column with more than 1 bit. The line labeled "rest" is the probability of a repeated column which contains between 3 and $N - 2$ bits.

B.3 Beta Binomial Distribution and the IBP

The beta-binomial distribution is a compound distribution on $0 \ldots n$ which arises in several situations as the beta and binomial distributions are conjugate. It can be defined directly by its p.m.f.

$$\mathbb{P}\left(X = k\right) = \binom{n}{k} \frac{\mathrm{B}(k+\alpha, n-k+\beta)}{\mathrm{B}(\alpha, \beta)}$$

for $0 \leq k \leq n$ with parameters n a positive integer and $\alpha > 0$,$\beta > 0$.

Often it is derived as a mixture where the parameter p in the binomial distribution is drawn from a beta distribution so

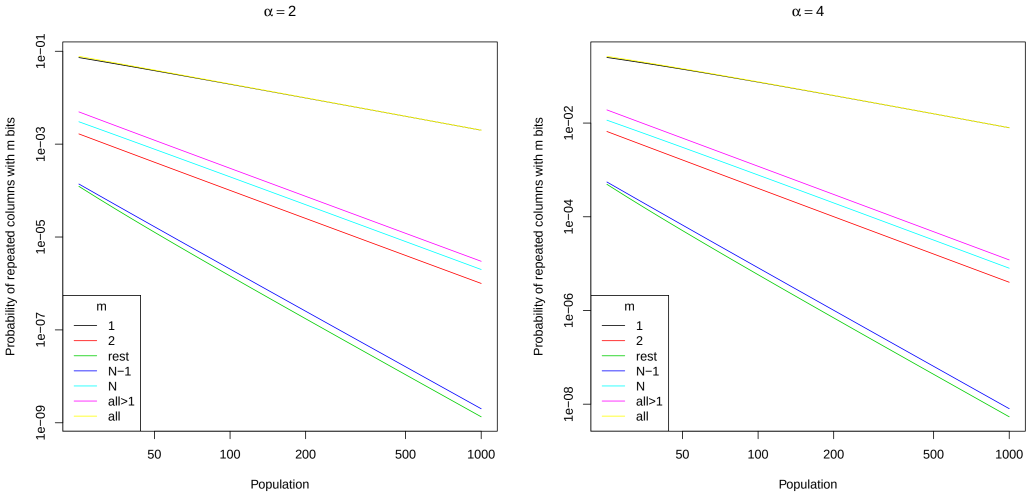


Figure B.2.2: Repeated column probabilities in the Indian Buffet Process

$$\mathbb{P}\left(X = k|p\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

and the density of p is $\frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$.
The mean is $\frac{n\alpha}{\alpha+\beta}$ and variance is $\frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
The c.d.f. is only available in terms of a generalized hypergeometric function.

IBP Row Sum Distribution

The distribution of the rowsums of $\mathbf{Z}$ is of interest, where $\mathbf{Z} \sim \text{IBP}\left(\alpha, \beta, N\right)$, we denote the rowsums by $X_i = \sum_k z_{ik}$. The distribution of the number of isolated individuals is bounded by the distribution of the number of $X_i = 0$, simulations show that the mean final size of epidemics conditioned on $\mathbf{Z}$ are approximately proportional to the sample mean of $\bar{X} = N^{-1} \sum_i X_i$. From the definition of the IBP the expected value of $\bar{X}$ is α, but although the distribution of an individual X_i is Poisson(α) they are strongly correlated. The joint distribution can be obtained recursively from the sequential representation as:

$$\begin{aligned}
X_1 &\sim \text{Poisson}\left(\alpha\right) \\
X_2 &\sim \text{binomial}(X_1, 1/(\beta+1)) + \text{Poisson}\left(\alpha\beta/(\beta+1)\right) \\
&\dots \\
X_i &\sim \text{binomial}(X_{i-1}, 1/(\beta+i-1)) + \text{Poisson}\left(\alpha\beta/(\beta+i-1)\right)
\end{aligned}$$

which can be evaluated numerically.

IBP Column Sum distribution

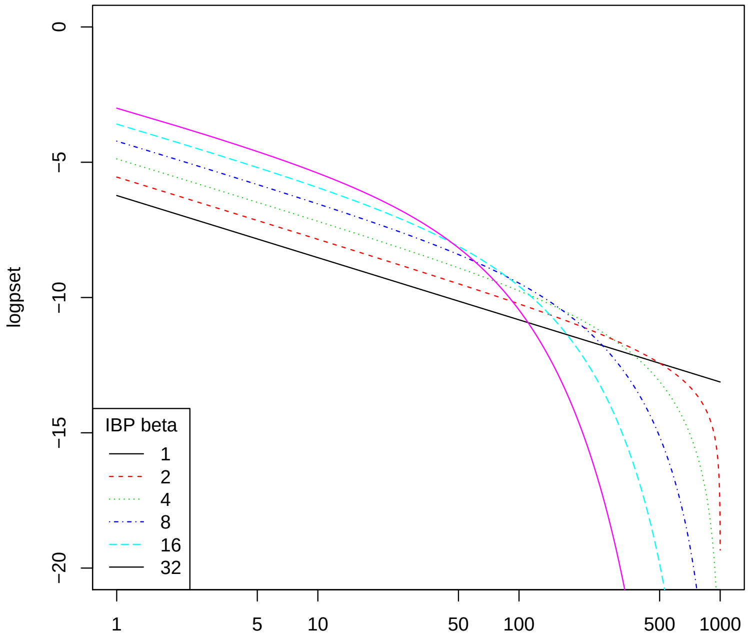


Figure B.3.1: Distribution of IBP column sums

The distribution of the column sums of the IBP are obtained by consideration of the finite K representation, where they have a beta binomial distributed with parameters $(N, \alpha\beta/K, \beta)$. As K increases the probability of zero increases and it is the distribution conditional on being greater than zero that is of interest.

Appendix C

MCMC Algorithms for Hidden Markov Models

C.1 Introduction

A wide variety of MCMC algorithms have been developed for inference in HMM, differences in the sizes of the state space, parameter space and observed data, can have a significant effect on the relative performance of these algorithms.

Many authors have considered inference in HMM, the book Cappé et al. (2005) provides a comprehensive treatment of the theory. Other relevant literature includes Scott (2002) and a recent review by Fearnhead (2011) and for state space models, Doucet and Andrieu (2001).

This appendix briefly describes candidate algorithms, including GIMH algorithms, for use in inference for the GSE or the binomial epidemic model using the exact transition matrix building on the likelihood calculations described in section 4.4.

C.2 Full posterior algorithms

The basic MCMC algorithms which have been widely studied in a range of applications perform separate θ and $X_{1:T}$ steps within a Gibbs framework with target $\pi(\theta, X_{1:T}|Y_{1:T})$. The distribution of the end states X_1 and X_T depends on the modelling assumptions. For the SIR model if the epidemic is assumed complete then $X_T = (n_p - Y_T, 0)$ the initial state is usually taken from a specified initial prior distribution.

C.2.1 Forward recursion backwards sampling algorithm

We use the forward recursions to calculate the 'alphas' which we then sample from in reverse to give a value of $X_{1:T} \sim \mathbb{P}\left(X_{1:T}|\theta, Y_{1:T}\right)$, this step is alternated with an update of $\theta \sim \mathbb{P}\left(\theta|X_{1:T}\right)$, in a standard MCMC algorithm. In continuous time epidemic models a conjugate prior is often used for θ, however there do not appear to be equivalent priors for the regular observation model, so a Metropolis-Hastings (MH) step will be used to update θ. The main limitation of this algorithm is that the calculation of the 'alphas' for large state spaces is slow $O(n_s^2 T)$.

C.2.2 Backward recursion forwards sampling algorithm

A variation on the previous algorithm, we use the backwards recursions to calculate the 'betas' which we then sample from to give a value of $X_{1:T} \sim \mathbb{P}\left(X_{1:T}|\theta, Y_{1:T}\right)$, this step is alternated with an update of $\theta \sim \mathbb{P}\left(\theta|X_{1:T}\right)$, in a standard MCMC algorithm. The performance of this algorithm will be very similar to the previous and the same comments apply, any differences will be largely due to the boundary conditions, it is expected that this algorithm may be better for complete epidemics and the previous for incomplete.

C.2.3 Single site Gibbs algorithm

A single site Gibbs algorithm where again updates of X and θ are alternated with X_t sampled as $X_t \sim \mathbb{P}\left(X_t|Y_t, X_{-t}, \theta\right) = \mathbb{P}\left(X_t|Y_t, X_{t-1}, X_{t+1}, \theta\right)$ where

$$\mathbb{P}\left(X_t = k|y_t, X_{t-1} = i, X_{t+1} = j, \theta\right) \propto p_{ik} p_{kj} g_k(y_t) \tag{C.2.1}$$

This algorithm is widely used in other HMM, and the usual choices of deterministic versus random scan choices of t must be made. The main problem reported is that of slow mixing, in the case of SIR epidemics this does not appear to be true. The need to recalculate the p_{ij} for each θ makes it unusable for epidemics in this simple form, however it forms the core of the algorithm described below C.3.5.

C.3 Marginal algorithms

In the case of epidemics there is usually no interest in the distribution of states and so marginal algorithms which target $\pi(\theta|Y_{1:T})$ directly are appropriate. This also has the advantage that one time costs of calculations for a given value of θ are used efficiently. The algorithms are based on the grouped independence Metropolis-Hastings (GIMH) algorithm Beaumont et al. (2002). In these algorithms several entire hidden paths are simulated, to simplify notation and correspond with that in Andrieu and Roberts (2009) we use $Z = X_{1:T}$ and Z_k for one of a set of n_z simulated paths through $\mathcal{X}^T$, which are then used in an importance sampler. The augmented target $\pi(\theta, X_{1:T}|Y_{1:T}) \propto \mathbb{P}\left(X_{1:T}|\theta\right)\mathbb{P}\left(Y_{1:T}|\theta, X_{1:T}\right)\pi(\theta)$ where $\pi(\theta)$ is the prior not the target. As the constant of proportionality $\mathbb{P}\left(Y_{1:T}\right)$ always cancels in a ratio we define $\pi(\theta, Z) = \mathbb{P}\left(X_{1:T}|\theta\right)\mathbb{P}\left(Y_{1:T}|\theta, X_{1:T}\right)\pi(\theta)$, the full data likelihood × the prior on θ.

C.3.1 Exact marginal

This algorithm is a variant of C.2.1 and should have the same statistical properties but be much quicker, for each proposed θ we sample several paths $X_{1:T}$.

Given a previous θ and the corresponding $\tilde{\pi}^N(\theta)$ repeat the following steps:

1. sample $\theta' \sim q(\theta, .)$
2. calculate α s using θ^*
3. for $k = 1 \ldots n_z$ sample $Z_k \sim q_{\theta^*}(.) = \mathbb{P}\left(X_{1:T}|\theta^*, Y_{1:T}\right)$ using the forward recursion backward sampling algorithm
4. compute $\tilde{\pi}^N(\theta^*) = n_z^{-1}\sum_{k=1}^{n_z}\pi(\theta^*, Z)/q_{\theta^*}(Z_k)$
5. accept θ^* based on the MH ratio
$$\frac{\tilde{\pi}^N(\theta^*)q(\theta^*, \theta)}{\tilde{\pi}^N(\theta)q(\theta, \theta^*)}$$

C.3.2 GIMH fixed $\hat{\theta}$

This is described in Fearnhead (2006), and uses one value of θ to generate all proposed Z, so they are exact for one value and used as importance samples for other values of θ.

After the initialization in steps 1 and 2 steps 3,4,5,6 are repeated many times.

1. choose $\hat{\theta}$ typically an approximation to the MLE.
2. calculate α s once for $\hat{\theta}$
3. propose $\theta^* \sim q(\theta, .)$
4. for $k = 1 \ldots n_z$ sample $Z_k \sim q_{\theta^*}(.) = \mathbb{P}\left(X_{1:T}|\hat{\theta}, Y_{1:T}\right)$ using the forward recursion backward sampling algorithm, the proposal for Z is independent of θ and $q_{\theta^*}(.) = q_{\hat{\theta}}(.)\ \forall\theta$.

5. compute $\tilde{\pi}^N(\theta^*) = n_z^{-1} \sum_{k=1}^{n_z} \pi(\theta^*, Z_k)/q_{\theta^*}(Z_k)$

6. accept θ^* based on the MH ratio

$$\frac{\tilde{\pi}^N(\theta^*)q(\theta^*, \theta)}{\tilde{\pi}^N(\theta)q(\theta, \theta^*)}$$

In many other HMM this algorithm would appear to have significant advantages, these are reduced significantly for the case we are considering as it is necessary to calculate $\exp(Q_{\theta^*})$ in order to calculate the importance weights.

Attempts to obtain smooth estimates of the likelihood Pitt (2002) are often hampered by the dependence of the proposal on θ, this approach could possibly combined with those techniques.

C.3.3 GIMH 1 step

This algorithm is the simplest GIMH using a simple forward simulation, although quick and simple to implement it is unlikely to be efficient and is described for completeness.

Given a previous θ and the corresponding $\tilde{\pi}^N(\theta)$ repeat the following steps:

1. propose $\theta^* \sim q(\theta, .)$

2. for $k = 1 \dots n_z$ sample $Z_k \sim q_\theta(.)$ i.i.d. where $q_\theta(.)$ samples $\mathbb{P}(X_1|Y_1) \propto \mathbb{P}(X_1, Y_1)$ then recursively for $t = 2 \dots T$ from $\mathtt{P}(X_t|X_{t-1}, Y_t) \propto \mathtt{P}(X_t, Y_t|X_{t-1})$

3. compute $\tilde{\pi}^N(\theta^*) = n_z^{-1} \sum_{k=1}^{n_z} \pi(\theta^*, Z_k)/q_{\theta^*}(Z_k)$

4. accept θ^* based on the MH ratio

$$\frac{\tilde{\pi}^N(\theta^*)q(\theta^*, \theta)}{\tilde{\pi}^N(\theta)q(\theta, \theta^*)}$$

Particularly at early stages of the epidemic this is likely to lead to cases where the epidemic dies out although the observations continue. An alternative is to use the look ahead algorithm below.

C.3.4 GIMH look ahead

This is a novel algorithm[1], although it could be applied to a general HMM it is designed for the SIR to prevent early termination of the simulated epidemic while remaining easy to calculate.

Given a previous θ and the corresponding $\tilde{\pi}^N(\theta)$ repeat the following steps:

[1] although the auxiliary particle filter approximates the same distribution

1. propose $\theta^* \sim q(\theta, .)$

2. for $k = 1 \dots n_z$ sample $Z_k \sim q_\theta(.)$ i.i.d. where $q_\theta(.)$ samples $\mathbb{P}\left(X_1|y_1, y_2\right)$ then for $t = 2 \dots T$

$$\mathbb{P}\left(X_t|x_{t-1}, y_t, y_{t+1}\right) \propto \mathbb{P}\left(X_t, y_t, y_{t+1}|x_{t-1}\right) = \sum_{x_{t+1}} \mathbb{P}(X_t|x_{t-1})\mathbb{P}(x_{t+1}|X_t)\mathbb{P}\left(y_t|X_t\right)\mathbb{P}\left(y_{t+1}|X_t\right) \tag{C.3.1}$$

3. compute $\tilde{\pi}^N(\theta^*) = n_z^{-1} \sum_{k=1}^{n_z} \pi(\theta^*, Z_k)/q_{\theta^*}(Z_k)$

4. accept θ^* based on the MH ratio

$$\frac{\tilde{\pi}^N(\theta^*)q(\theta^*, \theta)}{\tilde{\pi}^N(\theta)q(\theta, \theta^*)}$$

C.3.5 Marginal Gibbs

This is a novel algorithm which replaces the sampling of n_z i.i.d. paths each sequentially sampled for $t = 1 \dots T$ with a Markov sequence of n_z paths obtained from the single site Gibbs sampler described in C.2.3.

Given a previous θ and the corresponding $\tilde{\pi}^N(\theta)$ repeat the following steps:

1. propose $\theta^* \sim q(\theta, .)$

2. Sample Z_1 from an initial feasible distribution by,

   (a) set Z_0^p as a feasible path, e.g. 1 infective remains at each time step [2]

   (b) run n_b burn-in Gibbs steps, for $k = 1 \dots n_b$ $Z_k^b \sim q_\theta(.|Z_{k-1}^b)$ [3]

   (c) set $Z_1 = Z_{n_b}^b$

3. for $k = 2 \dots n_z$ sample $Z_k \sim q_\theta(.|Z_{k-1})$ where the proposal q_θ consists of a number of single site Gibbs steps, these can be fixed or random scan. Initial work has used a random ordering of all sites $1 \dots T$.

4. Calculate $\tilde{\pi}^N(\theta^*) = n_z^{-1} \sum_{k=1}^{n_z} \pi(\theta^*, Z_k)w_i$

5. accept θ^* based on the MH ratio

$$\frac{\tilde{\pi}^N(\theta^*)q(\theta^*, \theta)}{\tilde{\pi}^N(\theta)q(\theta, \theta^*)}$$

[2] until the final removal

[3] using the same sampler as described below

The Gibbs Markov chain ensures that the Z_k have an invariant distribution of $\mathbb{P}\left(X_{1:T}|\theta^*, Y_{1:T}\right)$ so using $w_i = 1$ gives an approximate algorithm, which appears to work. The weights w_i can also be derived from the Gibbs proposal Markov chain.

C.4 Particle MCMC algorithms

A related approach to inference in Markov jump processes is that described by Golightly and Wilkinson (2011) where a particle MCMC algorithm is reported to work well for noisy measurements. The SIR models considered here are assumed to be without observation noise and so the posterior distribution of $X_{1:T}$ is very concentrated. Preliminary investigations of this and similar algorithms using particle filters encountered difficulties in the particle filter stage as many of the sampled epidemics terminate early. Use of the auxiliary particle filter may alleviate this difficulty.

Appendix D

Notation

D.1 Abbreviations

EMC embedded Markov chain, see section 2.3.1,

lof left ordered form, see section 2.7.1,

GIMH Grouped Independence Metropolis-Hastings algorithm

GSE General Stochastic Epidemic

HMM Hidden Markov model,

KMH Kernel Metropolis-Hastings algorithm,

MCMC Markov chain Monte-Carlo, see chapter 3,

MCWM Monte Carlo within Metropolis algorithm

MH Metropolis-Hastings algorithm

RWM Random walk Metropolis algorithm

SAMH Stochastic Approximate Metropolis-Hastings algorithm, the generalisation of MCWM

SEMH Stochastic Exact Metropolis-Hastings algorithm, the generalisation of GIMH

SIR Susceptible Infectious Removed epidemic model

D.2 Basic statistical notation

- $\sim$ is used to indicate "is distributed as" e.g. $X \sim f$ where X is a random variable and f is some specification of a distribution
- $\Phi(x)$ the c.d.f. of the standard normal distribution $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-x^2/2)dx$
- $\Gamma(x)$ with a single argument is the gamma function.
- $\Gamma(x; \lambda, \nu)$ is the p.d.f. of the gamma distribution $\propto x^{\nu-1}\lambda^{\nu}\exp(-\lambda x)$.
- $\mathrm{B}(x, y)$ is the beta function $\mathrm{B}(x, y) = \dfrac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)}$
- beta$(x; a, b)$ is the p.d.f. of the beta distribution beta$(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a,b)}$
- Bernoulli distribution
- binomial(x;n,p) is the binomial distribution
- log(x) natural logarithms are used throughout
- $\pi(.)$ used in the MCMC chapter for an unspecified target probability measure, for which the interpretation is clear from the context.
- $\mathbb{P}(.)$ a general unspecified probability measure, for which the interpretation is clear from the context, often a discrete distribution.
- $\mathbb{E}_X$ to denote expectation with respect to the random variable X.
- $\mathbf{1}[A]$ the indicator function

D.3 Epidemic model notation

- λ the raw infection rate in an SIR epidemic
- ρ the recovery rate in an SIR epidemic
- $\mathcal{R}_0$ the basic reproduction number, for the GSE $\mathcal{R}_0 = n_p\lambda/\rho$
- $S(t)$ the number susceptible at time t
- $I(t)$ the number infectious at time t
- $C(t) = n_p - S(t)$ the total number that have been infected by time t , (this includes the initial infective)

- $R(t)$ the number recovered by time t
- $\mathfrak{R}_\infty = R(\infty)$ the total number infected at the end of the epidemic

Note that $R(t)$ and $C(t)$ are counting processes, while $I(t)$ is the difference of two counting processes. The associated times are

- T_j^I, t_j^I the time of the jth infection r.v. and a value
- $T_j^R, t_j^R$the time of the jth recovery r.v. and a value
- T_1^R, t_1^R the time of the first recovery r.v. and a value
- $I\left(T_1^R\right)$ the number infectious immediately after the first recovery at T_1^R
- l a value of $I\left(T_1^R\right)$ the number infectious immediately after the first recovery

D.4 Matrix notation

- matrices are generally in bold, and may be defined by their components as $\mathbf{P} = (p_{ij}\, i,, j \in \mathcal{S})$
- a term of a matrix or matrix expression is denoted by $[\mathbf{P}]_{ij}$to indicate the i, jth term
- $\mathbf{I}$ indicates an identity matrix of appropriate dimension
- diag($\mathbf{A}$) is a vector formed from the diagonal of a matrix $\mathbf{A}$
- diag(v) is a matrix with v on the diagonal and 0 elsewhere

D.5 IBP

- $\mathcal{Z}_{\text{lof}}$ the set of equivalence classes of binary matrices with distinct left ordered forms
- $\mathcal{Z}_{\text{seq}}$ the set of possible binary matrices generated by the sequential IBP
- $\mathcal{Z}$ one of $\mathcal{Z}_{\text{lof}}$ or $\mathcal{Z}_{\text{seq}}$, which will be clear from the context
- lof$(\mathbf{Z})$ left ordered form, a many to one mapping of all binary matrices → $\mathcal{Z}_{\text{lof}}$
- $\mathcal{Z}_{\text{uc}}$ the set of equivalence classes of matrices with the same unique columns

- $\mathcal{Z}_{\text{u2}}$ the set of equivalence classes of matrices with the same unique columns, and all column counts ≥ 2.
- $\text{IBP}\,(\alpha, \beta, N)$ the IBP distribution
- H_N^β a generalisation[1] of the standard harmonic number, $H_N^\beta = \sum_{j=1}^N \frac{\beta}{j+\beta-1}$

D.5.1 Indian Buffet Epidemic

- $\xi\,(n, \lambda)$ the infection rate scaling function defines the within group infection rate for a group of size n, e.g. $\xi\,(n, \lambda) = \lambda/n$ or $\xi\,(n, \lambda) = \lambda/\sqrt{n}$
- a $\mathcal{Z}_{\text{bip}}^{\boldsymbol{\lambda}}$ the set of equivalence classes of matrices which give the same bipartite epidemics with infection rate vector $\boldsymbol{\lambda}$.
- a $\mathcal{Z}_{\text{IBufE}}^{\xi}$ the set of equivalence classes of matrices which give the same bipartite epidemics with scaling function ξ.
- g_i a distribution over the population for the initial infective.

D.6 MCMC acceptance terminology and notation

Authors differ on both the words and notation for the many closely related quantities, here we define them as used in this thesis.

D.6.1 Definitions

invariant π is an invariant distribution for a Markov chain with transition kernel $\mathcal{K}$ if $\pi(dy) = \int_{\mathcal{X}} \pi(dx)\mathcal{K}(x, dy)$. If the state space is finite then π is a left eigenvector of the transition matrix with eigenvalue 1.

D.6.2 Metropolis Hastings

A key difference between authors is whether a conditional or measure theory type notation is used for proposals and Markov chain densities, eg from x to y can be written as q(y|x) or Q(x,dy), q(x,y)

proposal distribution The density of a proposed move from x to y (w.r.t. an implied measure usually Lebesgue on $\mathbb{R}^d$ or counting on $\mathbb{Z}^d$)

$q(y|x)$

[1] not to be confused with the standard generalisation $\sum_j j^{-m}$

acceptance ratio a non negative number

$\mathcal{A}(y|x) = \frac{\pi(y)\, q(x|y)}{\pi(x)\, q(y|x)}$

specific acceptance probabilty The probability that in a MH step a proposed move from x to y is accepted

alternative values are possible e.g. Barker

$\alpha(y|x) = \min(1, \mathcal{A}(y|x))$

acceptance probabilty The probability that in a MH step a move from x is accepted, depends on x and proposal

$\mathfrak{a}(x) = \int q(y|x) \min(\mathcal{A}(y|x), 1) dy$

in SEMH the state is x, w so $\mathfrak{a}(x, w)$ is used.

rejection probabilty The probability that in a MH step a move from x is accepted, depends on the state x and proposal

$\mathfrak{r}(x) = 1 - \mathfrak{a}(x)$

in SEMH the state is x, w so $\mathfrak{r}(x, w)$ is used.

acceptance rate The probability that at equilibrium a proposal is accepted, it is the expected value of $\mathfrak{a}(X)$.

$a = \int \pi(x) \int q(y|x) \alpha(y|x) dy dx$

usually dependence on parameters of the proposal or target will be indicated e.g. for scaling

$a(s) = \int \pi(x) \int q_s(y|x) \min(\mathcal{A}(y|x), 1) dy dx$

acceptance bound An upper bound on $\mathfrak{a}(x)$

$\zeta(x) = \int \frac{\pi(y)}{\pi(x)} q(y|x) dy$

D.6.3 SEMH and SAMH

expected acceptance/rejection probabilty The expectation over W of the probability that a move from x is accepted.

$\bar{\mathfrak{a}}(x) = \mathbb{E}(\mathfrak{a}(x, W))$

small move acceptance/rejection probabilty The probability that a very small move from x is accepted, depends on the state x, w and proposal $q(y|x) = \delta_x(y)$

$\mathfrak{a}_\delta(x) = \int_{\mathcal{X}} \int_{\mathcal{W}} \min(\mathcal{A}(y, w'|x, w), 1) f_W(w'|y) q(y|x) dw' dy$

$\mathfrak{r}_\delta(x) = 1 - \mathfrak{a}_\delta(x)$

specific acceptance probabilty The probability that in SAMH a proposed move from x to y is accepted
$\bar{\alpha}(y|x) = \int_{\mathcal{W}} \int_{\mathcal{W}} \min(\frac{z}{w}\mathcal{A}(y|x), 1) f_W(z|y) f_W(w|x) dw dz$

- $\tilde{\pi}$ to indicate an estimate of π which will be calculated at a finite set of points,
- $\mathfrak{r}(x, w)$ the rejection probability for Stochastic exact Metropolis-Hastings when the sampled state is x, w ,
- $\mathfrak{a}(x)$ expected acceptance rate for Stochastic exact Metropolis-Hastings when the sampled weight is w,
- $f_W(w|x)$ p.d.f. of weight W
- $\mathcal{K}(x, dx)$ a Markov chain transition kernel,

Bibliography

Al-Mohy, A. and Higham, N. (2008). Computing the Fréchet derivative of the matrix exponential, with an application to condition number estimation. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1639–1657.

Allen, L. (2008). An introduction to stochastic epidemic models. In *Mathematical epidemiology*, chapter 3. Springer.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Springer.

Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Bailey, N. T. J. (1953). The total size of a general stochastic epidemic. *Biometrika*, 40(1-2):177–185.

Bailey, N. T. J. (1964). *The elements of stochastic processes, with applications to the natural sciences*. Wiley.

Ball, F. (1986). A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Advances in Applied Probability*, 18(2):289–310.

Ball, F., Britton, T., and Sirl, D. (2010a). Household epidemic models with varying infection response. *Journal of Mathematical Biology*.

Ball, F. and Nåsell, I. (1994). The shape of the size distribution of an epidemic in a finite population. *Mathematical Biosciences*, 123(2):167–181.

Ball, F., Sirl, D., and Trapman, P. (2010b). Analysis of a stochastic sir epidemic on a random network incorporating household structure. *Mathematical Biosciences*, 224(2):53–73.

Ball, F., Sirl, D., and Trapman, P. (2014). Epidemics on random intersection graphs. *Annals of Applied Probability*, 24:1081–1128.

Barber, D. (2008). Clique matrices for statistical graph decomposition and parameterising restricted positive definite matrices. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 26–33.

Barbour, A. (1975). The duration of the closed stochastic epidemic. *Biometrika*, 62(2):477.

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate Bayesian Computation in population genetics. *Genetics*, 162(4):2025.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Becker, N. and Hasofer, A. (1997). Estimation in epidemics with incomplete observations. *Journal of the Royal Statistical Society: Series B*, 59(2):415–429.

Becker, N. and Yip, P. (1989). Analysis of variations in an infection rate. *Australian & New Zealand Journal of Statistics*, 31(1):42–52.

Becker, N. G. (1989). *Analysis of infectious disease data.* Chapman & Hall.

Bédard, M. (2008). Optimal acceptance rates for metropolis algorithms: Moving beyond 0.234. *Stochastic Processes and their Applications*, 118(12):2198–2222.

Begon, M., Bennett, M., Bowers, R., French, N., Hazel, S., and Turner, J. (2002). A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiology and infection*, 129(01):147–153.

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, 76(4):633.

Bladt, M. and Sorensen, M. (2005). Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B*, 67:395–410.

Brauer, F., Van den Driessche, P., and Wu, J. (2008). *Mathematical epidemiology.* Springer.

Britton, T. (1998). Estimation in multitype epidemics. *Journal of the Royal Statistical Society: Series B*, 60(4):663–679.

Britton, T., Deijfen, M., Lageras, A., and Lindholm, M. (2008). Epidemics on random graphs with tunable clustering. *Journal of Applied Probability*, 45:743–756.

Britton, T., Kypraios, T., and O'Neill, P. (2011). Inference for epidemics with three levels of mixing: Methodology and application to a measles outbreak. *Scandinavian Journal of Statistics*, 38(3):578–599.

Britton, T. and O'Neill, P. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29(3):375–390.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag.

Charleston, B., Bankowski, B., Gubbins, S., Chase-Topping, M., Schley, D., Howey, R., Barnett, P., Gibson, D., Juleff, N., and Woolhouse, M. (2011). Relationship between clinical signs and transmission of an infectious disease and the implications for control. *Science*, 332(6030):726–729.

Chen, Y., Diaconis, P., Holmes, S., and Liu, J. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–120.

Clancy, D. and O'Neill, P. (2008). Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis*, 3(4):737–758.

Daley, D. and Gani, J. (1999). *Epidemic Modelling*. Cambridge Univ. Press.

Danon, L., Ford, A., House, T., Jewell, C., Keeling, M., Roberts, G., Ross, J., and Vernon, M. (2011). Networks and the epidemiology of infectious disease. *Interdisciplinary perspectives on infectious diseases*, 2011.

Decreusefond, L., Dhersin, J.-S., Moyal, P., and Tran, V. (2010). Large graph limit for an SIR process in random network with heterogeneous connectivity.

Demiris, N. and O'Neill, P. (2005a). Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal of the Royal Statistical Society: Series B*, 67(5):731–745.

Demiris, N. and O'Neill, P. (2006). Computation of final outcome probabilities for the generalised stochastic epidemic. *Statistics and Computing*, 16(3):309–317.

Demiris, N. and O'Neill, P. (2005b). Bayesian inference for epidemics with two levels of mixing. *Scandinavian Journal of Statistics*, 32(2):265–280.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

Diekmann, O., Heesterbeek, J., and Metz, J. (1990). On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):365–382.

Diggle, P. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical methods in medical research*, 15(4):325.

Doucet, A. and Andrieu, C. (2001). Iterative algorithms for state estimation of jump Markov linear systems. *Signal Processing, IEEE Transactions on*, 49(6):1216–1227.

Durrett, R. (2007). *Random graph dynamics*. Cambridge Univ. Press.

Eubank, S., Guclu, H., Kumar, V., Marathe, M., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature(London)*, 429(6988):180–184.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213.

Fearnhead, P. (2011). *MCMC for State Space Models*, pages 513–529. Chapman & Hall/CRC Handbook of Modern Statistical Methods.

Fearnhead, P. and Meligkotsidou, L. (2004). Exact filtering for partially-oberved continuous-time models. *Journal of the Royal Statistical Society: Series B*, pages 771–789.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85:398–409.

Ghahramani, Z., Griffiths, T., and Sollich, P. (2007). Bayesian nonparametric latent feature models. In *Bayesian Statistics*. Oxford Univ. Press.

Gibson, G. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, 15(1):19–40.

Golightly, A. and Wilkinson, D. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820.

Griffiths, T. and Ghahramani, Z. (2011). The Indian buffet process: an introduction and review. *Journal of Machine Learning Research*, 12(April):1185–1224.

Griffiths, T. and Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process (tech. rep. no. 2005-001). Technical report, Gatsby Computational Neuroscience Unit, University College London.

Groendyke, C., Welch, D., and Hunter, D. (2011). Bayesian inference for contact networks given epidemic data. *Scandinavian Journal of Statistics*, 38(3):600–616.

Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods,*. Wiley.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hethcote, H. W. (1994). A thousand and one epidemic models. In *Frontiers in mathematical biology*, pages 504–515. Springer.

Higham, N. J. (2008). *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics.

Hohle, M., Jorgensen, E., and O'Neill, P. (2005). Inference in disease transmission experiments by using stochastic epidemic models. *Journal of the Royal Statistical Society: Series C*, 54(2):349–366.

Jarner, S. and Roberts, G. (2007). Convergence of heavy-tailed Monte Carlo Markov chain algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815.

Jewell, C., Keeling, M., and Roberts, G. (2008). Predicting undetected infections during the 2007 foot-and-mouth disease outbreak. *Journal of the Royal Society Interface*.

Jewell, C., Kypraios, T., Christley, R., and Roberts, G. (2009a). A novel approach to real-time risk prediction for emerging infectious diseases: A case study in avian influenza H5N1. *Preventive Veterinary Medicine*, 91(1):19–28.

Jewell, C. and Roberts, G. (2012). Enhancing Bayesian risk prediction for epidemics using contact tracing. *Biostatistics*, 13(4):567–579.

Jewell, C., Kypraios, T., Neal, P., and Roberts, G. (2009b). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4:465–496.

Jonkers, A., Sharkey, K., and Christley, R. (2010). Preventable H5N1 avian influenza epidemics in the British poultry industry network exhibit characteristic scales. *Journal of The Royal Society Interface*, 7(45):695–701.

Keeling, M. and Ross, J. (2008). On methods for studying stochastic disease dynamics. *Journal of The Royal Society Interface*, 5(19):171.

Kemeny, J. and Snell, J. (1976). *Finite Markov Chains.* Springer Verlag.

Kermack, W. and McKendrick, A. (1991a). Contributions to the mathematical theory of epidemics–i. *Bulletin of Mathematical Biology*, 53:33–55.

Kermack, W. and McKendrick, A. (1991b). Contributions to the mathematical theory of epidemics–ii. the problem of endemicity. *Bulletin of Mathematical Biology*, 53:57–87.

Kermack, W. and McKendrick, A. (1991c). Contributions to the mathematical theory of epidemics–iii. further studies of the problem of endemicity. *Bulletin of Mathematical Biology*, 53:89–118.

Kurtz, T. G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and their Applications*, 6(3):223–240.

Kurtz, T. G. (1981). *Approximation of Population Processes.* Society for Industrial and Applied Mathematics, .

Kypraios, T. (2009). A note on maximum likelihood estimation of the initial number of susceptibles in the general stochastic epidemic model. *Statistics & Probability Letters*, 79(18):1972–1976.

Latapy, M., Magnien, C., and Vecchio., N. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*, 30:31–48.

Maire, R., Reibman, A., and Trivedi, K. (1987). Transient analysis of acyclic Markov chains. *Performance Evaluation*, 7(3):175–194.

Masuda, N., Miwa, H., and Konno, N. (2005). Geographical threshold graphs with small-world and scale-free properties. *Physical Review E*, 71(3):36108.

Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30.

Medlock, J. and Galvani, A. (2009). Optimizing influenza vaccine distribution. *Science*, 325:1705–1708.

Moler, C. and Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1):3–49.

Mollison, D. (1995). *Epidemic models: their structure and relation to data.* Cambridge Univ. Press.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS.Med.*, 5.

Neal, P. and Roberts, G. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, 5(2):249–261.

Neal, P. and Roberts, G. (2005). A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing*, 15(4):315–327.

Newman, M. (2003). Properties of highly clustered networks. *Physical Review E*, 68(2):026121.

Norris, J. R. (1998). *Markov Chains.* Cambridge Univ. Press.

O'Neill, P. and Roberts, G. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A*, pages 121–129.

O'Neill, P., Balding, D., Becker, N., Eerola, M., and Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series C*, 49(4):517–542.

Papaspiliopoulos, O., Roberts, G., and Skold, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, 22:59–73.

Parzen, E. (1962). *Stochastic Processes.* Holden-Day.

Pellis, L., Ball, F., and Trapman, P. (2012). Reproduction numbers for epidemic models with households and other social structures. i. definition and calculation of R0. *Mathematical Biosciences*, 235(1):85–97.

Piessens, R., Doncker-Kapenga, D., Überhuber, C., Kahaner, D., et al. (1983). *QUADPACK, A subroutine package for automatic integration.* Springer.

Pitt, M. (2002). Smooth particle filters for likelihood evaluation and maximisation. *Economics Research Paper*, 651. URL `http://wrap.warwick.ac.uk/1536/1/WRAP_Pitt_twerp651.pdf`.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–284.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26:102–115.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods, 2nd ed.* Springer-Verlag.

Roberts, G. and Rosenthal, J. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367.

Roberts, G. and Smith, A. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49(2):207–216.

Roberts, G. and Tweedie, R. (1996). Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83(1):95–110.

Rosenthal, J. (2011). *Handbook of Markov chain Monte Carlo*, chapter 4, Optimal Proposal Distributions and Adaptive MCMC. Chapman & Hall.

Scalia-Tomba, G. (1985). Asymptotic final-size distribution for some chain-binomial processes. *Advances in Applied Probability*, 17(3):477–495.

Scott, S. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st Century. *Journal of the American Statistical Association*, 97(457):337–351.

Severo, N. C. (1969a). A recursion theorem on solving differential-difference equations and applications to some stochastic processes. *Journal of Applied Probability*, pages 673–681.

Severo, N. C. (1969b). Generalizations of some stochastic epidemic models. *Mathematical Biosciences*, 4(3-4):395–402.

Stewart, W. (1991). *Numerical solution of Markov chains*. CRC Press.

Teh, Y., Gorur, D., and Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 556–563.

van de Liefvoort, A. and Heindl, A. (2005). Approximating matrix-exponential distributions by global randomization. *Stochastic Models*, 21(2):669–693.

Whittle, P. (1955). The outcome of a stochastic epidemic - a note on Bailey's paper. *Biometrika*, 42(1-2):116–122.

Wilkinson, R. D. (2011). The pseudo-marginal approach to exact approximate MCMC algorithms. URL `http://darrenjw.wordpress.com/2010/09/20/the-pseudo-marginal-approach-to-exact-approximate-mcmc-algorithms/`.