Estimating Parameters for Stochastic Epidemics

Michael Höhle
Department of Animal Science and Animal Health, Royal Veterinary and Agricultural University,
Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark

Erik Jørgensen
Department of Agricultural Systems, Danish Institute of Agricultural Sciences,
Research Centre Foulum, P.O. Box 50, 8830 Tjele, Denmark

Dina Research Report No. 102 • November 2002

Estimating Parameters for Stochastic Epidemics

By Michael Höhle and Erik Jørgensen

Dina Research Report No. 102 • November 2002

Abstract:
Understanding the spread of infectious disease is an important issue in order to prevent major outbreaks. In this report mathematical modeling is used to gain insight into the dynamics of an epidemic. A process model, the SIR model, exploiting knowledge about population dynamics serves as framework. Key interest is in adapting the stochastic model to observed data – especially from animal production. Observing all events of an epidemic is not feasible in practice, hence estimation of model parameters has to be done from missing data. We give a rigorous treatment of an existing technique to handle estimation in partially observed epidemics using Markov Chain Monte Carlo (MCMC). The aim of this report is to extend the basic SIR model to handle two common situations in animal production: interaction into the course of the epidemic and population heterogeneity due to the spatial layout of confinement. Handling partially observed epidemics in these contexts we do by extending the above described MCMC method. A programming environment has been developed to exemplify the model extensions at a proof of concept level. It is made available for download for others to confirm our results or try the extensions on their own data.

Keywords: Infectious disease, SIR model, partial observability, Markov chain Monte Carlo, multi-type epidemic.

This report is also available on WWW @ URL:
http://www.dina.dk/∼hoehle/pubs/dina102.pdf

Danish Informatics Network in the Agriculture Sciences
The Royal Veterinary and Agricultural University
Thorvaldsensvej 40
1871 Frederiksberg C
Denmark

Contents

Chapter 1

Introduction

It does not take more than a couple of pages of Stephen King's "The Stand" before one realizes that malicious epidemics seem to be a favored way to scare the interested reader. Unfortunately, the literary universe of e.g. King is not far ahead of reality. Stories in the news range from potential terrorist threats by smallpox, HIV in Africa, to foot and mouth disease in England and spreading of e-mail viruses. The list indicates that epidemics are definitely not confined to humans – animals, computers, etc. are also targets. Understanding the spread of infectious disease is an important issue in order to prevent major outbreaks of an epidemic. Exploiting mathematics and especially statistics to gain this insight has a long tradition [Cliff and Haggett, 1993] and boils down to the design of one or more parametric models, which are investigated by fitting them to epidemics observed in the real world. Here one distinguishes between time-series models and process-based models. Where as the former can be described as a black-box approach doing well to detect patterns in the data, they lack explanatory power. Process based models, on the other hand, are white-box oriented; prior knowledge about population dynamics from clinical investigations or earlier studies, etc. is put into the model giving the parameters a physical interpretability. For a person to person transmitted disease, such prior knowledge could consist of a division of the population into compartments (being distinct stages of the disease) and a model for the transmission between the stages. Each transmission again consists of sub-components, e.g. to go from healthy to sick requires a healthy to have an infectious contact. Thus, assumptions are made on the contact behavior of the population and how the disease might be transmitted in case of an encounter. For example in a *homogeneous population*, two random individuals have an equal probability of meeting. For a population living in a large area this is not a good assumption, because they are more inclined to meet near neighbors as individuals far away.

The crucial step, when using a model to describe an infectious disease, is the adaptation to the specific disease by *estimation* of model parameters from data. In our case the data are the observations available for the infectious disease under study. Here, the advantage of interpretable parameters is that a fitted model can serve as a tool to gain both insight on the process and evaluate different control strategies.

This report deals with a commonly used process-model the SIR-model; a compartment model with three states: susceptible, infected, and recovered [Cliff and Haggett, 1993]. A susceptible individual becomes a disease transmitter by changing to the infectious state. When an infectious individual is cured or in other ways cannot contribute to the spread of the disease anymore (e.g. dies or is isolated), it is regarded as recovered. To keep it simple the population is often assumed to be homogeneous. SIR-modeling can be done using a variety of methods; using a deterministic or stochastic setting, regarding individuals as continuous or discrete quantities, and letting time scale be continuous or discrete, see [Diekmann and Heesterbeek, 2000; Andersson and Britton, 2000]. Common is a desire to establish a sufficient framework for understanding and investigating the epidemics. Fitting the model to observations from epidemics by estimating the model parameters depends on the data available. If the time and type of every transition between compartments is observed this is easy. Such ideal situations although never occur in practice, because here data are collected retrospectively and not as part of designed experiments. Onset of infection is especially hard to determine – often only secondary information, such as clinical symptoms or seroconversion can be detected. The connection in time between onset of infection and the secondary information is often highly stochastic. It is often easier to detect the recoveries, because they refer to the time point, where an individual becomes inactive with respect to disease-spreading. Detection of the disease by symptoms can be assumed equal to recovery, because bed-rest (humans), stamping out (herds), or other isolating and terminating actions are performed. The assumption that detection equals recovery is although not always reasonable; for an only moderately injurious disease, measurements range from doing nothing to applying some sort of medical treatment. At best, the latter speeds up recovery time, but none of the measures stop the detected infective immediately. Another problem is that a non-negligible amount of falsely detected cases of a disease might occur.

Our aim is to use the SIR-model as building block for decision support handling infectious diseases in animal production. Application of SIR-models within this domain is subject to a set of characteristics worth noting.

- Populations under study in this application can be small, e.g. the members of a pen or section in a slaughter pig production unit. Handling the population as a continuous quantity would be too coarse an approximation in this case.
- The smaller the unit the more distinct the influence of the chance element. Hence, stochasticity of the model is a must.
- Observations are likely to be infrequent compared to the speed of the epidemic. Hence, long chains of infections can occur between observation times, violating basic independence assumptions in the SIR-model. A continuous time model would be better suited for infrequent data. Furthermore, it allows for any-time predictions, which are of interest for a warning system.

Whereas the above argues for the choice of time scale, population granularity and stochasticity within the framework of single population SIR-model, two further characteristics require extensions to the simple model.

- Stalling of the animals into pens and sections introduces a spatial dimensions causing an inhomogeneity in the contact process not covered by the simple model. The hierarchy of pens, sections and population should thus be part of the model.
- Interaction into the epidemic process by e.g. authorities implementing control measures influences contact behavior and recovery process. Examples from the veterinarian field are the introduction of a vaccination policy or the decision to carry out pre-emptive slaughtering of farms neighboring an infected farm.

The aim of this report is to give a detailed description of parameter estimation in the simple SIR-model and integrate the necessary extensions induced by the animal production domain into the

SIR-model. That is handling spatial arrangements of the population causing an inhomogeneity in the contact process and interaction into the both contact and recovery processes by control measures. With a fitted SIR-model the way is paved for applications such as a disease warning system predicting the location of new cases in a pig production facility as described in [Höhle, 2002]. A secondary goal is to present a programming environment developed to exemplify the model extensions at a proof of concept level. Furthermore, this implementation allows others to try the extensions on their own data.

Chapter 2

Parameter estimation in simple SIR epidemics

We begin this chapter by introducing the concept of the continuous time SIR-model, the class of SIR model best suited to meet the demands mentioned in Chapter 1. The model will constitute our framework for describing infectious diseases and estimate its parameters to fit specific observations on recovery and infection times. At first, well-known formulas on how to obtain maximum likelihood estimates for β and γ in case of full observations are derived. Hereafter, an approach for handling estimation in case of missing infection times is presented: Bayesian analysis using Markov Chain Monte Carlo (MCMC). Goal is to obtain a sufficient understanding for creating an own implementation of the various SIR-estimation procedures. The Bayesian part of this chapter is based on work of [O'Neill and Roberts, 1999], but extended in an applied manner by convergence testing, discussion of implementational issues, and method validation through simulated datasets.

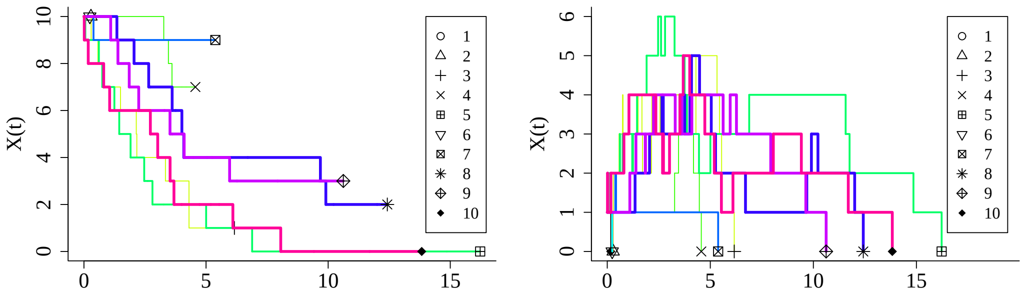
2.1 Introducing the stochastic continuous time SIR-model

The following establishes the notational framework using the terms of [O'Neill and Roberts, 1999]. A population of initially N susceptible and a infectious is considered. Note that in general a can be arbitrarily large, but to keep things simple only $a = 1$ is considered. $X(t)$ and $Y(t)$ denote respectively the number of susceptible and the infected individuals at time $t \geq 0$. It is sufficient to keep track of $(X(t), Y(t))$, because for all t the equality $X(t)+Y(t)+R(t) = N+a$ must hold, with $R(t)$ denoting the number of recovered individuals. Evolution of the epidemic process $(X(t), Y(t))$ can be described by a continuous time Markov chain with the following transition rates.

$$\begin{aligned}(i,j) \to (i-1, j+1) \quad &: \quad \beta X(t)Y(t) \\ (i,j) \to (i, j-1) \quad &: \quad \gamma Y(t)\end{aligned}$$

Beginning with the single initially infected individual, the epidemic runs until there are no more infected left in the population. In case the initial infective recovers before any other individuals are infected the epidemic ceases. We will denote such an epidemic a *singleton*.

EXAMPLE 2.1 (A SIMPLE EPIDEMIC)
Consider a SIR epidemic with parameters $\beta = 0.1$, and $\gamma = 0.3$ in a population with $(N, a) = (10, 1)$, i.e. 10 susceptibles and one initial infected. Figure 2.1 shows 10 realizations of this model underlining the stochasticity of the model. In an estimation context, usually only one realization of the epidemic is available – from which we then try to estimate the parameters.



(a) $X(t)$ (b) $Y(t)$

Figure 2.1: $X(t)$ and $Y(t)$ for 10 simulations of the example epidemic. In the runs 1,2, and 6, the initial infectious recovers before any new infections occur, thus stopping the epidemic.

◇

One way to exploit SIR-models is to use them as a tool to simulate the outcome of epidemics to get an understanding of how the model parameters influence the outcome. For example, Figure 2.1 illustrates the stochasticity in Example 2.1. Another way is to use the model to explain observed epidemic data by finding the model parameters β and γ giving the best fit to the available data. Depending on the specific case, this could be both recovery infection and recovery times, just recovery times or in some situation just the final total number of susceptibles infected at the end of the epidemic, i.e. the *final size*, Z, of the epidemic. Sections 2.2 and 2.3 treat the first two scenarios in detail. In case only final size data is available, no individual estimate of β and γ can be made – only an estimate of the *basic reproduction ratio*, i.e.

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

can be obtained. This number is often of interest, because major outbreaks can occur only if $R_0 > 1$, see [Andersson and Britton, 2000, Chapter 4]. To illustrate the final size, Figure 2.2 shows the final size distribution of the epidemic in Example 2.1. Further information on how to estimate

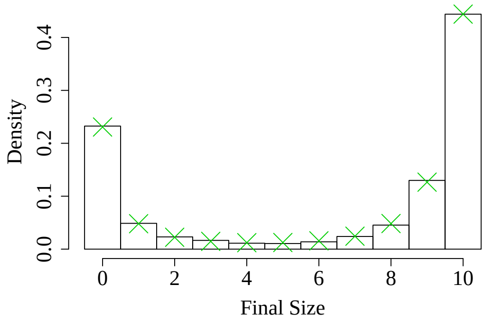


Figure 2.2: Final size distribution of the epidemic in Example 2.1. The bar-plot shows the result of 10,000 simulations, whereas crosses show exact value obtained by solving the equation system in [Andersson and Britton, 2000, Theorem 2.2].

this basic reproduction ratio from final size data can be found in [Becker, 1989; Andersson and Britton, 2000] but is otherwise not treated in this report.

To describe event data, let $I_0 < 0$ be the time of the first infection, i.e. the initial infection making $a = 1$. Assume that the epidemic is observed in the time interval $[I_0, T]$, if at time T all infected are recovered, the epidemic is said to be observed to its end. Data thus consists of observed removal times $\boldsymbol{\tau}$, and subsequent infectious times $\boldsymbol{I}$ observed during $[I_0, T]$. Here, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)$, is in time wise order with the first recovery defined to be observed at time zero. That is $\tau_1 \equiv 0$ and $\tau_i \leq \tau_{i+1}$. Similarly, $\boldsymbol{I} = (I_1, \dots, I_{m-1})$, with $I_j \leq I_{j+1}$. In general $n \leq m \leq N$, with $m = n$ if the epidemic is observed for its entire duration. To make the analysis simpler we decided to only consider epidemics observed until end, i.e. in our cases we always have $m = n$. This is though not a general requirement before parameter estimation is possible [O'Neill and Roberts, 1999].

As the duration of being infectious is stochastic, I_i and τ_i do not necessarily have to refer to the same individual. On the other hand we know that once the total number of infected individuals drops to zero no more infections can occur and the epidemic stops. This knowledge combined with the observed data can be translated to a relationship between the I's and the τ's. Let $I(t) = |\{I \in (\{I_0\} \cup \boldsymbol{I}) \,|\, I \leq t\}|$ and $R(t) = |\{\tau \in \boldsymbol{\tau} \,|\, \tau \leq t\}|$ be the number of infective and recovered at time t. For each observed recovery τ_i, the correspondent infection has to occur prior to τ_i, i.e. for all $t \in [I_0, \tau_n[$ we should have $Y(t) = I(t) - R(t) > 0$. In case of $m = n$ this is equivalent to the constraint

$$I_i < \tau_i \text{ for all } 1 \leq i \leq m-1. \tag{2.1}$$

2.2 ML-estimation in case of full data

As a first step, the likelihood expression for a fully observed epidemic with data $(\boldsymbol{\tau}, I_0, \boldsymbol{I})$ is computed. Instead of absolute time the waiting time between the different events is modeled, i.e. $t_i = t'_i - t'_{i-1}$, where t'_i denotes events in absolute time. Likelihood of each waiting time can be found using survival analysis methodology in a setup with multiple modes of failure, see [Fahrmeir and Tutz, 1994, p.331]. Two events of "failure" are possible: becoming infected or recovering. The two hazard functions can be written as

$$\begin{aligned}\lambda_{\text{inf}}(t_i|\beta) &= \beta X(t_i) Y(t_i),\\ \lambda_{\text{rec}}(t_i|\gamma) &= \gamma Y(t_i).\end{aligned}$$

with $X(t)$ and $Y(t)$ being the number of susceptibles and infectious at time t, respectively. The overall hazard function is thus given by

$$\lambda(t_i|\beta,\gamma) = \beta X(t_i) Y(t_i) + \gamma Y(t_i).$$

Between events the hazard functions are constant, i.e. the advantage of using waiting times is that it allows operating with constant hazards.

The epidemic dataset is described by $\boldsymbol{D} = \{(t_i, e_i)\}$, where $e_i \in \{\text{inf}, \text{rec}\}$ denotes the event-type. $\boldsymbol{D}$ consists of n recovery events and m infectious events, with $m = n$ when the epidemic is observed to end. Interest is now in the density of $\boldsymbol{\tau}, \boldsymbol{I}$ given β, γ, I_0. Consider therefore the likelihood of a single arbitrary event e_i in $\boldsymbol{D}$ as failure time in a setup with constant hazard and non-informative censoring. If time is zero at the occurrence of the event just prior to e_i, the likelihood is

$$L_i = \lambda_{e_i}(t_i|\beta,\gamma)P(T_i \geq t_i|\beta,\gamma) = \lambda_{e_i}(t_i|\beta,\gamma)S(t_i|\beta,\gamma),$$

where T_i is a stochastic variable denoting failure time of e_i and

$$S(t|\beta,\gamma) = \exp\left(-\int_0^t \lambda(u|\beta,\gamma)du\right)$$

is the survival function based on the overall hazard. Assuming independence between the arrival times, the overall likelihood is given by looking at all infection and all recovery events except I_0. The latter is not part of the likelihood, because an epidemic always is conditioned on the existence of the first infective. Hence, L

$$\begin{aligned}
&= \left[\prod_{i=1}^{n} \gamma Y(\tau_i^-)\exp\left(-\int_{\tau_{i-1}}^{\tau_i}\lambda(t|\beta,\gamma)\,\mathrm{dt}\right)\right]\left[\prod_{i=1}^{m-1}\beta X(I_i^-)Y(I_i^-)\exp\left(-\int_{I_{i-1}}^{I_i}\lambda(t|\beta,\gamma)\,\mathrm{dt}\right)\right]\\
&= \prod_{i=1}^{n}\gamma Y(\tau_i^-)\prod_{i=1}^{m-1}\beta X(I_i^-)Y(I_i^-)\left[\prod_{i=1}^{n}\exp\left(-\int_{\tau_{i-1}}^{\tau_i}\lambda(t|\beta,\gamma)\,\mathrm{dt}\right)\right]\left[\prod_{i=1}^{m-1}\exp\left(-\int_{I_{i-1}}^{I_i}\lambda(t|\beta,\gamma)\,\mathrm{dt}\right)\right]\\
&= \prod_{i=1}^{n}\gamma Y(\tau_i^-)\prod_{i=1}^{m-1}\beta X(I_i^-)Y(I_i^-)\left[\exp\left(-\sum_{i=1}^{n}\left(\int_{\tau_{i-1}}^{\tau_i}\lambda(t|\beta,\gamma)\,\mathrm{dt}\right)-\sum_{i=1}^{m-1}\left(\int_{I_{i-1}}^{I_i}\lambda(t|\beta,\gamma)\,\mathrm{dt}\right)\right)\right],
\end{aligned}$$

where $X(t^-), Y(t^-)$ denotes the situation just prior to time t, i.e. $Y(t^-) = \lim_{t\to t^-} Y(t)$. As the i and r events are consecutive in time the sums of integrals can be simplified into a single integral. Hence,

$$f(\boldsymbol{\tau},\boldsymbol{I}|\beta,\gamma,I_0) = \prod_{i=1}^{n}\gamma Y(\tau_i^-)\prod_{i=1}^{m-1}\beta X(I_i^-)Y(I_i^-)\exp\left(-\int_{I_0}^{T}\beta X(t)Y(t)+\gamma Y(t)\,\mathrm{dt}\right) \quad (2.2)$$

Differentiating this likelihood with respect to either β and γ and solving for zero reveals that the ML estimates are given as

$$\hat{\beta} = \frac{m-1}{\int_{I_0}^{T} X(t)Y(t)\,\mathrm{dt}}, \qquad \hat{\gamma} = \frac{n}{\int_{I_0}^{T} Y(t)\,\mathrm{dt}}, \quad (2.3)$$

which corresponds to the results from [Andersson and Britton, 2000, Section 9.2]. Here, using counting processes the ML-estimators are also proven to be asymptotically Gaussian with the true parameter values as mean and consistent estimates of the standard errors being

$$\text{s.e.}(\hat{\beta}) = \frac{\hat{\beta}}{\sqrt{m-1}}, \qquad \text{s.e.}(\hat{\gamma}^{-1}) = \frac{\hat{\gamma}^{-1}}{\sqrt{n}}.$$

From (2.3) it is also easy to see that if all observations, i.e. $I_0, \boldsymbol{I}, \tau$, are scaled in time by a factor k, the resulting ML estimates for β and γ are just k times the value of the non-scaled estimators. This might come handy to avoid numeric overflow when working with epidemics having large t values.

2.3 MCMC-estimation in case of partial observability

As mentioned in the introduction, infectious times are hardly ever observed in practice, but we are still interested in estimation β, γ, or just R_0 despite of missing $\{I_0\} \cup \boldsymbol{I}$ observations. Based on the observed removal times Markov Chain Monte Carlo can be used to deal with the partial observed epidemic process. Our goal is to obtain estimates of the β and γ process parameters. Due to the non-observed infection times our state vector will thus be $(\beta, \gamma, I_0, \boldsymbol{I})$. This section follows the work described in [O'Neill and Roberts, 1999].

Operating within the Bayesian framework prior distributions are assumed for the unknown parameters β, γ, I_0.

$$\beta \sim \Gamma(\nu_\beta, \lambda_\beta), \quad \gamma \sim \Gamma(\nu_\gamma, \lambda_\gamma) \quad -I_0 \sim \mathrm{Exp}(\theta),$$

where $\Gamma(\nu, \lambda)$ denotes the Gamma distribution with mean ν/λ and variance ν/λ^2, and $\mathrm{Exp}(\theta)$ denotes the exponential distribution with mean $1/\theta$. Prior for $\boldsymbol{I}$ is just assumed to be uniform on the set of valid configurations. Using Equation 2.2 the full-conditional posterior distributions can be obtained for β, γ, and I_0, which allows us to use Gibbs updating of these quantities. For $\boldsymbol{I}$ the full-conditionals cannot be as easily established, therefore a Metropolis sampler will be used here. Calculation of the full-conditional posteriors happens according to Bayes Formula,

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

As interest is only up to proportion all terms not containing the variables of interested are ignored, i.e. for β we have

$$\begin{aligned}
\pi(\beta|\boldsymbol{\tau}, I_0, \boldsymbol{I}, \gamma) &\propto f(\boldsymbol{\tau}, \boldsymbol{I}|\beta, \gamma, I_0)\pi(\beta) \\
&\propto \left(\prod_{i=1}^{m-1} \beta X(I_i^-)Y(I_i^-)\right) \exp\left(-\int_{I_0}^{T} \beta X(t)Y(t) + \gamma Y(t)\,\mathrm{d}t\right) \beta^{\nu_\beta - 1} \exp(-\lambda_\beta \beta) \\
&\propto \beta^{m-1} \exp\left(-\int_{I_0}^{T} \beta X(t)Y(t)\,\mathrm{d}t\right) \beta^{\nu_\beta - 1} \exp(-\lambda_\beta \beta) \\
&= \beta^{\nu_\beta + m - 2} \exp\left(-\lambda_\beta \beta - \beta \int_{I_0}^{T} X(t)Y(t)\,\mathrm{d}t\right) \qquad (2.4) \\
&\sim \Gamma\left(\nu_\beta + m - 1, \lambda_\beta + \int_{I_0}^{T} X(t)Y(t)\,\mathrm{d}t\right).
\end{aligned}$$

In case of non-informative priors, $\nu_\beta = \lambda_\beta = 0$, the mean of the posterior will be identical to the ML estimates in Equation (2.3). If the epidemic of interest consists of a singleton, a setting with non-informative priors makes the first parameter of the Γ-distribution zero. This leads to an incorrectly specified distribution causing problems in case we want to sample from it. If instead a proper prior is specified, sampling occurs directly from this prior, because the data of a singleton epidemic provide no additional information about β.

Using similar computations, the posterior of γ is found to be

$$\pi(\gamma|\boldsymbol{\tau}, I_0, \boldsymbol{I}, \beta) \sim \Gamma\left(\nu_\gamma + n, \lambda_\gamma + \int_{I_0}^{T} Y(t)\,\mathrm{d}t\right).$$

To compute the posterior distribution of I_0, we use that the prior for $-I_0$ is $\mathrm{Exp}(\theta)$ distributed.

Letting the zero indicator function of the exponential distribution be implicit, we obtain

$$\begin{aligned}
\pi(-I_0|\boldsymbol{\tau},\boldsymbol{I},\beta,\gamma) &\propto f(\boldsymbol{\tau},\boldsymbol{I}|\beta,\gamma,I_0)\pi(-I_0|\beta,\gamma,I_0) \\
&= \prod_{i=1}^{n}\gamma Y(\tau_i^-)\prod_{i=1}^{m-1}\beta X(I_i^-)Y(I_i^-)\exp\left(-\int_{I_0}^{T}\lambda(t|\beta,\gamma)\,\mathrm{dt}\right)\theta\exp(-\theta(-I_0)) \\
&\propto \exp\left(-\int_{I_0}^{I_1}\beta X(t)Y(t)+\gamma Y(t)\,\mathrm{dt}\right)\exp(-\theta(-I_0)).
\end{aligned}$$

As $X(t) = N$ for $I_0 \leq t < I_1$ and $Y(t) = 1$ for $I_0 \leq t < I_1$ the necessary integrals are,

$$\int_{I_0}^{I_1} Y(t)\,\mathrm{dt} = (I_1 - I_0), \quad \text{and} \int_{I_0}^{I_1} X(t)Y(t)\,\mathrm{dt} = N(I_1 - I_0).$$

yielding $\pi(-I_0|\boldsymbol{\tau},\boldsymbol{I},\beta,\gamma) \propto \exp(-\beta N(I_1 - I_0) - \gamma(I_1 - I_0) - \theta(-I_0))$. As calculation is only up to proportionality, we can multiply by appropriate constants to obtain a proper pdf., i.e.

$$\pi(-I_0|\boldsymbol{\tau},\boldsymbol{I},\beta,\gamma) \sim \mathrm{Exp}(\beta N + \gamma + \theta), \tag{2.5}$$

which produces the desired posterior distribution of I_0.

The initial state vector

Initial values for β and γ are obtained by sampling from their respective priors. By fixing the seed value of the random generator it is possible to assure fixed initial values useful for debugging and comparison. Similarly, I_0 is obtained by sampling a value from $\mathrm{Exp}(\theta)$ and negating it. To find a $\boldsymbol{I}$ taken uniform from the set of valid configurations, we draw $m-1$ independent values from the uniform distribution on (I_0, τ_{n-1}), sort them in ascending order, and check whether they obey the constraint of $I_i \leq \tau_i$ for $1 \leq i \leq m-1$. This procedure is repeated until a valid sample is obtained.

Generating new states by Gibbs-within-Metropolis

A Gibbs within Metropolis sampling scheme is used to update the state vector $(\beta,\gamma,I_0,\boldsymbol{I})$. The first three components are updated by sampling from the respective full conditional distribution, whereas sampling from $\pi(\boldsymbol{I}|\boldsymbol{\tau},I_0,\beta,\gamma)$ is done using a Metropolis sampler. For notational convenience $\pi(\boldsymbol{\tau},\boldsymbol{I}|I_0,\beta,\gamma)$ is abbreviated as $f(\boldsymbol{I})$. Compared to the estimation scheme of O'Neill and Roberts [1999] we know know the size of $\boldsymbol{I}$, because the epidemic is assumed observed to end. Thus, there is no need to estimate m by being able to add or remove infection times to the state vector. The only operation is therefore the moving of an infection event in time: Choose at random one of the infection times $I_1,\ldots,I_{m-1}$ and denote this time s. Generate a replacement time t by sampling uniform on (I_0, T). This new infection candidate is accepted with probability

$$\alpha(X,Y) = \min\left(1, \frac{f((\boldsymbol{I}\backslash\{s\})\cup\{t\})}{f(\boldsymbol{I})}\right).$$

Proposal distribution of the new element in the state $\boldsymbol{I}' = (\boldsymbol{I}\backslash\{s\})\cup\{t\}$ is thus $U(I_0,T)$, which is independent of current state $\boldsymbol{I}$, corresponding to the independence sampler. Independence also automatically ensures that the proposal distribution is symmetric which is required for a Metropolis Algorithm, see [Gilks *et al.*, 1996]. Also, for MCMC to work it has to be ensured that the designed Markov chain is irreducible and recurrent, i.e. it must be able to visit relevant states in the state space and that infinitely often [Gilks *et al.*, 1996, Chapter 4]. Demonstrating these properties is not always non-trivial and O'Neill and Roberts [1999] do not discuss the matter in further detail. For a more rigorous proof consult the work of Gibson and Renshaw [1998], who prove the above properties for an almost similar model.

Implementationalwise the $\boldsymbol{I}$ vector is represented as an array. After t is written to s's position in $\boldsymbol{I}$, the array needs to be sorted before the acceptance probability can be computed. Let $t = I_j$ after sorting, if $t < \tau_j$ the configuration is invalid and will automatically be rejected because $f((\boldsymbol{I}\backslash\{s\}) \cup \{t\}) = 0$. The reason here fore, is that the data claims that new infectious case are generated, even though $Y(t) = 0$. It remains future research to investigate whether a better proposal distribution scheme can be devised, than the above proposed by [O'Neill and Roberts, 1999]. For example, Figure 2.3 shows the acceptance probability of replacing I_1 or I_9 in Epidemic no. 3 from Example 2.1. Note how the uniform proposal is fair in the case of I_9, but quite poor in case of I_1. The question is whether some adaptive method can yield less rejection, fulfill the criteria of an Metropolis proposal distribution, while having a higher mixing rate. A possible alternative could be to use a random walk sampler instead of the independence sampler. Updating again occurs by at random selecting an existing infection time I_i using it to generate a new value

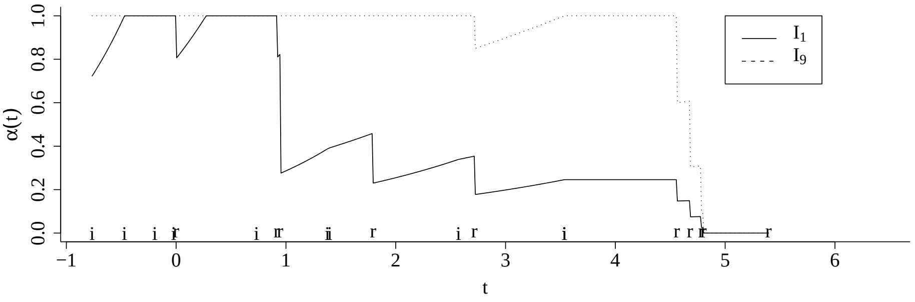


Figure 2.3: The acceptance probability α for replacement of I_1 and I_9 in Epidemic no. 3 from Example 2.1 with a $t \in (I_0 = -0.77, T = 5.40]$. The i and r letters on the x-axis denote current location of $I_0, \boldsymbol{I}$, and $\boldsymbol{\tau}$ events. Because a recovery means fewer infected and thus a lower probability for becoming infected, the acceptance probability drops at each recovery event.

$$I_i' = I_i + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

Tuning σ means finding a trade-off between a high acceptance rate and a low mixing, here adaptive methods, etc. from [Gilks *et al.*, 1996] might lead to inspiration.

Because only one component of $\boldsymbol{I}$ is updated per run, quite a large number of iterations are needed before stationarity can be expected. A possible workaround could be to update a larger number of components. This gives quicker mixing, but at the cost of a lower acceptance rate. Finally, some efficiency gains might be obtained by re-parameterizing the $\boldsymbol{I}$ parameter to contain waiting time between infection times instead. This would make time-consuming sorting superfluous.

2.4 Testing the implementation

A program, `LadyBug`, implementing the theory of the preceding sections has been developed. Providing information about population size, $\boldsymbol{I}$ and $\boldsymbol{\tau}$, the program computes the desired parameter estimates; Appendix A contains a thorough description of the program together with information on invocation, etc. In this section we will apply it to three examples: a small simulated epidemic containing five cases from [O'Neill and Roberts, 1999], data from a smallpox epidemic in Nigeria, and finally the simulated datasets generated by us in Figure 2.1. Aim is to check the correctness of the implementation by validating against the results by [O'Neill and Roberts, 1999] and datasets, where we know the correct value of the parameters. Such a validation is important: The iterative and stochastic nature of the methods makes introducing programming errors easy – locating and debugging them a matter of balancing between optimism and perplexity!

2.4.1 The five case SIR epidemic

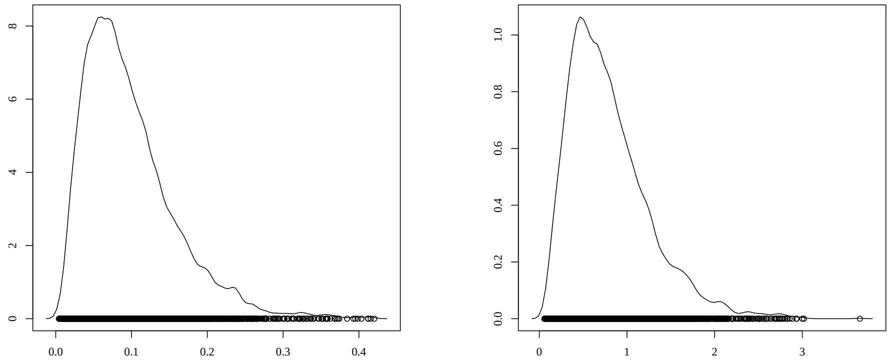
First test-case contains simulated data from a SIR-model with parameters $\beta = 0.12$ and $\gamma = 1$ in a population with $(N, a) = (9, 1)$. Data consists of five observed removal times[1], where it is assumed that the epidemic has been observed for its entire length, see O'Neill and Roberts [1999] for details. Prior distribution parameters are $(\nu_\beta, \lambda_\beta) = (0.1, 1)$ and $(\nu_\gamma, \lambda_\gamma) = (0.1, 0.1)$. After a burn-in of 1000 samples 5,000 samples are generated by running the chain for an additional 1,000,000 samples and thinning it by taking each 200th value. A thinning scheme is applied, in order to reduce the auto-correlation of the samples, this feels necessary, because only one infection time besides I_0 is changed per run. Despite of thinning, a larger number of generated samples also provides better convergence results in e.g. Geweke plots. Output of the program is now analyzed by the BOA package in R, see B. [2001]. The overall statistics for the two posterior distributions are as follows.

	Mean	SD	Naive SE	MC Error	Batch SE	Batch ACF
β	0.0978	0.0608	0.0008600	0.0008612	0.0008556	-0.0570
γ	0.7995	0.4667	0.0065997	0.0063301	0.0068253	-0.1854

Here, SD corresponds to the standard error, σ, of the posterior distribution, and Naive SE error is a measure of the precision of the sample mean as a point estimate of the true posterior mean. It is naively derived by assuming independence between samples, i.e. $\sigma/\sqrt{5000}$. The next two columns also estimate the above mentioned precision, but take the autocorrelation of the samples into account – either using spectral estimators or dividing calculations into batches, see [BUG, 2001; B. , 2001] for details. After the 101,000 samples the acceptance rate of the $\boldsymbol{I}$ component of the Metropolis sampler is 42.75%. In [O'Neill and Roberts, 1999] posteriors means of β and γ are 0.098 and 0.780 are reported with std. error 0.0632 and 0.451, respectively. Our estimates are thus in agreement with their findings. Furthermore, Figure 2.4 shows density plots of the two parameters obtained by a kernel-filter.

To get an idea about chain convergence we consider in Figure 2.5 the Geweke convergence diagnostic plot [BUG, 2001; B. , 2001] and in Figure 2.6(a) a trace of the β values. Out of the plotted 50 points in the figure approximately 1-2 are lying outside the confidence bands for the β parameter. The question is whether this indicates that the chain has converged? Heidelberger and

[1] Note that all data are available as `LadyBug` specification files from the program homepage.

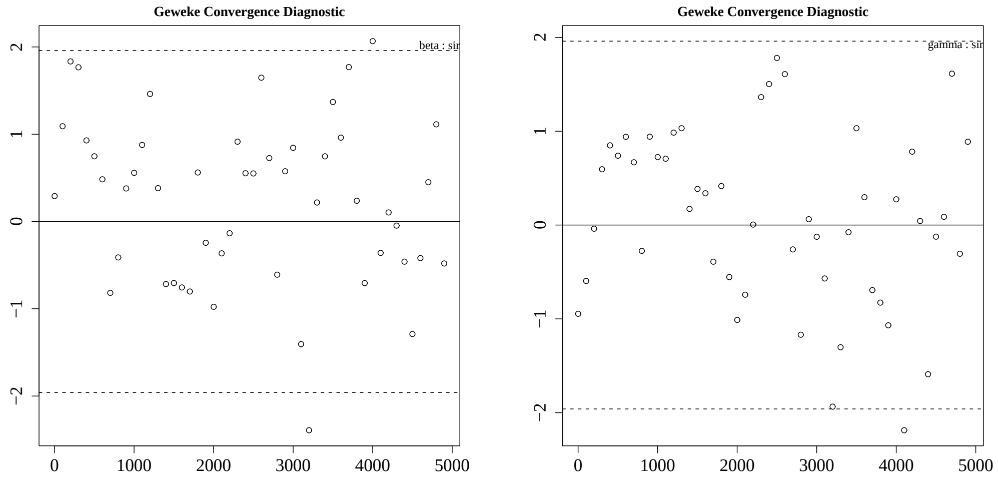


(a) β density

(b) γ density.

Figure 2.4: Kernel densities for β and γ, respectively.

Welch's test for stationarity (see [BUG, 2001]) is passed for both parameters, without having to throw any samples away. From the results of the above diagnostics, we play the daredevils and conclude that the chain appears to have converged.

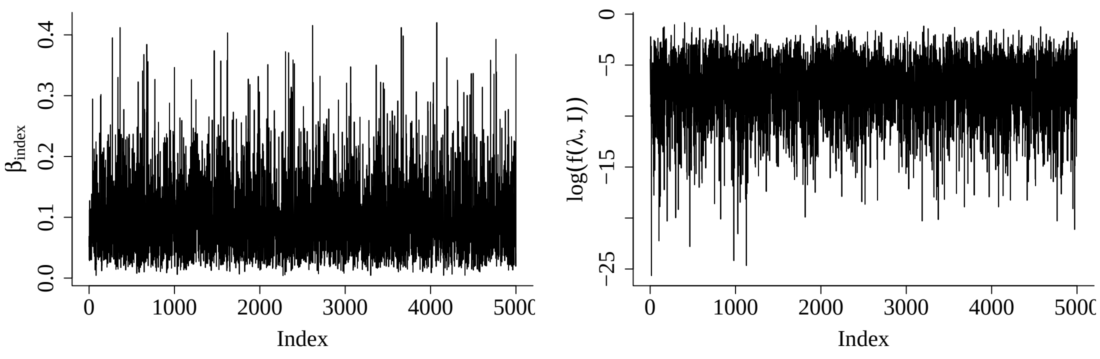


(a) β parameter

(b) γ parameter

Figure 2.5: The Geweke convergence diagnostic for β, γ, i.e Z-score together with 95% confidence intervals for the two parameters, see [BUG, 2001].

As an additional diagnostic tool Figure 2.6(b) shows the development of $f(\boldsymbol{\tau}, \boldsymbol{I}| \ldots)$ as a function of the sample index. If we instead of a Bayesian analysis think of MCMC as maximizing the f function, the trace in the Figure could be regarded as the output of stochastic optimizer of which the best result could be taken. If a brute-force maximization of Expression (2.2) with $(\beta, \gamma, I_0, \boldsymbol{I})$ as unknowns is desired, more sophisticated methods such as conjugate gradient or stochastic simulated annealing should be applied, see [Press *et al.*, 1992].



(a) (b)

Figure 2.6: (a) The β trace of the 5,000 samples. (b) Trace of $\log(f(\boldsymbol{\tau}, \boldsymbol{I}| \ldots))$ obtained by the same samples.

2.4.2 Sketching a Marginalized Likelihood method

As the small five-case epidemic constitutes our first application, a consideration about model-identifiability and alternative estimation methods are in place. We note, that the total number of model parameters is $|\boldsymbol{I}|+3$, which we are trying to estimate from $|\boldsymbol{\tau}|$ (in our case equal to $|\boldsymbol{I}|+1$) recovery times and knowledge about the start population (N, a). Having more model parameters than data is certainly problematic, but the constraint on $I_0 \cup \boldsymbol{I}$ given by $I_{i-1} \leq I_i \leq \tau_i$ restricts the parameter space substantially. To visualize this, consider the following intuitive procedure.

1. Select a valid configuration of $I_0 \cup \boldsymbol{I}$ with respect to Equation (2.1) by drawing it uniformly from the space of valid of configurations.
2. Estimate β, γ using ML-Estimation from $\boldsymbol{\tau}$ and the sample of $I_0 \cup \boldsymbol{I}$.

Repeating the procedure k times, yields k independent estimates for β, γ of which we can take the mean to obtain a qualified guess on the desired estimates.

Pondering, one realizes that the above procedure reminds about Monte Carlo optimization of an augmented likelihood function. Augmenting the likelihood $L(\beta, \gamma|\boldsymbol{\tau})$ with fictions $\{I_0\} \cup \boldsymbol{I}$ makes estimation easy. By integrating this augmented likelihood over all possible configurations of $\{I_0\} \cup \boldsymbol{I}$ given $\boldsymbol{\tau}$ the desired likelihood is obtained. In other words; the ML-estimate of β and γ based on the observable data $\boldsymbol{\tau}$ is found as the maximizer of the marginal density

$$L(\beta, \gamma|\boldsymbol{\tau}) = f(\boldsymbol{\tau}|\beta, \gamma) = \int \int f(\boldsymbol{\tau}, \boldsymbol{I}|I_0, \beta, \gamma) \, \mathrm{d}\boldsymbol{I} \, \mathrm{d}I_0,$$

where integration happens over all valid configurations $\{I_0\} \cup \boldsymbol{I}|\boldsymbol{\tau}$. Calculating the above high dimensional integral is certainly infeasible analytical or by ordinary numerical solutions. Instead, the objective function can be expressed as calculation of the expectation

$$L(\beta, \gamma|\boldsymbol{\tau}) = \int f(\boldsymbol{\tau}, \boldsymbol{I}|I_0 \boldsymbol{\tau}|\beta, \gamma) \; dP_{\tau}(\boldsymbol{I}, I_0),$$

where dP_{τ} is the distribution of $\{I_0\} \cup \boldsymbol{I}$ given $\boldsymbol{\tau}$, i.e. uniform over the set of valid configuration. This allows approximate the marginalized likelihood using Monte Carlo integration. Based on a sample of size k from dP_τ, denoted $\{I_0^i\} \cup \boldsymbol{I}^i, 1 \le i \le k$, calculate

$$L_k(\beta, \gamma|\boldsymbol{\tau}) = \frac{1}{k}\sum_{i=1}^{k} f(\boldsymbol{\tau}, \boldsymbol{I}^{(i)}|I_0^{(i)}, \beta, \gamma),$$

and let $(\hat{\beta}, \hat{\gamma})$ and $(\hat{\beta}_k, \hat{\gamma}_k)$ be the arguments maximizing L and L_k respectively. The obtained Monte Carlo estimate $(\hat{\beta}_k, \hat{\gamma}_k)$ has the property that $(\hat{\beta}_k, \hat{\gamma}_k) \rightarrow (\hat{\beta}, \hat{\gamma})$ for $k \rightarrow \infty$.

The difference between the intuitive method and Monte Carlo optimization is that we maximize β and γ for each configuration, instead of maximizing β and γ once by minimizing the likelihood sum of all configurations.

Crux of the above procedures is the sampling from P_τ, i.e. in our case to draw uniform from the space of valid $I_0 \cup \boldsymbol{I}|\boldsymbol{\tau}$ configurations. Problem is that I_0 is not bound to the left, making the space of possible configurations enormous, most of them with ridiculous low likelihood values. Figure 2.7 investigates the dependence of $f(\boldsymbol{\tau}, \boldsymbol{I}|I_0, \beta, \gamma)$ on I_0. The values are obtained by conditioning I_0 to be in $\{-1 - 5i \mid 0 \le i \le 20$ and then sample 100 configurations of $\boldsymbol{I}|I_0, \boldsymbol{\tau}$ for the five-case epidemic. The box-and-whisker plots illustrates the distribution of the log-likelihood for these 100 cases. The linear effect on the log-scale is clear from Equation (2.2).

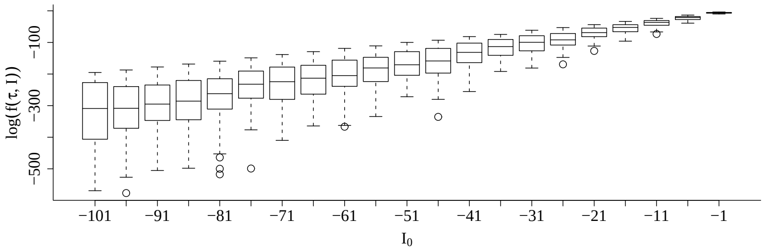
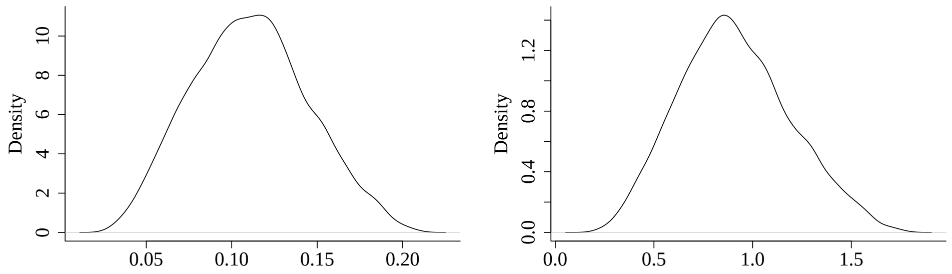


Figure 2.7: Box-and-whisker plot illustrating $\log(f(\boldsymbol{\tau}, \boldsymbol{I}|I_0, \beta, \gamma))$ as a function of I_0 for a set of fixed values. For each box, the median of the 100 values is shown, while box hinges show the borders of the 1st and 3rd quartile of the values, and whiskers extend to the most extreme data-point no more than 1.5 times away from these quartiles borders.

Figure 2.7 can be used to motivate that certain configurations are not worth exploring. To make estimation feasible but unfortunately only approximate, we replace the simultaneous sampling of I_0 and $\boldsymbol{I}$ by a stepwise procedure. Draw m values from $U(\theta, \tau_n)$, where θ is chosen such that an appropriate truncation of the interval $(-\infty, \tau_n)$ is obtained. By monitoring how $f(\boldsymbol{\tau}, \boldsymbol{I}|I_0, \beta, \gamma)$ depends on $\boldsymbol{I_0}$ it is often possible to get a good idea on how to choose θ in order to get a fair truncation. Sort the obtained values and check whether the constraints are satisfied with respect to $\boldsymbol{\tau}$. If not, repeat the procedure until a valid sample sample is obtained. Precision of the estimates depends very much of the shape of f, number of samples used and the fairness of the truncation. It remains a task of future research to replace the above coarse approximation with a more sound procedure.

Figure 2.8 shows the kernel smoothed density obtained from $10,000$ samples for the five-case epidemic, where we maximize for each configuration. The β samples have mean 0.1109 and variance of 0.0012, while the γ samples have mean 0.9093 and variance 0.0800. Running 100 runs of the Monte Carlo Optimization method each based on a sample size of $k = 1,000$ yields the following



(a) β density

(b) γ density.

Figure 2.8: Kernel densities for β and γ, obtained by from 10,000 independent and valid samples for $I_0 \cup \boldsymbol{I}$.

mean and sample variances: $\hat{\beta} = (0.0991, 1.4025e - 6)$ and $\hat{\gamma} = (0.8129, 9.2571 \cdot 10^{-5})$. Note that this uncertainty is just a measure of the effect of sampling error on the point estimate. The results from the two approaches illustrate the effect of the constraint at least for small epidemics: Missing infection times are handled well, because the constraint allows to deduce quite narrow intervals for the missing values.

2.4.3 Smallpox data

The next example concerns the classic smallpox epidemic example occuring in a closed community of 120 individuals in Abakaliki, Nigeria mentioned in both [Becker, 1989; O'Neill and Roberts, 1999]. With none-informative priors, the following results are obtained using a setup with 5,000 samples obtained by a burn-in of 1000 and a thin of 200.

	Mean	SD	Naive.SE	MC.Error	Batch.SE	Batch.ACF
β	0.00093	0.00030	9.365 $\cdot 10^{-6}$	9.288$\cdot 10^{-6}$	1.133 $\cdot 10^{-5}$	0.2940
γ	0.09765	0.03015	9.533 $\cdot 10^{-4}$	9.806$\cdot 10^{-4}$	1.013 $\cdot 10^{-3}$	0.1106

Acceptance rate is at 47.89%. O'Neill and Roberts [1999] reports estimates of 0.0009 and 0.098 with variances 0.00019 and 0.02074, which according to them corresponds to maximum likelihood estimates found by some method not described in further detail. The Geweke convergence plot for β is shown in Figure 2.9.

For comparison consider the results of the marginalized likelihood approach. Based on 100 samples each using 1,000 values for the Monte Carlo integration with $\theta = 10$ we obtain $\hat{\beta} = (0.0006, 1.566 \cdot 10^{-11})$ and $\hat{\gamma} = (0.0642, 1.979 \cdot 10^{-7})$, which unfortunately is somewhat away from both the obtained MCMC values and the reported ML values. It appears that in a larger epidemic missing infection times are harder to restore, because constraints do not restrict the space as much as e.g. in the five case epidemic.

2.4.4 Value of Infection data

So far the implementation has only been validated by comparing results for well-known datasets used in the literature. In this section we will use the 10 simulated epidemics with $(N = 10, a = 1)$ shown in Figure 2.1, providing a dataset containing both $\boldsymbol{\tau}$ and $\{I_0\} \cup \boldsymbol{I}$. Like that we can compare

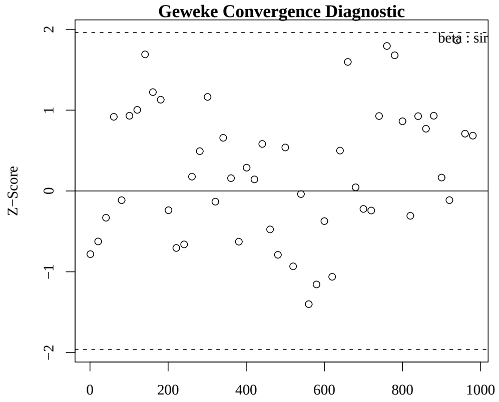


Figure 2.9: The Geweke convergence diagnostic (50 bins) for β in the Abakaliki epidemic.

the estimates obtained using both recovery and infectious times by ML to the ones obtained using MCMC from recovery times alone, see Table 2.4.4. In the latter case, posterior means of 10,000 samples (non-informative priors) together with naive standard error estimates are shown. In case of a singleton epidemic it is not possible to use the MCMC method, because the method requires at least two infection times. ML estimation of $\hat{\beta}$ and $\hat{\gamma}$ in case of singletons is possible, but there is not a lot of information to use; $\hat{\beta}$ will be zero and $1/\hat{\gamma}$ will be equal to the length of the single infectious period. The formula for asymptotic variance will have a zero in the dominator and can thus not be used and the standard deviation of $\hat{\gamma}$ will be equal to $\hat{\gamma}$.

		Full information (ML)				Partial information (MCMC)			
#	m	$\hat{\beta}$	$\hat{\gamma}$	$s.e.(\hat{\beta})$	$s.e.(\hat{\gamma})$	β	γ	$s.e.(\beta)$	$s.e.(\gamma)$
1	1	0.0000	5.8617	NA	5.8671	NA	NA	NA	NA
2	1	0.0000	3.4892	NA	3.4892	NA	NA	NA	NA
3	10	0.1289	0.5293	0.0430	0.1674	0.1282	0.5632	0.0006	0.0027
4	4	0.0506	0.5805	0.0292	0.2903	0.2760	2.9930	0.0021	0.0210
5	12	0.1330	0.2157	0.0421	0.0650	0.4943	0.1580	0.0044	0.0005
6	1	0.0000	4.1580	NA	4.1580	NA	NA	NA	NA
7	2	0.0204	0.3693	0.0204	0.2611	0.0074	0.1357	0.0001	0.0015
8	9	0.0641	0.3563	0.0227	0.1188	0.0658	0.3214	0.0003	0.0015
9	8	0.0522	0.2791	0.0197	0.0987	0.0541	0.3240	0.0003	0.0016
10	11	0.1194	0.3330	0.0378	0.1004	0.1872	0.2090	0.0012	0.0009

Table 2.1: Comparing ML estimation based on access to both $\boldsymbol{\tau}$ and $\boldsymbol{I}$ with MCMC estimation from just $\boldsymbol{\tau}$. The first two columns show, which run of Fig 2.1 is considered together with its number of infections. For MCMC, posterior means together with a naive std. error estimates are shown.

If singleton epidemics are discarded the mean of the remaining seven $\hat{\beta}$ estimates is 0.081 and 0.17 for ML and MCMC, respectively. That the mean is higher in case of partial information is mostly due to the large deviation for run four and five, otherwise the calculations roughly yield the same parameter estimates. Why runs four, five and possibly also 10 yield high $\hat{\beta}$ estimates is hard to say. A possible explanation might be that they all are epidemics with a high number of observed recoveries/infections – not knowing the infection times thus corresponds to throwing a lot of information away thus introducing more parameters. The hypothesis is although contradicted by the results of run 3. Not knowing anything about the infectious times should in principle allow for multiple solution structures – one of them having a high number of infections in the beginning (β high) followed by a long recovery time (with large variance). A more straightforward explanation could be some type of convergence difficulty, which a more detailed analysis could reveal.

It is hard to draw well-founded conclusions from just 10 simulations in a very small population. Of course lack of infection times is loss of information, but a more detailed description of the impact is very context dependent. For some runs it appears that the limited information makes a solution with a high infectious rate followed by a long recovery time more attractive. One thing is although clear - the asymptotic variance estimates obtained in the ML framework are not feasible in the case of a population of size 11! Here, the results obtained from the sampling appear more precise.

The three examples should have illustrated the functionality of the `LadyBug` program in connection with simple SIR epidemics. We are thus ready to move on to extended SIR-models.

Chapter 3

Estimation in generalized SIR models

This chapter discusses two extensions of the SIR model, aimed at making it more applicable to the domain of animal production, together with how to do parameter estimation in these settings. After presenting the necessary theory, the methods are illustrated using both simulated epidemics and data from a classical swine fever epidemic in the Netherlands. As in the last chapter, these illustrations concurrently works as a test for the crafted implementation.

3.1 The extensions

The first extensions allows for switching between k different sets of parameters in the SIR model. Focus is on a model, where the active regime at time t is completely determined by the value of t. An example is to divide (I_0, T) into two disjunct intervals expressing that in the first interval SIR-parameters are (β_1, γ_1) while being (β_2, γ_2) in the second interval. Advanced switching such as e.g. switched Markov processes [Holst and Madsen, 2000], where the active regime depends on past values, could be relevant, but is beyond the scope of this report. To introduce the necessary notation, let the ordered set of switch points

$$\boldsymbol{S} = (s_{12}, s_{23}, \ldots, s_{k-1k}), \text{ such that } s_{12} < s_{23} < \ldots < s_{k-1k},$$

determine the active regime at time t. To keep things simple we assume that $\boldsymbol{S}$ is known – in many real world applications it might be of interest to estimate $\boldsymbol{S}$ by itself. Now, let the function $\text{cr}(t) : t \to \{1, \ldots, k\}$ determine the active regime at time t, i.e.

$$\text{cr}(t) \equiv \begin{cases} 1 & \text{if } t < s_{12} \\ 2 & \text{if } s_{12} \le t < s_{23} \\ \ldots & \\ k & \text{if } t \ge s_{(k-1)k} \end{cases}$$

As before the population consists initially of N susceptibles and $a = 1$ infected. Transition intensities of $(X(t), Y(t)) = (i, j)$ at time t are given by

$$(i, j) \to (i-1, j+1) \quad : \quad \beta_{\text{cr}(t)} ij \tag{3.1}$$

$$(i, j) \to (i, j-1) \quad : \quad \gamma_{\text{cr}(t)} j. \tag{3.2}$$

Note that the above continuous time process no longer satisfies the Markov property. A switched SIR model allows handling situations, where e.g. different control measures and management routines have been attempted over time to reduce the spread of the disease and limit its consequences.

An example is the classical swine fever epidemic in the Netherlands [Pluimers *et al.*, 1999; Stegeman *et al.*, 1999] analyzed in Section 3.5.3. Here, authorities increased control measures stepwise in an attempt to eradicate the disease. Each regime would then correspond to a fixed set of control measures practiced in a specified time interval. Parameter estimation from data can then be used to see how the measures had an effect on the β and γ parameter. A less comprehensive example with only two regimes is used to illustrate the idea.

EXAMPLE 3.1 (TWO-REGIME SIR-MODEL)
Consider an epidemic in a population with $N = 34$, $a = 1$, and $\boldsymbol{S} = (5.1216)$. We let $\beta_1 = 0.01, \beta_2 = 0.1, \gamma_1 = 0.3$, and $\gamma_2 = 0.3$. Figure 3.1(a) shows a single realization of this epidemic. Notice how the increase in β after the switch causes the second top in the number of infected at $t \approx 12$. In (b) the effect on the final size can be seen from a histogram based on 10,000 samples. Final size of the switching model (bars) tends to be larger compared to a single regime model with Regime 1 parameters, but smaller than a single model with Regime 2 parameters. 27.24% of the 10,000 simulations end up in regime 2.

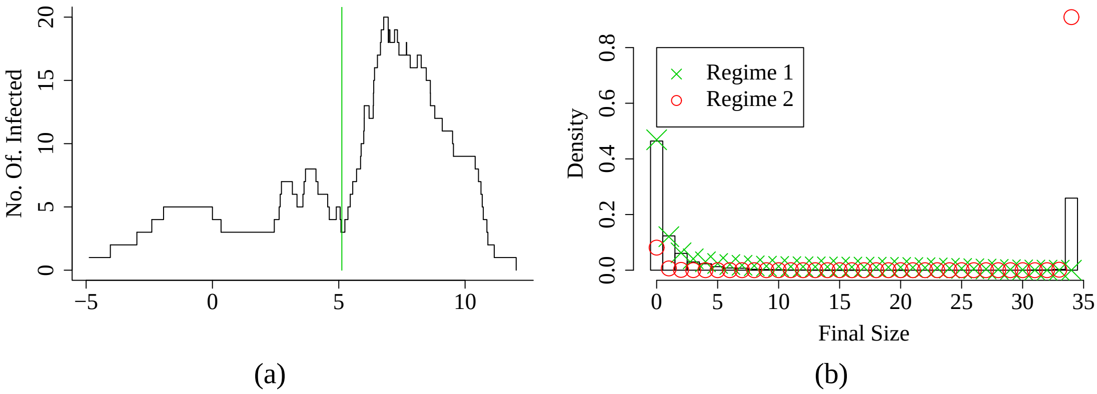


Figure 3.1: (a) Number of infected as a function of time for a single simulation of the model. The vertical line at 5.1216 indicates the switch point. (b) Bars show final size distribution of the epidemic in Example 3.1 obtained by 10,000 samples. Crosses and circles show probabilities of a single model with parameters as in Regime 1 and Regime 2, respectively.

◇

Whereas the basic model assumes homogeneity of the entire population, the second extension allows handling a set U, $|U| = l$, of sub-populations. Within each sub-population homogeneity is still the case, but we now enable heterogeneity between the interacting populations. Such a pattern could for example originate from a spatial arrangement of the populations giving different conditions to spread because contact behavior is limited. To keep it simple, the spatial structure is assumed to be of grid tiling as e.g. in [Marshall *et al.*, 2001; Höhle, 2001], with the further assumption that only neighboring populations immediately interact. More advanced spatial structures and interaction schemes are of course imaginable, e.g. counties of a country where exchange of infectious material is proportional to the Euclidean distance.

In the spatial setup, each unit $u \in U$ has its own set of parameters β^u, γ^u. For notational convenience all involved quantities are formulated in vectors of length l. That is, let $\boldsymbol{N} = (N^1, \ldots, N^l)$ be a vector denoting the different initial number of susceptibles and let $\boldsymbol{a}$ be zero except for the component corresponding to the unit hosting the initial infective. At each time time-point t, $(\boldsymbol{i}, \boldsymbol{j})$

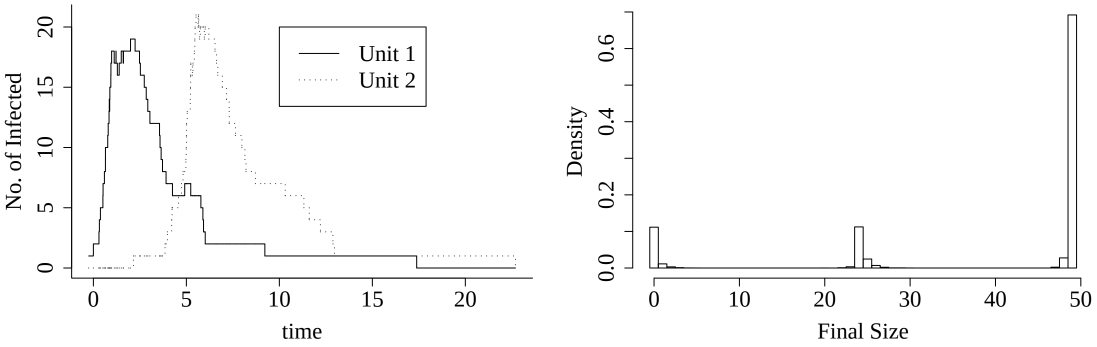
denotes the number of susceptibles and infectious in each unit. Transition intensities for unit u are now given as

$$(\boldsymbol{i},\boldsymbol{j}) \to (\boldsymbol{i}[i^u-1],\boldsymbol{j}[j^u+1]) \quad : \quad \beta^u i^u j^u + \sum_{s\in N_4(u)} \beta^{u\to s} i^u j^s \tag{3.3}$$

$$(\boldsymbol{i},\boldsymbol{j}) \to (\boldsymbol{i},\boldsymbol{j}[j^u-1]) \quad : \quad \gamma^u j^u, \tag{3.4}$$

where $\boldsymbol{i}[i^u-1] \equiv (i^1,\ldots,i^u-1,\ldots,i^l)$ and $N_4(u)$ denotes the city-block neighborhood of u. Notice the additional $\beta^{u\to s}$ parameters introduced to handle interaction between neighbors. The formulated model is although still just an instance of a *multi-type epidemic*, see [Andersson and Britton, 2000, Chapter 6]. To reduce the number of parameters one could assume that all β^u and γ^u are equal and that one parameter is sufficient to describe the interaction between any two units, i.e. a total of three parameters needs to be estimated.

EXAMPLE 3.2 (TWO UNIT MODEL)
Consider a one-regime model with two units each having their own set of (β,γ) parameters but $\beta^{1\to2} = \beta^{2\to1} = \beta^n$. Let $N = (24,25), a = (1,0), \beta^1 = \beta^2 = 0.1, \beta^n = 0.001$, and $\gamma^1 = \gamma^2 = 0.3$. Figure 3.2(a) shows the trajectory of an epidemic with such a configuration, in (b) the final distribution is shown based on 10,000 simulations. Note, how the epidemic either ceased straight away, infects all animals within the first unit (without spreading to the second), or infects practically all animals in the two units.



(a)

(b)

Figure 3.2: (a) Trajectory of the number of infected in each unit as a function of time. Note how the disease jumps to Unit 2, when there are a maximal of infected in Unit 1. (b) Shows the final size distribution of the chosen setup based on 10,000 simulations.

◇

Instead of dividing the population according to spatial membership one could divide the overall population into populations with varying susceptibility as in [Hayakawa *et al.*, 2000]. Here, the number of susceptibles in each group, $X(t)$, is tracked whereas the number of infected $Y(t)$ is the sum of the number of infected in each sub-population. Transition intensity for an infectious event in population i is then $\beta^i X^i(t)Y(t)$. This yields nice formulas both for ML-Estimation in case of full data and MCMC-Estimation in case of just recovery times, see [Hayakawa *et al.*, 2000]. The derived equations from this chapter can with only slight modifications be used to perform the necessary estimation.

Finally, a model containing both regime switching and spatial extensions is possible. Here, each unit u consists of k different SIR-regimes with common switch point vector $\boldsymbol{S}^u$. In the following, we will use subscripts to denote regime and superscripts to denote unit, i.e. β_1^2 is the β parameter of the first regime of unit two.

3.2 Estimation in the extended models with full data

Despite of different regimes and units, focus is still on an epidemic starting from a single initial infective. Knowledge about the location of the initial infective is an assumption, which in practice can be hard to achieve. Similar derivations as in Section 2.2 are now performed to obtain a likelihood expression. It is possible to handle the two extension from the previous framework within one framework if we in addition to infection and recover events introduce switch events. Furthermore, each event now needs to carry information about which unit it occured in. That is, an event ϵ consists of the entries (t, e, u), where $t \in \mathbb{R}$ denotes time of occurrence, $e = \{\text{inf}, \text{rec}, \text{sw}\}$ is the event type, and $u = \{1, \dots, l\}$ denotes the location. The functions $\mathrm{t}(\epsilon)$, $\mathrm{e}(\epsilon)$, and $\mathrm{u}(\epsilon)$ are introduced to extract this information from an event record. Note that the active regime at ϵ can be deduced by using the $\mathrm{cr}(\mathrm{t}(\epsilon))$ function. Some additional notation is necessary to get through the derivation of the likelihood. Let the ordered sets

$$\boldsymbol{I}_1^1, \boldsymbol{\tau}_1^1, \dots, \boldsymbol{I}_k^1, \boldsymbol{\tau}_k^1, \boldsymbol{I}_1^2, \boldsymbol{\tau}_1^2, \dots, \boldsymbol{I}_k^2, \boldsymbol{\tau}_k^2, \dots \boldsymbol{I}_1^l, \boldsymbol{\tau}_1^l, \dots, \boldsymbol{I}_k^l, \boldsymbol{\tau}_k^l$$

contain all i and r events associated to their respective regimes and units. Note that the first infection, I_0, is handled separately and is therefore not part of any other set. For notational convenience let $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_k\}, \boldsymbol{I} = \{\boldsymbol{I}_1, \dots, \boldsymbol{I}_k\}, \boldsymbol{\beta} = \{\beta_1^1, \dots, \beta_k^l\}$, $\boldsymbol{\beta}^n = \{\beta_1^n, \dots, \beta_k^n\}$ and $\boldsymbol{\gamma} = \{\gamma_1^1, \dots, \gamma_k^l\}$. Furthermore, for each unit u we artificially generate a switch event for each element of $\boldsymbol{S}$; the ordered set $\boldsymbol{S^u}$ is then used to denote the k switch events generated for section u. A switch event does not change $X(t)$ nor $Y(t)$ and is only used to handle changes in the hazard functions due to regime shifts.

As before, the waiting time between events is modeled, i.e. let

$$\boldsymbol{\xi} = (\epsilon_0, \dots, \epsilon_{\text{last}}) = I_0 \cup \left(\bigcup_{j=1,\dots,l} \boldsymbol{S^j} \cup \left(\bigcup_{i=1,\dots,k} \boldsymbol{I_i^j} \cup \boldsymbol{\tau_i^j} \right) \right)$$

be the ordered set of all events sorted according to their event-time. The likelihood for the waiting time $t_i - t_{i-1}$ can again be expressed by a multiple failure model using several cause-specific hazard functions. For a single location $u \in U$ the overall-hazard function is given by

$$\lambda^u(t|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda^u_{\text{inf}}(t|\boldsymbol{\beta}, \boldsymbol{\gamma}) + \lambda^u_{\text{rec}}(t|\boldsymbol{\beta}, \boldsymbol{\gamma}) + \lambda^u_{\text{sw}}(t|\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where $\lambda^u_{\text{inf}}(t|\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\lambda^u_{\text{rec}}(t|\boldsymbol{\beta}, \boldsymbol{\gamma})$ are given by equations (3.1-3.2), (3.3-3.4), or a combination. In case of switching, the $\mathrm{cr}(t)$ function will determine, the appropriate regime to use. For the artificial switch events the following hazard is used.

$$\lambda^u_{\text{sw}}(t|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \left\{ \begin{array}{ll} 1 & \text{if } t \in \boldsymbol{S^u} \\ 0 & \text{otherwise} \end{array} \right. .$$

The overall-hazard for all units is then given as $\lambda(t|\boldsymbol{\beta},\gamma) = \sum_{u\in U} \lambda^u(t|\boldsymbol{\beta},\gamma)$. Hence, the likelihood for all observations can be derived as

$$L = f(\boldsymbol{\tau}, \boldsymbol{I}|\boldsymbol{S}, \boldsymbol{\beta}, \gamma, I_0) = \prod_{\epsilon_i \in \boldsymbol{\xi}\setminus\{\epsilon_0\}} \lambda^{u(\epsilon_i)}_{e(\epsilon_i)}(t(\epsilon_i)|\boldsymbol{\beta},\gamma) \exp\left(- \int_{t(\epsilon_{i-1})}^{t(\epsilon_i)} \lambda(t|\boldsymbol{\beta},\gamma)\,\mathrm{dt} \right). \tag{3.5}$$

Due to its shape, all $\lambda^u_{\text{sw}}(t|\boldsymbol{\beta},\gamma)$ terms can be ignored in the integral of the overall hazard. As expected, the likelihood boils down to (2.2) in case of only one unit and one regime. In case of a fully observed epidemic, ML estimates can be obtained using numerical optimization of Equation (3.5). In special cases such as the regime switching model — a closed form can although be derived. Two examples illustrate this point.

EXAMPLE 3.3 (ESTIMATION IN A TWO REGIME MODEL)
Consider a two-regime model with only one unit as in Example 3.1. This allows dropping the unit superscripts. Let the switch point list be given by $\boldsymbol{S} = (s_{12})$, where $I_0 < s_{12} < T$. Then the ML-estimates can be derived from (3.5) as

$$\hat{\beta}_1 = \frac{|\boldsymbol{I}_1|}{\int_{I_0}^{s_{12}} X(t)Y(t)\,\mathrm{dt}}, \quad \hat{\gamma}_1 = \frac{|\boldsymbol{\tau}_1|}{\int_{I_0}^{s_{12}} Y(t)\,\mathrm{dt}}, \quad \hat{\beta}_2 = \frac{|\boldsymbol{I}_2|}{\int_{s_{12}}^{T} X(t)Y(t)\,\mathrm{dt}}, \quad \hat{\gamma}_2 = \frac{|\boldsymbol{\tau}_2|}{\int_{s_{12}}^{T} Y(t)\,\mathrm{dt}}.$$

The above is easily generalized to a general k-regime switching models, as long as $l = 1$. To derive standard error for the ML-estimators one would have to go through the counting process derivations of [Andersson and Britton, 2000, Section 9.2]. Furthermore, the easy derivation of ML-estimates also encourages Gibbs updating for the various β, γ parameters of a regime SIR-model in case of a partially observed epidemic.

◇

EXAMPLE 3.4 (ESTIMATION IN TWO UNIT MODEL)
This example covers estimation in a two-unit model as in Example 3.2. Using Equation 3.5 the ML-estimate for e.g. β^1 is found by solving the following equation.

$$\frac{\partial}{\partial \beta^1} \left[\prod_{\epsilon \in \boldsymbol{I}^1} X^1(\mathrm{t}(\epsilon)) \left(\beta^1 Y^1(\mathrm{t}(\epsilon)) + \beta^n Y^2(\mathrm{t}(\epsilon)) \right) \exp\left(- \int_{I_0}^{T} \beta^1 X^1(\mathrm{t}(\epsilon)) Y^1(\mathrm{t}(\epsilon))\,\mathrm{dt} \right) \right] = 0$$

Unfortunately, the term $\beta^1 Y^1(\mathrm{t}(\epsilon)) + \beta^n Y^2(\mathrm{t}(\epsilon)$ makes it impossible to obtain a nice expression from the above differentiation, because the sum within the product corresponds to a polynomial of degree $|\boldsymbol{I}^1|$, with all terms being non-zero. Rephrasing the two population/unit model using counting process methodology along the lines of [Britton, 1998] yields a more compact representation. To do this, two counting process are formulated.

$$\begin{aligned} I_i(t) = n_i - Y_i(t) \quad &\text{with intensity} \quad X_i(t-)\boldsymbol{\beta}^{i\bullet}\boldsymbol{Y}(t-) \\ R_i(t) \quad &\text{with intensity} \quad \gamma^i Y_i(t-), \end{aligned}$$

where $i = \{1,2\}, \boldsymbol{Y(t)} = (Y_1(t), Y_2(t))^T, \boldsymbol{\beta}^{1\bullet} = (\beta^1, \beta^n)$, and $\boldsymbol{\beta}^{2\bullet} = (\beta^n, \beta^2)$. In other words I_i counts the number of infected in unit i and R_i counts the number of recovered. Using Andersen *et al.* [1993, p.402] the log-partial-likelihood of the data observed up to time t can be written as

$$\begin{aligned} l_t(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \sum_{i=1}^{2}\int_{I_0}^{t} \log\Big(X_i(u-)\boldsymbol{\beta}^{i\bullet}\boldsymbol{Y}(u-)\Big)\, dI_i(u) - X_i(u)\boldsymbol{\beta}^{i\bullet}\boldsymbol{Y}(u-)du \\ &+ \sum_{i=1}^{2}\int_{I_0}^{t} \log(\gamma^i Y_i(u-))dR_i(u) - \gamma^i Y_i(u-)du \end{aligned} \tag{3.6}$$

ML estimators for the data are now found by differentiating $l_t(\boldsymbol{\beta},\boldsymbol{\gamma})$ with respect to each parameter, setting each such derivative to zero, and solving the obtained equation system. Introducing the shorthand notation $G_{ij}(t) = \int_{I_0}^{t} X_i(u-)Y_j(u-)du$ the derivates are as follows.

$$\begin{aligned} \frac{\partial l_t(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\beta^1} &= \int_{I_0}^{t} \frac{Y_1(u-)}{\boldsymbol{\beta}^{1\bullet}\boldsymbol{Y}(u-)} dI_1(u) - G_{11}(t) \\ \frac{\partial l_t(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\beta^2} &= \int_{I_0}^{t} \frac{Y_2(u-)}{\boldsymbol{\beta}^{2\bullet}\boldsymbol{Y}(u-)} dI_2(u) - G_{22}(t) \\ \frac{\partial l_t(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\beta^n} &= \int_{I_0}^{t} \frac{Y_2(u-)}{\boldsymbol{\beta}^{1\bullet}\boldsymbol{Y}(u-)} dI_1(u) - G_{12}(t) + \int_{I_0}^{t} \frac{Y_1(u-)}{\boldsymbol{\beta}^{2\bullet}\boldsymbol{Y}(u-)} dI_2(u) - G_{21}(t) \end{aligned}$$

Because $dI_i(t)$ is 1 at times t, where the counting process increases and 0 otherwise, the integrals involving $dI_i(u)$ are equal to sums. Because β^n is part of the for both β^1 and β^2 it is necessary to solve the system

$$\{\frac{\partial l_t(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\beta^1} = 0, \frac{\partial l_t(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\beta^2} = 0, \frac{\partial l_t(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\beta^n} = 0\}$$

simultaneously, which complicates the solution dramatically and makes a closed form solution impossible. Lack of finding a closed solution for $\boldsymbol{\beta}$ also indicates that in case of partially observability it might not be possible to keep the Gibbs updating of the β parameters — instead Metropolis-Hasting has to be used. Instead a numeric solution is required. Getting estimates for $\boldsymbol{\gamma}$ on the other hand is easy.

$$\frac{\partial l_t(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\gamma^i} = \int_{I_0}^{T} \frac{1}{\gamma^i} dR_i(u) - \int_{I_0}^{T} Y_i(u-)du = 0 \quad \Leftrightarrow \quad \gamma^i = \frac{|\boldsymbol{\lambda}_i|}{\int_{I_0}^{T} Y_i(u)du}, \tag{3.7}$$

corresponding to our earlier results. This also indicates that it will be easy to adopt the Gibbs updating for $\boldsymbol{\gamma}$ in case of partial observability.

◇

Concluding from the derivations of the above examples: Operating with several interacting spatial units complicates the picture in a degree, where solution to the ML equations appears to become hard if not intractable. Modeling with switched SIR regimes, on the other hand, is feasible.

3.3 Estimation with infection times missing

After the initial discussion in the previous section, this section takes a closer look on parameter estimation based on only recovery times in the model extensions. For a setup with l units each having k regimes the parameters to estimate are the $|l \times k|$ sets of (β_i^j, γ_i^j) and the k β_i^n's. Again we consider Bayesian approach using MCMC and sketch an marginalized likelihood approach. Derivations are of similar nature as in Section 2.2.

3.3.1 Bayesian analysis using MCMC

Our dataset is $\boldsymbol{D} = \boldsymbol{\tau}_1^1 \cup \ldots \cup \boldsymbol{\tau}_k^l$, therefore all the missing infection times, $I_0, \boldsymbol{I}_1^1, \ldots, \boldsymbol{I}_k^l$ again become parameters. To calculate posterior distributions we again look for a Markov Chain Monte Carlo method as in the last chapter. Here, we were fortune that full-conditionals could be specified for β and γ allowing for a Gibbs-within-Metropolis construction. As before prior gamma distributions are assumed for the β and γ parameters, i.e.

$$\beta_i^j \sim \Gamma(\nu_{\beta_i^j}, \lambda_{\beta_i^j}), \qquad \gamma_i^j \sim \Gamma(\nu_{\gamma_i^j}, \lambda_{\gamma_i^j}).$$

Because the derivation of the full conditionals are tightly connected with the maximum likelihood equations it is clear that it will not be possible to establish full conditionals in case of more than one unit. Future work has to determine whether we can do something more efficient than a cumbersome Metropolis Hasting for the β parameters. In the following we will only consider a SIR model with k regimes, switch-points $\boldsymbol{S} = (s_{12}, \ldots, s_{k-1k})$ and a *single* unit. Deriving the posterior for some β_i is similar to (2.5) except that the likelihood from (3.5) is used.

$$\begin{aligned}
\pi(\beta_i|\boldsymbol{\tau}, \boldsymbol{I}, \boldsymbol{S}, I_0, \boldsymbol{\beta}\backslash\{\beta_i\}, \boldsymbol{\gamma}) &\propto f(\boldsymbol{\tau}, \boldsymbol{I}|\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\gamma}, I_0)\pi(\beta_i) \\
&\propto \beta_i^{|\boldsymbol{I}_i|} \exp\left(-\int_{s_{(i-1)i}}^{s_{i(i+1)}} \beta_i X(t)Y(t)\,\mathrm{dt}\right) \beta_i^{\nu_{\beta_i}-1} \exp(-\lambda_{\beta_i}\beta_i) \\
&\propto \beta_i^{\nu_{\beta_i}+|\boldsymbol{I}_i|-1} \exp\left(-\lambda_{\beta_i}\beta_i - \beta_i \int_{s_{(i-1)i}}^{s_{i(i+1)}} X(t)Y(t)\,\mathrm{dt}\right) \\
&\sim \Gamma\left(\nu_{\beta_i} + |\boldsymbol{I_i}|, \lambda_{\beta_i} + \int_{s_{(i-1)i}}^{s_{i(i+1)}} X(t)Y(t)\,\mathrm{dt}\right)
\end{aligned}$$

Notice how the results in a setup with non-informative priors corresponds to the result of the maximum likelihood derivations from Example 3.3. If regime i does not contain any infection events, i.e. $|I_i| = 0$, $\nu_{\beta_i} > 0$ is required – otherwise we would have an improper Γ distribution. In case of many regimes it his very likely that one or more of the regimes will be empty as infection times are moved around, i.e. it is wise to assign informative priors to all regimes. Similar calculations for γ_i yield

$$\pi(\gamma_i|\boldsymbol{\tau}, \boldsymbol{I}, \boldsymbol{S}, I_0, \boldsymbol{\beta}, \boldsymbol{\gamma}\backslash\{\gamma_i\}) \sim \Gamma\left(\nu_{\gamma_i} + |\boldsymbol{\tau_i}|, \lambda_{\gamma_i} + \int_{s_{(i-1)i}}^{s_{i(i+1)}} Y(t)\,\mathrm{dt}\right).$$

Again it is wise to specify informative priors for γ_i in case $\boldsymbol{\tau}_i$ should become empty.

Finally, the posterior of I_0 given all other infection times is to be calculated. To do this, it is sufficient to consider the time up to ϵ_1, the first event occuring after I_0 — this could either be the next infection, $(\boldsymbol{I}_u)_1$, occuring in some unit u, the switch point s_{12}, or the first recovery $(\boldsymbol{\lambda}_1)_1$ in case of a singleton epidemic. Using ϵ_1 instead of I_1, derivations are just as in Equation (2.5), thus

$$\pi(-I_0|\boldsymbol{\tau}, \boldsymbol{I}, \boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \sim \operatorname{Exp}(\beta_1 N + \gamma_1 + \theta). \tag{3.8}$$

Finding the full conditional in case of several units is problematic, because β^u for unit u is part of the hazard function of other units as well. This makes Gibbs sampling from the full-conditional for $\boldsymbol{\beta}$ impossible, instead a random-walk Metropolis updating scheme is applied to each component of $\boldsymbol{\beta}$. Proposing $\beta_i'^u$ as the new value of β_i^u in $\boldsymbol{\beta}$ we accept it with probability

$$\alpha(\beta_i^u, \beta_i'^u) = \min\left(1, \frac{f(\boldsymbol{\tau}, \boldsymbol{I}|I_0, \boldsymbol{S}, \boldsymbol{\gamma}, (\boldsymbol{\beta}\backslash\{\beta_i^u\}) \cup \{\beta_i'^u\})\pi(\beta_i'^u)}{f(\boldsymbol{\tau}, \boldsymbol{I}|I_0, \boldsymbol{S}, \boldsymbol{\gamma}, \boldsymbol{\beta})\pi(\beta_i^u)}\right).$$

The already described Gibbs-sampling for $\boldsymbol{\gamma}$ can be directly adopted for the spatial setting by adding appropriate sums over all units, where as the full-conditional for I_0 is changed to

$$\pi(-I_0|\ldots) \sim \text{Exp}\left(\theta + \beta^{u(I_0)}N(u(I_0)) + \gamma^{u(I_0)} + \sum_{u' \in N_4(u(I_0))} \beta^n N^{u'}\right),$$

where $u(I_0)$ is the unit, where I_0 occurs, and $N(\cdot)$ denotes the number of initial susceptibles in this unit. Note that the location of I_0 remains fixed due to our assumptions. It is although not too hard to relax this by treating I_0 as any other element in $\boldsymbol{I}$. Should a proposal – possibly from a different unit – become the new first infection one just needs to take the prior of I_0 into account when calculating the acceptance probability of the Metropolis step.

3.3.2 Sketch of a Marginalized Likelihood approach

Again a Monte Carlo approximation of the marginalized likelihood is used to find its maximizer $(\hat{\boldsymbol{\beta}}, \hat{\gamma})$. Derivations are similar to the ones performed in Section 2.4.2 except that we now operate with $f(\boldsymbol{\tau}, \boldsymbol{I}|I_0, \boldsymbol{\beta}, \gamma)$ as defined by Equation (3.5). Sampling uniformly over the space of all valid $\{I_0\} \cup \boldsymbol{I}$ also becomes a little more complicated with the introduction of units, because we only need to ensure that

- $Y^u(t) \geq 0$ while $t \in [I_0, T]$, for each unit $u \in U$
- $\sum_{u \in U} Y^u(t) > 0$ while $t \in [I_0, \tau_{\texttt{last}}[$, where $\tau_{\texttt{last}}$ is the last observed recovery.

Opposite to the single unit setup, $Y^u(t)$ may drop to zero, because this does not mean the epidemic necessarily dies in the unit; a new infection can occur by transmission from another unit $u' \in N_4(u)$ where $Y^{u'}(t) > 0$. Again, the crucial problem is how to sample the first value I_0. Once this value is found, the remaining values should be sampled independently, sorted, and compared to the above constraints until a valid sample is found. We will refrain from a discussion on how to sample I_0 and just be content with an approximative solution as in Section 2.4.2.

3.4 Model Selection

When choosing an appropriate extensions of the basic SIR-model to model available data, the important question is how much of an extension is actually needed to fit the data. A classical choice is to select a full model and then try to investigate whether sub-models hereof are sufficient by use of hypothesis testing. Other ways of performing model selection is to consider a tradeoff between number of parameters and fit by using an information criterion such as AIC or BIC [DeGroot, 1989]. We will only discuss model selection within a Bayesian context by use of posterior sampling such as MCMC. To test model M_0 against model M_1 given data D the Bayesian choice is to compute the *Bayes Factor*

$$B_{21} = \frac{P(D|M_1)}{P(D|M_2)},$$

i.e. the ratio of the marginal likelihoods

$$P(D|M_k) = \int P(D|\theta_k, M_k)P(\theta_k|M_k)d\theta_k,$$

where θ_k, $k = 1, 2$, are the parameters of model M_k and $P(\theta_k|M_k)$ is the prior density of these parameters. Hence, B_{21} can be interpreted as the evidence for M_2 against M_1 provided by the data. The following table taken from [Gilks *et al.*, 1996, p.165] quantifies the evidence as follows.

B_{21}	evidence for M_2
< 1	negative (supports M_1)
1 to 3	barely worth mentioning
3 to 12	positive
12 to 150	strong
> 150	very strong

The question is now how to calculate $P(D|M_k)$ based on posterior samples obtained by e.g. an MCMC method suffering from samples not independent and only approximatively drawn from the desired distribution. Gilks *et al.* [1996, p.169] suggests using the harmonic mean of the likelihood computed for each posterior sample, i.e.

$$\hat{P}_2(D|M_k) = \left(\frac{1}{k} \sum_{i=1}^{k} \frac{1}{L_{M_k}(\theta_k^{(i)})} \right)^{-1},$$

where $L_{M_k}(\theta_k^{(i)})$ denotes the likelihood under M_k calculated from the i'th posterior sample of the parameter vector. The estimator converges almost surely to the correct value as $k \to \infty$, but it has some instability problems, because $\theta_k^{(i)}$'s with small likelihood tend to dominate it. Various proposals exist to overcome this problem, e.g. using a mixture of prior and posterior samples etc., but we will stick with $\hat{P}_2(D|M_k)$ and refer to [Gilks *et al.*, 1996, Chapter 10] for a discussion of improvements.

An alternative way to guide the model selection is to use the *Deviance Information Criterion* (DIC) [Spiegelhalter *et al.*, 2002], but the problem is that this measure does not readily take restrictions in the state space into account. In our case, the constraints on the infection times infer problems when calculating the deviance of the posterior mean.

3.5 Testing the implementation

To investigate the applicability of theory presented in this chapter, the implementation from Section 2.4 is extended and three examples are shown. Because programming and debugging the stochastic algorithms has a great potential of being faulty, the capability to generate simulated data from an extended SIR-regime model is a necessity. Contrary to Section 2.2 no analysis (at least not to our knowledge) with switched regimes exists in the literature, which could be used to validate the implementation. By providing both infection and recovery times, simulated data allows estimation in both the full and partial data scenario enabling a comparison and hence an idea about the value of information embedded in the infection times.

3.5.1 The two regime model

Maximum likelihood estimation in the two regime from Example 3.3 using both infection and recovery times shown in Figure 3.1 are $\hat{\beta}_1 = 0.0107, \hat{\gamma}_1 = 0.2514, \hat{\beta}_1 = 0.0862, \hat{\gamma}_1 = 0.3492$ which looks quite good compared to the true values.

To perform estimation in case of only recovery times in the Bayesian framework we base the analysis on the following priors $\nu_{\beta_1} = 0.01, \lambda_{\beta_1} = 1, \nu_{\gamma_1} = 0.3, \lambda_{\gamma_1} = 1, \nu_{\beta_2} = 0.1, \lambda_{\beta_2} = 1, \nu_{\gamma_2} = 0.3$, and $\lambda_{\gamma_1} = 1$. After 1,000 burn-in samples, 5,000 samples are generated by taking every 100th sample. Results are as follows.

	Mean	SD	Naive SE	MC Error	Batch SE	Batch ACF
β_1	0.0225	0.009864	0.0001395	0.0002638	0.0002949	0.1444
β_2	0.0904	0.09458	0.001338	0.0023174	0.002882	0.003100
γ_1	0.4912	0.2492	0.003524	0.006798	0.007954	0.2292
γ_1	0.4392	0.1476	0.002088	0.004848	0.005664	0.1557

Acceptance rate at the end is at 32.81%. Also of interest is how many infection times are present in the two regimes, see Figure 3.3. Here, the first regime always contains at least 10 infection times, i.e. $|\boldsymbol{I_1}| > 10$ at all times, because 10 recovery times are situated to the left of the switch point, forcing the corresponding infection times to be in $\boldsymbol{I_1}$.

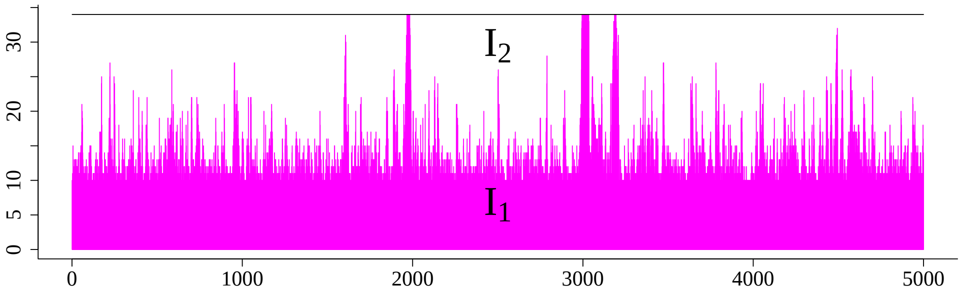


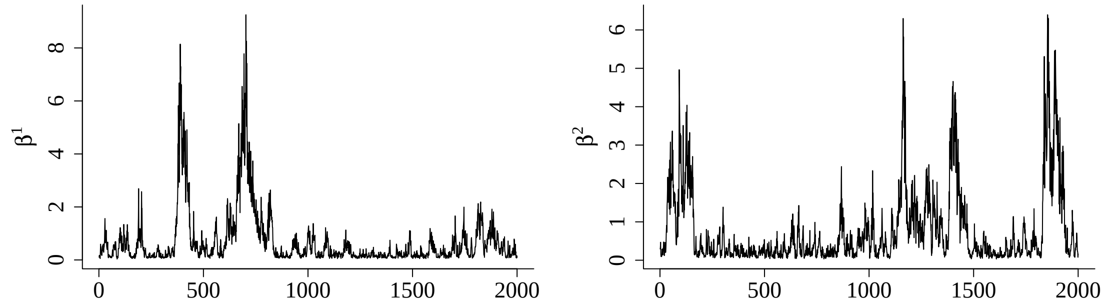
Figure 3.3: Size of $|\boldsymbol{I}_1|$ and $|\boldsymbol{I}_2|$ as a function of the sample number. Note that at all times $|\boldsymbol{I}_1| + |\boldsymbol{I}_2| = 34$.

If we want to investigate whether a single regime model is sufficient to fit the data, one could consider the Bayes factor between a model with one and two regimes. Such an analysis based on 1,000 samples yields $\hat{P}_2(D|M_2) = 334.26$ against $\hat{P}_2(D|M_1) = 304.83$ thus (barely) supporting the two regime model. But it also becomes clear that the harmonic mean metric is dominated by a few likelihood values.

3.5.2 Two unit model

In the two-unit of Example 3.2 finding $\hat{\boldsymbol{\gamma}}$ is done by Equation (3.7), yielding $\gamma^1 = 0.3172$ and $\gamma^2 = 0.2642$, which is in good agreement with the desired values. To find $\hat{\boldsymbol{\beta}} = (\beta^1, \beta^2, \beta^n)$ Equation 3.6 is numerically maximized, which proved to be easier than solving the three dimensional non-linear equation system of the derivates. Solutions for $\boldsymbol{\gamma}$ are inserted and a gradient descent method with line optimizations, see [Press *et al.*, 1992, Chapter 10], is used to find a maxima of the log-partial-likelihood. Assuming that each parameter is greater than zero we optimize the $\log$ of the parameter vector, which is unconstrained. Hereby we obtain $\hat{\beta}^1 = 0.1165$, $\hat{\beta}^2 = 0.0939$, $\hat{\beta}^n = 0.0009$, a solution, which nicely agrees with the true values and also is surprisingly stable to different starting values.

Estimation from recovery times alone is done using the above described MCMC method. The analysis is based on 2,000 samples using a filtration factor of 5,000 and a burn-in of 100,000 samples. The acceptance rates are 4.39%, 73.33%, and 1.26% for $\boldsymbol{I}$, average of $\beta^i, i \in \{1, 2\}$, and β^n respectively. The low acceptance rate for $\boldsymbol{I}$ indicates that the proposal distribution is not very good requiring a large number of samples before an actual jump in state space is performed. Despite the large number of samples used, the chain does not appear to have converged as trace plots of the β's show in Figure 3.4. Looking at the posterior distribution of the parameters, the values for $\boldsymbol{\gamma}$ look nice (mean 0.3257 and 0.2982) despite of non-convergence, whereas the values for $\boldsymbol{\beta}$ of course are doubtful as revealed by Figure 3.4. Even by tuning option parameters to give slightly higher acceptance rates the convergence problem remains. Convergence could of course just be a matter of a (time-wise tractable) extra number of samples, but a characteristic of the example draws attention: in both sections all susceptibles eventually become infected. This might cause problems, because the great extent of the epidemic does not provide any upper bounds for the β's. Hence, the posterior distribution for the high values will be quite flat making the chain mix badly. A different example provides additional support for this hypothesis.



(a) β^1

(b) β^2

Figure 3.4: Trace plots of β^1 and β^2 values. Note the high serial correlation indicating non-convergence.

Example 3.5 (Two unit model)
Still considering a two unit model with $N = (24, 25)$ and $a = (1, 0)$ we alter the parameters slightly to obtain a less categorical final size distribution as in Figure 3.2(b). Using $\beta^1 = \beta^2 = \beta^n = 0.1$ and $\gamma^1 = \gamma^2 = 3$ a final size distribution as in Figure 3.5 is obtained.

◇

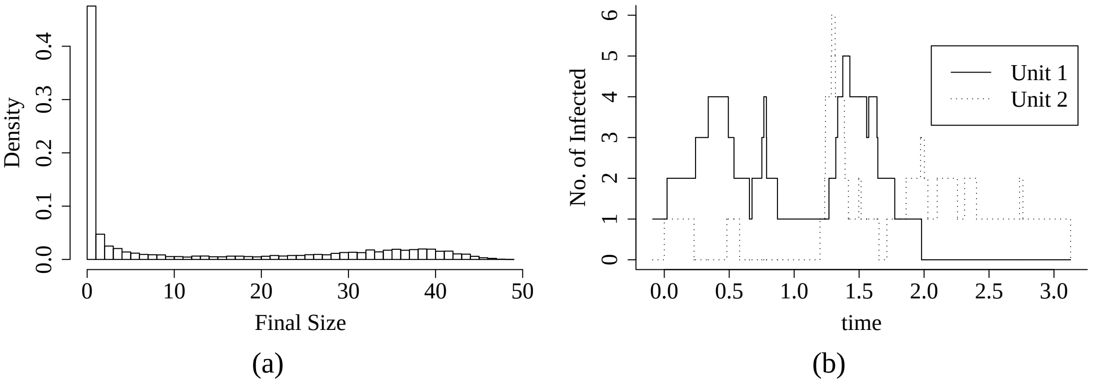


Figure 3.5: (a) Final size of the model obtained by 10,000 simulation. (b) Trajectory from the model selected for estimation with a final size of 26, i.e. an epidemic spreading to both units but stopping before everything is infected.

Estimation for the epidemic shown in Figure 3.5(b) is done as before, except that 2,000 samples with a filter of 500 and burn-in of 100,000 is sufficient. Posterior distribution of the parameters are as follows, where we also given the lower and upper bounds of a 95% credibility interval.

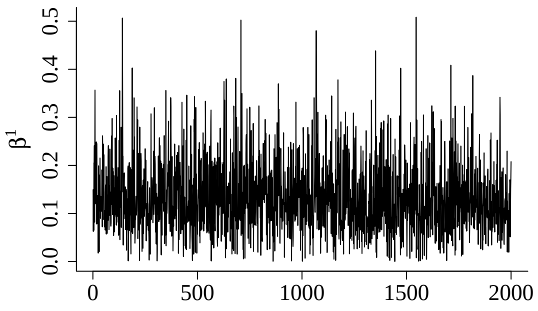
	Mean	SD	CI-lower	CI-upper
β^1	0.1334	0.07407	0.0170	0.311
β^2	0.1379	0.06523	0.0432	0.299
β^n	0.0661	0.03207	0.0161	0.143
γ^1	4.6728	2.48680	1.6150	10.978
γ^2	2.8982	1.24720	1.2287	6.076

For comparison, the ML estimate when the true infection times are know are as follows. $\hat{\beta}^1 = 0.1036$, $\hat{\gamma}^1 = 2.6215$, $\hat{\beta}^2 = 0.1983$, $\hat{\gamma}^2 = 4.6569$, and $\hat{\beta}^n = 0.0568$. That is, the MCMC $\boldsymbol{\gamma}$ looks good, where as $\boldsymbol{\beta}$ reflects that in a spatial setup it is hard to distinguish between within and between unit infections – especially if infection times are missing. But more important; this time the trace plots show a more desired pattern with respect to convergence, see Figure 3.6. Providing a single example to support the hypothesis that convergence is easier to obtain if the extent of the epidemic is moderate of course is not sufficient, but we discovered the phenomena more than once while using other trajectories.

3.5.3 The Classical Swine Fever epidemic in the Netherlands

To move beyond toy-epidemics using simulated numbers, this section looks at data from the classical swine fever (CSF) epidemic in Holland 1997-1998 [Pluimers *et al.*, 1999]. The analysis is made to emphasize the applicability of the method rather than thoroughly analyzing the data and interpret the results. If the latter is of interest, analysis should be performed on a more detailed dataset than the one available to us.

Classical swine fever is a viral disease where symptoms such as dullness and anorexia may be observed within the first days after infection. In its acute form, which need not occur, CSF has a rapid development towards mortality, sometimes even without showing any clinical signs. Often only secondary symptoms such as diarrhea or respiratory problems are observed [Elbers *et al.*,

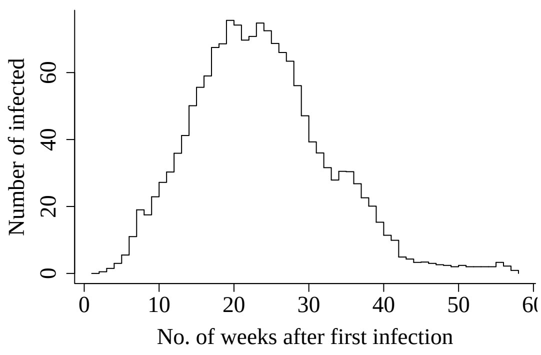


(a) β^1

(b) β^2

Figure 3.6: Trace plots of the 2,000 β^1 and β^2 samples for the trace shown in Figure 3.5(b). Note the apparent improved convergence.

1999]. On the 4th of February 1997 an outbreak of CSF on a pig farm in the southern part of the Netherlands was laboratory wise confirmed. Before the epidemic was proclaimed as ceased in May 1998, a total of 429 infected farms were detected and approximately 700,000 pigs from these farms were slaughtered. Figure 3.7(a) shows the number of newly infected herds per week as a function of time. During the progress of the epidemic, several control strategies were applied to dike the outbreaks. Among them *pre-emptive slaughtering*, i.e. slaughtering of herds who have been in contact with or are located within close proximity of an infected herd. In the CSF case, a total of 1286 herds (approximately 1.1 million pigs) were pre-emptively slaughtered - only 44 of the herds could later be diagnosed as infected - either by clinical signs or by blood samples taken just before depopulation. This makes pre-emptive slaughtering a highly controversial measure due to the killing of many apparently healthy animals. Both Elbers *et al.* [1999]; Stegeman *et al.* [1999] although argue that pre-emptive slaughtering is a necessary measure: It was first when this measure was applied that the epidemic started to cease! They argue that signs of CSF take time to develop — some of the terminated herds would ultimately have proven as infected had no interaction occured.



(a)

(b)

Figure 3.7: (a) Number of newly infected herds per week for each week after the occurrence of the first case in week 51 1996 [Stegeman *et al.*, 1999] (b) Distribution of the interval (number of days) between CSF laboratory diagnosis and depopulation. Negative numbers originate from some cases detected through pre-emptive slaughtering. Taken from [Elbers *et al.*, 1999, p.173].

If an analysis could show that pre-emptive slaughtering – or other measurements – have an actual effect, this would be a great advantage when deciding, which measures to apply next time. To investigate on the subject, the strategies applied by the authorities in the detection and fight against the CSF outbreak are divided into five consecutive phases, each implemented from a known date [Stegeman *et al.*, 1999].

Phase 1 Before detecting the first case. Obligation to report clinical suspicion of CSF and investigation of tonsils from pigs sent in for routine post-mortem examinations.

Phase 2 After the first case, EU required measures came into force: Stamping out infected herds, establishing of surveillance zones around infected herds, no transportation of pigs allowed within zone, and all herds within zone are inspected clinically once a week. Same applied to all herds known to have been in contact with the infected herd.

Phase 3 Added pre-emptive slaughtering of all herds in contact with infected herds or within a 1km radius. Due to logistical problem it could take up to 4 weeks between detection of infected an herd and the pre-emptive slaughtering of the neighboring herds.

Phase 4 Interval between detection of infection and pre-emptive slaughtering brought down to a maximum of seven days. Breeding ban and killing of 3-17 day old piglets instead of transporting them.

Phase 5 Included monthly serological survey for enclosed areas to detect sub-clinical cases. Compartmentalized transportation, i.e. trucks were only allowed to travel in specific zones.

It is now of interest to see, what type of effect the different measurements had on the course of the epidemic. To do this Stegeman *et al.* [1999] used detailed knowledge about each case, to determine for each week i the number of newly infected C_i, number of susceptible S_i, and number of infectious I_i. This allows calculating β_i from the expression

$$C_i = \frac{\beta_i}{N} I_i S_i,$$

which is a discretized version of the SIR-model. The population size N is set equal to 21,500 — the total number of pig herds in the Netherlands. In other words, each β_i is seen as independent from the others given information about C_i. Forming averages for each phase and testing for equality brings them to the conclusion that the periods are significantly different and that the basic reproduction ratio $R_0 = \beta N/\gamma = (6.76, 1.33, 0.91, 0.53, 0.59)$ is below one in phases three to five [Stegeman *et al.*, 1999]. It is not possible to argue from the numbers, whether the measures had an effect, due to lack of comparison opportunities, but it is clear that introduction of new measures and reduction in R_0 coincides.

Instead of the individual weekly estimation of parameters, we will estimate parameters directly in a switched SIR model. Within the context of a SIR-epidemic, we let the pig herd be the unit of modeling and a removal time corresponds to the day, where the herd was depopulated. Unfortunately, the exact time of occurrence is not available to us, instead data from [Stegeman *et al.*, 1999, Table 1] specifying the number of newly infected herds for each week shown in Figure 3.7(a), is used. Based on Figure 3.7(b) we claim that depopulation approximately happens within the same week.

That is, a number of 13 new infections for week 97-06 is thus translated to 13 recoveries that week, where the exact time within the week is found by sampling 13 numbers uniformly within the week. This is not as good as knowing the exact depopulation time, but feels sufficient for our purposes. Compared to the total length of the epidemic, small unreliabilities on recovery time should not be of importance. Based on 5,000 samples obtained by taking each 100th sample after a burn-in of 100,000 samples yields the following results.

	Mean	SD	Naive SE	MC Error	Batch SE	Batch ACF
β_1	0.79492	0.216765	3.066e-03	9.253e-03	0.0121006	0.1943
β_2	0.62654	0.194427	2.750e-03	1.073e-02	0.0149983	0.5089
β_3	0.38605	0.128587	1.818e-03	7.048e-03	0.0103225	0.5077
β_4	0.06816	0.070617	9.987e-04	3.981e-03	0.0057916	0.5594
β_5	0.08776	0.074351	1.051e-03	2.491e-03	0.0030370	0.1903
γ_1	0.56086	0.193367	2.735e-03	8.566e-03	0.0114645	0.3291
γ_2	0.53966	0.177245	2.507e-03	9.746e-03	0.0137178	0.5136
γ_3	0.39699	0.110225	1.559e-03	5.980e-03	0.0087810	0.5493
γ_4	0.22770	0.065201	9.221e-04	3.194e-03	0.0045453	0.5346
γ_5	0.19200	0.079728	1.128e-03	2.205e-03	0.0027275	0.1058

The total of 600,000 samples takes 1143 seconds to complete, with the acceptance rate for $\boldsymbol{I}$ being 32.10%. Trace plots for two selected β's are found in Figure 3.8. Especially β_4 looks problematic, but the variations might be explained by the small number of infection residing in that particular regime resulting in great variations, when the infection times change. For the other parameters the trace plots look fine.

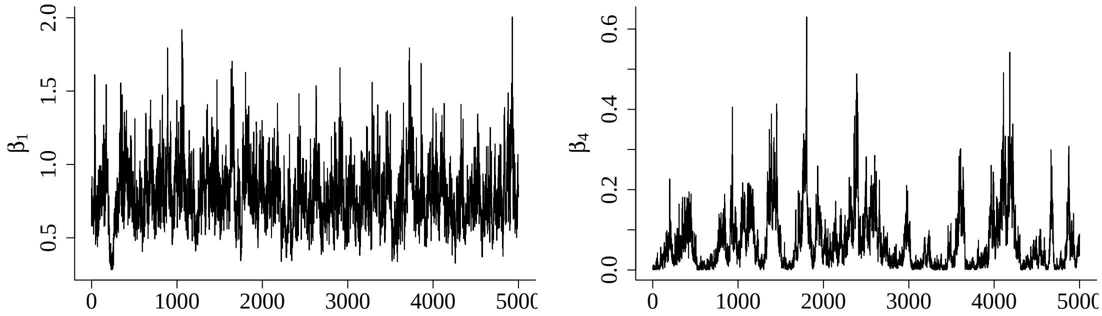


Figure 3.8: Trace plots for β_1 and β_4 for the CSF outbreak.

Based on the above estimates, the basic reproduction ratio for the 5 phases are 1.42, 1.16, 0.97, 0.30, 0.46. Compared to the numbers from Stegeman *et al.* [1999] especially our first R_0 is substantially lower, the others are comparable. Furthermore, our results agree on whether $R_0 > 1$ for a phase or not. Again, we can only conclude that there is a certain correlation between different measures and R_0 reductions. Investigating, whether using a five regime model, M_5, is reasonable for the CSF data, we tested it against a model with three regimes (i.e. using only the first two switch points of M_5) and one with only one regime. Testing was done using Bayes Factors based on 1000 samples (Thinning 100, 10,000 Burn-in samples) we obtained

$$\hat{P}_2(D|M_5) = 1.08 \cdot 10^{39} \cdot k_2, \hat{P}_2(D|M_3) = 2.23 \cdot 10^{27} \cdot k_2, \hat{P}_2(D|M_1) = 9.62 \cdot 10^{11} \cdot k_2,$$

with $k_2 = \exp(1222)$. This means that the data provides strong evidence of M_5 against both M_3 and M_1, respectively.

To get convincing ideas about effects an analysis should not work with a fixed switch time occurrence, but instead have $\boldsymbol{S}$ as a parameter as well. That way, both number of regimes, k, and location of their switch-points are to to be determined, e.g. by a reversible-jump MCMC algorithm [Waagepetersen and Sorensen, 2001]. Only if the estimated $\boldsymbol{S}$ is in concordance with the known measure-implementational time, an effect could be claimed. A further extension might be to improve the discrete-time to continuous-time-data transformation scheme. Discrete time observations providing the number of recoveries happening on e.g. daily basis is a quite common situation when collecting epidemic data. Letting the exact recovery times be parameters, subordinate to the constraint that the number occuring within a specified interval corresponds to the discrete time observation, one could handle the situation at the expense of even more unknowns.

Chapter 4

Conclusions and Future Work

This report dealt with parameter estimation in various settings of the continuous time stochastic SIR-model – a process oriented model to describe infectious diseases. Aim has been to extend the basic model to deal with the domain of animal production. This implied handling inhomogeneity in contact introduced by stalling and handling interaction into the course of the epidemic by control measures. Central point of the tale has been the adaptation of the models to actual observations.

In our context this meant that parameter estimation had to be based on just recovery times. Modeling interactions into the epidemic through the concept of regime switching proved via MCMC to be relatively easy, even for the case of partial observable epidemics. Here, a Gibbs updating scheme could be used for the parameters of the epidemic. Introducing spatially arranged populations, on the other hand, required more work. Gibbs updating was not possible for the transmission rates, for which reason Metropolis-Hastings steps were introduced. In general the estimation worked well for non-extensive epidemics, i.e. epidemics where not all susceptibles got infected. While investigating a case of an all-embracing epidemic we although experienced convergence problems. Future investigations have to reveal, whether this is common feature for all-embracing epidemics. Also, a connection could exist to the assumption about observing the epidemic to its end. If the number of recoveries equals the number of susceptible the epidemic definitely ceases at the last recovery time. On the other hand, if there still are susceptible left at the last recovery, declaring the epidemic as terminated means we either waited for a while with nothing happening or that we simply assumed nothing more is happening. Extending the framework to handle non-terminated epidemics as in [O'Neill and Roberts, 1999] would provide valuable insights.

Distinguishing between within and between section transmissions is not always easy in a spatial setup, which was reflected by the parameter estimates obtained. The possibility for an infection to originate from several places brings along fewer constraints on the location of the infection times. A trick would be to augment each infection with an additional parameter indicating the origin. This would make the likelihood expression nicer enabling Gibbs updating for the transmission parameters at the expense of more parameters.

4.1 Application

In order to use an extended SIR-model as core of an on-site disease prediction system operation in a pig production facility, more thoughts on how to gather the steps of estimation and prediction are required. Initially, the prediction problem corresponds to estimation in an partially observed epidemic not monitored till end: we do not know the exact number of infections occured, but can use the already seen recoveries to estimate the parameters with some certainty. Interest is although not in the parameter values; simultaneously we would like to find distributions for future values of infected and recovered and report these predictions to the user in an intuitive manner [Höhle, 2002].

Application of the spatial extension is highly relevant to many other already existing data from animal production, e.g. disease transmission experiments conducted at the Royal Veterinary and Agricultural University, Copenhagen, Denmark. Two containers, each hosting a specified number of pigs, were connected using a controllable ventilation system. One or more pigs in the first container were inoculated with pneumonia strains after which the course of the epidemic is observed in order to quantify the air-borne spread of the disease. Information was collected by testing weekly blood samples for zero-conversion, i.e. presence of anti-bodies against the disease. This common data situation unfortunately does not immediately fit into our estimation framework, because anti-bodies are usually found a – individually varying – number of weeks after infection. Using some distributional assumption about time between infection and creation of antibodies we would be able to deduce information about infection times instead of recovery times. For non-lethal diseases, diagnosed pigs are not culled and hence remain infectious throughout the entire experiment, not providing any informative recovery times at all.

Because monitoring for zero-conversion is the preferred way to perform experiments of disease transmission between individuals in pig production, accommodation of the estimation procedure would open the doors to a large number of interesting data sets [Vraa-Andersen, 1994; Stärk *et al.*, 2000]. A recurrent feature of these (and many other) data sets is although that recordings are performed on a discrete time scale, e.g. daily or weekly. In our opinion a continuous time model better captures the dynamics of the epidemic and hence provides more explanatory power. But, depending on the application, it might be beneficial to take the discrete time nature of the data into account when estimating. For example by letting the appropriate event times be parametric but subject to the constraint that the number of events in each discrete time interval corresponds to the data.

Adding an additional exposed phase to the model to cope for the non-negligible latency time of the disease would probably give a better fit to the data and correspond better to the physiological knowledge available about the disease. Also, more realistic distributional assumptions regarding length of latency and infection time as in [O'Neill and Becker, 2001] would enhance realism and thus the integration of prior knowledge. Such model extensions definitely intensify a demand to perform model-comparisons. Bayes factors seem an obvious choice, but better approaches to compute them than ours seem necessary, e.g. the method by Carlin and Chib or reversible-jump MCMC [Han and Carlin, 2000].

Another hot topic of application could be the 2001 UK food and mouth outbreak [Ferguson *et al.*, 2001; Keeling *et al.*, 2001]. Operating on e.g. county level it might be interesting to quantify spread according to airborne transmission. Taking contact patterns when exchanging and transporting animals into account a nearest neighbor approach might although not be sufficient, here introduction of an additional background risk or an entire contact hierarchy could perhaps improve the realism.

4.2 Implementation

Implementation of MCMC methods is a tedious task – especially debugging is cumbersome, because all results are a mixture of possible "systematic" (read: programming) error and random error. Tiny programming errors can lead to convergence problems, where one does not know, whether this lack is due to methodology, too short chains or due to a simple bug. As a help to others, the software together with the simulated datasets is made available for download[1]. It has been tested by confirming the results of [O'Neill and Roberts, 1999] and by verifying it against simulated data. By adopting an open-source policy with respect to our code we do not only invite others to confirm our results, it also forces us as developers to pay regard to decent documentation. We are very interested in receiving feedback about its use should anyone decide to give it a try.

4.3 Summing up

Formulation of process models and their fitting to actual epidemics is definitely a valuable tool to gain insight into disease dynamics. Selecting a useful model and collecting good data is certainly not trivial. Even done with success, it might not save your computer from the newest virus or reduce health management on the farm to pure routine, but at least we end up knowing a little more on the how and why. Scary tales on the other hand, will nevertheless still wait for you on bookshelves or in the media.

[1] Visit `http://www.dina.kvl.dk/~hoehle/software/sir/`

Appendix A

Implementation

The SIR estimation routines are gathered in the program `LadyBug` implemented in Java JDK 1.3, see [Sun, 2002]. Class files, source, and documentation is is available for download from the url

`http://www.dina.kvl.dk/~hoehle/software/sir/`

Class documentation is available through the JavaDoc generated HTML documentation. To invoke the `LadyBug` program from the prompt three files are required: a recovery file specifying the environment of the epidemic together with observed recovery times, an optional file containing infectious time, and an options file contains information about the estimation method to use.

A.1 The recovery and infection time files

The recovery file contains a $(3 + l)$ line header, where the dimension of the grid formed by the spatial layout is given by $l_1 \cdot l_2 = l$. The header specifies the environment of the epidemic, i.e. unit layout and possible regime switches, using the following format.

`switch` $s_{12} \quad \ldots \quad s_{k-1k}$

`unit` $1\ 1\ \ N_{1,1}\ \ a_{1,1}\ \ \nu_{\beta_1^{1,1}}\ \lambda_{\beta_1^{1,1}}\ \nu_{\gamma_1^{1,1}}\ \lambda_{\gamma_1^{1,1}} \quad \ldots \quad \nu_{\beta_k^{1,1}}\ \lambda_{\beta_k^{1,1}}\ \nu_{\gamma_k^{1,1}}\ \lambda_{\gamma_k^{1,1}}$

$\vdots$

`unit` $l_1\ l_2\ \ N_{l_1,l_2}\ \ a_{l_1,l_2}\ \ \nu_{\beta_1^{l_1,l_2}}\ \lambda_{\beta_1^{l_1,l_2}}\ \nu_{\gamma_1^{l_1,l_2}}\ \lambda_{\gamma_1^{l_1,l_2}} \quad \ldots \quad \nu_{\beta_k^{l_1,l_2}}\ \lambda_{\beta_k^{l_1,l_2}}\ \nu_{\gamma_k^{l_1,l_2}}\ \lambda_{\gamma_k^{l_1,l_2}}$

`betan` $\nu_{\beta_1^n}\ \lambda_{\beta_1^n} \quad \ldots \quad \nu_{\beta_k^n}\ \lambda_{\beta_k^n}$

`theta` θ

If there is only one regime use `switch NA` as the first line. Currently, `LadyBug` only supports models, where all neighbor effects are equal, i.e. $\forall i, j : i \neq j : \beta^{i \to j} \equiv \beta^n$. The prior for β^n is specified in the `betan` line – in case of only one unit, the specified values are of course ignored. Following the header, the recovery times are provided one per line together with the unit they occur in.

The above syntax is definitely cumbersome for larger spatial specifications, but as `LadyBug` currently only handles moderate sized layouts in tractable time the syntax is sufficient. Future versions should allow for a direct symbolic specification of the $\boldsymbol{\beta}$-matrix, e.g in order to specify that infectious material can travel from unit one to two, but that the opposite is impossible. Furthermore, the current specification requires specification of priors — even for non MCMC analysis. It might therefore be beneficial to make prior specification part of the MCMC method in the options file instead.

Using the 5-case epidemic described in Section 2.4.1 as an example, the `oneill.rec` looks as follows.

```
switch NA
unit 1 1 9 1 0.1 1 0.1 0.1
betan 0 0 0 0
theta 0
1     1       0
1     1       1.52292
1     1       1.55004
1     1       1.93064
1     1       2.67492
```

Here, the recovery times are sorted according to time and written one per line together with the unit the event occured in. A similar approach is used to specify infection times in the infection-file. No header is necessary here, infection times are just listed in time wise order one per line together with the unit they occured it.

A.2 The options file

Using the option file, the user can control the estimation method to use together with its various parameters. Specification of this happens according to the following grammar.

```
OPTIONSFILE   ←  OPTIONS METHOD
OPTIONS       ←  (class seed=SEED timescale=float)
SEED          ←  NA | int
METHOD        ←  (method METHODTYPE )
METHODTYPE    ←  ml
                 | mml samples=int mcsamples=int sumscale=double
                 | mcmc samples=int thin=int burnin=int betaRWsigma=double
                   betaNRWsigma=double hmeanscale=double
```

The `options` class allows fixing the seed value of the random generator, this might be beneficial in debugging situations, where one wants to ensure operating on the same numbers. Also by setting `timescale` to a constant k_1 all recovery and infection times are divided by k_1, which could be due to modeling reasons or because it is numerically practical. Note although that the obtained parameter estimates constitute k_1 times the non-scaled estimates.

In case of an MCMC analysis one has to specify the number of samples to base the estimation on. These samples will be generated by taking each `thin`'th sample after an initial `burnin` samples. Further control options are the standard deviation of the gaussian proposal distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^n$ when these are updated using the Metropolis sampler, i.e. in case of $l > 1$. The `hmeanscale=`k_2 options allows intermediate scaling when computing $\hat{P}_2$. All contributions are then divided by $\exp(k_2)$, which might be necessary to avoid numeric overflow.

For a maximum marginalized likelihood analysis, one can control the number of samples used in the Monte Carlo approximation of the integral. In some situations it might be beneficial to get an idea of the effect of this Monte Carlo approximation. It is therefore possible to let `LadyBug` repeat the entire estimation procedure with new samples `samples` number of times. Monitoring the optimizer's convergence towards the maximum of the marginalized likelihood can be done by `veroboseOptimizer`.

Summarizing the above in an example, the option file, `oneill.sir`, of a MCMC estimation for the 5-case epidemic from Section 2.4.1 looks as follows.

```
//                              \./
//                             (o o)
//-------------------------oOOo-(_)-oOOo-------------------------
// Author:      Michael Höhle <hoehle@dina.kvl.dk>
// Description: Use MCMC to estimate parameters in the 5-case epidemic
//              presented in the O'Neill article.
//

(options
        seed=1999       //Fix seed value, so we can restore results.
        timescale=1.0 //No need to scale time.
)

(method mcmc
        samples=5000
        thin=100
        burnin=1000
        betaRWsigma=0.05   //not really used in a one unit setup
        betaNRWsigma=0.025 //not used in a one unit setup.
        hmeanscale=0       //exp(0)=1, i.e. no scaling necessary.
)
```

Invoking the analysis by the corresponding call to the Java engine yields the following program output.

```
$ java sir.estimate.LadyBug ../data/oneill.rec NA ../data/oneill.sir
...
I acceptance rate   : 36.74291417%
----- Parameter Estimates = (E,V) of posterior  --------
beta[1][1][1] = (0.09631495 , 0.00348988)
gamma[1][1][1]= (0.79477083 , 0.21700828)
\hat{P}_2(D|M) = 5.8253572170989247E-8 * exp(0.0)

Job done...used 28 s.
```

The three prompt parameters to the `LadyBug` class are the filenames of the recovery, infection, and option file, respectively. In case infection times are not available as in the above example, the string `NA` is used to indicate this. In the program output, the I-acceptance rate is the percentage of the proposed infection times in the Metropolis sampler that is accepted. For each parameter in the model, e.g. `beta[x][y][r]`=$\beta_r^{x,y}$, the mean and variance of the posterior distribution samples is reported. Finally, $\hat{P}_2$, i.e. the harmonic mean of the likelihood values calculated on the basis of the posterior samples is reported. For further analysis such as convergence checks with BOA [B. , 2001], etc. a log-file `log.txt` contains all parameter samples together with acceptance rate and the log likelihood of the sample.

A.3 Miscellaneous

Part of the program, implementing the MCMC estimation, takes advantage of the HYDRA package [Warnes, 2001]. Unfortunately, the Gibbs-within-Metropolis sampling scheme is not directly supported by the package. Only the basic structure of a MCMC chain, together with its `generate` method could be used from the package together with some modified listeners allowing for burnin and filtering. In our opinion the HYDRA package can be of great help for basic MCMC, but for our purposes the lack of file structure and documentation complicated its use. Most of the `LadyBug` program is therefore own code. An exception is the nonlinear numerical optimizer, which is integrated from the "Nonlinear Optimization Java Package", see Verrill [1998].

Besides the estimation program, a program was developed, which is able generate appropriate datasets for estimation by simulation from the extended SIR-models. Simulatation from continuous Markov chains follows the approach of [Hillier and Lieberman, 1995, Chapter 14].

Bibliography

P.K. Andersen, Ø. Borgan, R.D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, 1993.

H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*. Number 151 in Lecture notes in statistics. Springer, 2000.

B. Smith, Ph.D., Department of Biostatistics, The University of Iowa College of Public Health. *Bayesian Output Analysis Program*, 2001. `http://www.public-health.uiowa.edu/boa/`.

N. G. Becker. *Analysis of Infectious Disease Data*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1989.

T. Britton. Estimation in multitype epidemics. *Journal of the Royal Statistical Society, Series B*, 60(4):663–679, 1998.

The BUGS project. *CODA – Convergence Diagnosis and Output Analysis Software for Gibbs sampling output*, 2001. `http://www.mrc-bsu.cam.ac.uk/bugs/`.

A.D. Cliff and P. Haggett. Statistical modelling of measles and influenza outbreaks. *Statistical Methods in Medical Research*, 2:42–73, 1993.

M. H. DeGroot. *Probability and Statistics*. Addison-Wesley, Reading, Massachusetts, 2nd edition, 1989. Reprinted with corrections.

O. Diekmann and J. H. Heesterbeek. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Wiley Series in Mathmatical and Computational Biology. John Wiley & Son, Ltd., 2000.

A.R.W. Elbers, A. Stegeman, H. Moser, H.M. Ekker, J.A. Smak, and F.H. Pluimers and. The classical swine fever epidemic 1997-1998 in the Netherlands: descriptive epidemiology. *Preventive Veterinary Medicine*, 42(3–4):157–184, 1999.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, 1994.

N. M. Ferguson, C. A. Donnelly, and R. M. Anderson. The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science*, 292:1155–1160, 11th May 2001.

G. J. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics applied in Medicine & Biology*, 15:19–40, 1998.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

C. Han and B. Carlin. MCMC methods for computing Bayes factors: a comparative review. *Biometrika*, 82(4):711–732, 2000.

Y. Hayakawa, D. Upton, P.S.F. Yip, and P.D. O'Neill. Inference for a multitype epidemic model using Markov chain monte carlo methods. Technical Report 00-03, Statistics Division, School of Mathematical Sciences, University of Nottingham, 2000. Available from `http://www.maths.nottingham.ac.uk/statsdiv/research/reports.html`.

M. Höhle. Operational management of pneumonia in slaughter pig productions. Unpublished Ph.D. Exercise Report, 2001.

M. Höhle. Decision support for pneumonia management in pig production. In *Proceedings of the 1st European Workshop on Sequential Decisions under Uncertainty in Agriculture and Natural Resources*, pages 51–56, september 2002.

F.S. Hillier and G.J. Lieberman. *Introduction to Operations Research*. Industrial Engineering Series. McGraw-Hill International Editions, sixth edition edition, 1995.

J. Holst and H. Madsen. Lecture notes in advanced time series analysis. Handouts · IMM · DTU, December 2000.

M. J. Keeling, M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell. Dynamics of the 2001 uk foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, 294:813–817, 2001.

E. C. Marshall, A. Frigessi, N. C. Stenseth, M. Holden, V. S. Ageyev, and N. L. Klassovskiy. Plague in kazakhstan: a Bayesian model for the temporal dynamics of a vector-transmitted infectious disease. *Journal of the American Statistical Association*, 2001. Under revision – not published yet.

P.D. O'Neill and N.G. Becker. Inference for an epidemic when susceptibility varies. *Biostatistics*, 2(1):99–108, 2001.

P. D. O'Neill and G. O. Roberts. Bayesian inference for partially observed stochastic epidemics. *J. R. Statist. Soc.*, 162:121–129, 1999.

F.H Pluimers, P.W. de Leeuw, J.A. Smak, A.R.W. Elbers, and J.A. Stegeman. Classical swine fever in the Netherlands 1997-1998: a description of organisation and measures to eradicate the disease. *Preventive Veterinary Medicine*, 42(3–4):139–155, 1999.

W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, New York, NY, 1992.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society B*, 64(3):1–34, 2002.

A. Stegeman, A.R.W Elbers, A. Bouma, H. de Smit, and M. C. M. de Jong. Transmission of classical swine fever virus within herds during the 1997-1998 epidemic in the Netherlands. *Preventive Veterinary Medicine*, 42(3–4):201–218, 1999.

K.D.C. Stärk, D.U. Pfeiffer, and R.S. Morris. Within-farm spread of classical swine fever virus - a blueprint for a stochastic simulation model. *Veterinary Quarterly*, 22(1):36–43, jan 2000.

Sun Microsystems, Inc. *Java 2 Platform - Standard Edition*, 2002.

S. Verrill. *Nonlinear Optimization Java Package*. Forest Products Laboratory, USDA Forest Service, 1998. Available from `http://www1.fpl.fs.fed.us/optimization.html`.

L. Vraa-Andersen. *Respiratory Diseases in Danish Slaughterwine – Application of Epidemiological methods*. PhD thesis, Division of Ethology and Health, Department of Animal Science and Health, Royal Vertinary and Agricultural University, Frederiksberg, Denmark, 1994.

R. Waagepetersen and D. Sorensen. A tutorial on reversible jump MCMC with a view toward QTL-mapping. *International Statistical Review*, 69:49–61, 2001.

G.R. Warnes. HYDRA: A Java Library for Markov chain Monte Carlo. Technical Report 394, University of Washington, Department of Statistics, April 2001.