Bayesian methods to infer direct transmission using data from outbreaks in households

Solomon Christopher, M.Sc

Licentiate Thesis

Department of Mathematics and Statistics
University of Helsinki

June 3, 2020

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Faculty of Science	Koulutusohjelma – Utbildningsprogram – Degree Programme Licentiate (Doctoral Programme in Mathematics and Statistics - Domast)

Opintosuunta – Studieinriktning – Study Track
Biometry / Statistics

Tekijä – Författare – Author
Solomon Christopher

Työn nimi – Arbetets titel – Title
Bayesian methods to infer direct transmission using data from outbreaks in households

Työn laji – Arbetets art – Level Monograph	Aika – Datum – Month and year June, 2020	Sivumäärä– Sidoantal – Number of pages 75

Tiivistelmä – Referat – Abstract

The study of how transmissible an infectious pathogen is and what its main routes of transmission are is key towards management and control of its spread. Some infections which begin with zoonotic or common-source transmission may additionally exhibit potential for direct person-to-person transmission. Methods to discern multiple transmission routes from observed outbreak datasets are thus essential. Features such as partial observation of the outbreak can make such inferences more challenging.

This thesis presents a stochastic modelling framework to infer person-to-person transmission using data observed from a completed outbreak in a population of households. The model is specified hierarchically for the processes of transmission and observation. The transmission model specifies the process of acquiring infection from either the environment or infectious household members. This model is governed by two parameters, one for each source of transmission. While in continuous time they are characterised by transmission hazards, in discrete time they are characterised by escape probabilities.

The observation model specifies the process of observation of outbreak based on symptom times and serological test results. The observation design is extended to address an ongoing outbreak with censored observation as well as to case-ascertained sampling where households are sampled based on index cases. The model and observation settings are motivated by the typical data from Hepatitis A virus (HAV) outbreaks.

Partial observation of the infectious process is due to unobserved infection times, presence of asymptomatic infections and not-fully-sensitive serological test results. Individual-level latent variables are introduced in order to account for partial observation of the process. A data augmented Markov chain Monte Carlo (DA-MCMC) algorithm to estimate the transmission parameters by simultaneously sampling the latent variables is developed. A model comparison using deviance-information criteria (DIC) is formulated to test the presence of direct transmission, which is the primary aim in this thesis. In calculating DIC, the required computations utilise the DA-MCMC algorithm developed for the estimation procedures. \\

The inference methods are tested using simulated outbreak data based on a set of scenarios defined by varying the following: presence of direct transmission, sensitivity and specificity for observation of symptoms, values of the transmission parameters and household size distribution. Simulations are also used for understanding patterns in the distribution of household final sizes by varying the values of the transmission parameters.

From the results using simulated outbreaks, DIC$_6$ consistently indicates towards the correct model in almost all simulation scenarios and is robust across all the presented simulation scenarios. Also, the posterior estimates of the transmission parameters using DA-MCMC are fairly consistent with the values used in the simulation.

The procedures presented in this thesis are for SEIR epidemic models wherein the latent period is shorter than the incubation period along with presence of asymptomatic infections. These procedures can be directly adapted to infections with similar or simpler natural history. The modelling framework is flexible and can be further extended to include components for vaccination and pathogen genetic sequence data.

Avainsanat – Nyckelord – Keywords
Outbreak, household, direct transmission, final size, infectious disease model, Bayesian inference, Markov chain Monte Carlo, deviance information criteria, simulation

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Abstract

The study of how transmissible an infectious pathogen is and what its main routes of transmission are is key towards management and control of its spread. Some infections which begin with zoonotic or common-source transmission may additionally exhibit potential for direct person-to-person transmission. Methods to discern multiple transmission routes from observed outbreak datasets are thus essential. Features such as partial observation of the outbreak can make such inferences more challenging.

This thesis presents a stochastic modelling framework to infer person-to-person transmission using data observed from a completed outbreak in a population of households. The model is specified hierarchically for the processes of transmission and observation. The transmission model specifies the process of acquiring infection from either the environment or infectious household members. This model is governed by two parameters, one for each source of transmission. While in continuous time they are characterised by transmission hazards, in discrete time they are characterised by escape probabilities.

The observation model specifies the process of observation of outbreak based on symptom times and serological test results. The observation design is extended to address an ongoing outbreak with censored observation as well as to case-ascertained sampling where households are sampled based on index cases. The model and observation settings are motivated by the typical data from Hepatitis A virus (HAV) outbreaks.

Partial observation of the infectious process is due to unobserved infection times, presence of asymptomatic infections and not-fully-sensitive serological test results. Individual-level latent variables are introduced in order to account for partial observation of the process. A data augmented Markov chain Monte Carlo (DA-MCMC) algorithm to estimate the transmission parameters by simultaneously sampling the latent variables is developed. A model comparison using deviance-information criteria (DIC) is formulated to test the presence of direct transmission, which is the primary aim in this thesis. In calculating DIC, the required computations utilise the DA-MCMC algorithm developed for the estimation procedures.

The inference methods are tested using simulated outbreak data based on a set of scenarios defined by varying the following: presence of direct transmission, sensitivity and specificity for observation of symptoms, values of the transmission parameters and household size distribution. Simulations are also used for understanding patterns in the distribution of household final sizes by varying the values of the transmission parameters.

From the results using simulated outbreaks, DIC_6 consistently indicates towards the correct model in almost all simulation scenarios and is robust across all the presented simulation scenarios. Also, the posterior estimates of the transmission parameters using DA-MCMC are fairly consistent with the values used in the simulation.

The procedures presented in this thesis are for SEIR epidemic models wherein the latent period is shorter than the incubation period along with presence of asymptomatic infections. These procedures can be directly adapted to infections with similar or simpler natural history. The modelling framework is flexible and can be further extended to include components for vaccination and pathogen genetic sequence data.

Acknowledgements

I express my sincere gratitude to my supervisor Professor Kari Auranen from whom I have learned the most about scientific research and writing. I thank him for his diligent and patient guidance through each stage of this research work and for making a profound impact on my thinking in infectious disease modelling and statistics. I also thank him for providing me the opportunity to be a teaching assistant in his course *Statistical methods in medicine and epidemiology* at University of Helsinki.

I am deeply indebted to my second supvervisor Professor Sangita Kulathinal, without whose support, this research would have never seen light of day. I thank her for her continued guidance and support during this research and for her critical ideas that made this research more precise and rigorous. I also thank her for giving me opportunity to present this work in her courses related to event history modelling.

I am grateful to Associate Professors Matti Vihola and Yang Yang for accepting the invitation to act as examiners for this thesis.

I thank the Finnish Institute for Health and Welfare (THL) for having me as a visiting researcher throughout the course of this research. I especially enjoyed the research environment there and the good company of Bijoy, Hanna R-K, Jukka, Markku, Mikhail, Sanna and Ulrike.

I thank the Centre for International Mobility (formerly CIMO, now EDUFI) for the scholarship recieved during the first year which enabled to be a visiting student and researcher at the University of Helsinki.

I thank my colleagues at Medialis Ltd., (where I am currently working since 2018) for creating a good research environment. Special thanks are due to Dr. Ravi Jandhyala who sincerely cares for and has a genuine interest towards my academic research and training.

I also thank my former colleagues at EpidResearch/StatFinn (now part of IQVIA) where I worked as a statistician from 2012-2018. I especially thank Dr. Pasi Korhonen and Dr. Fabian Hoti for giving me the opportunity to work in mainstream biostatistics and pharmaco-epidemiology and also for providing flexible work schedule in the early years of my employment there to complete my required course-work.

Thanks to Mikhail Yuryev who encouraged me to complete this thesis by reminding me about it in interesting ways. But most of all, thank you for the friendship, our adventures together and memories to cherish.

I thank my friends Ahmed, Lena, Andrea and Maija for many years of friendship, roadtrips and conversations. Thank you to Markus, Satu, Sampo, Katka, Zoe, Darina, Kati and Sini for your friendship and support, you made my life brighter. Special thanks to Hannu for the awesome coffee and writing company at Kumpula in the past year.

I thank my parents for their continued love, support, encouragement and patience during the years of this research. I dedicate this work to them.

Solomon Christopher
June, 2020

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"Indeed, what is history but an interpretation of time?"
- Lindsey JK (2004)

Analysing data from infectious disease outbreaks poses two intrinsic challenges: (i) transmissibility of infection produces dependent observations and (ii) partial observation of the infection process often resulting in the size of missing data surpassing that of observed data (Kypraios & Minin, 2018). These challenges in the data can be accounted for, in part, by using mechanistic models to explicitly model dependencies and inference procedures to deal with latent state models.

As opposed to empirical models that directly address uncertainty in the observations, mechanistic models describe the underlying processes that generate the data. Mechanistic models explicitly specify the following: dependence in observations, dependence between observed and missing data, and the model in terms of epidemiologically meaningful parameters.

This chapter briefly presents the general framework of infectious disease modelling along with some associated terminology. Some modelling aspects of outbreaks in households are provided as a background for the methods presented in this thesis. Some selected literature on direct (or person-to-person) transmission is discussed. As the inference methods in the thesis are developed under the Bayesian framework, some relevant Bayesian ideas are presented. The chapter ends with summarising the scope and structure of the thesis.

1.1 Stochastic modelling of infectious diseases

1.1.1 Infectious disease modelling

The natural history of the infection in question defines the relevant compartments (infection states) through which an infected individual progresses. An SIR compartmental model indicates that individuals progress through susceptible (S), infected (and infectious) (I) and removed (R) states. Removal may refer to the individual not contributing to the infection risk process due to recovery from infection, quarantine or death. The model is specified for transitions between the states using a mechanistic formulation.

The contact structure between susceptible and infectious individuals along with contact rates and the probability of infection given such a contact defines the rate of transition $S \rightarrow I$. For a homogeneously mixing population, the transmission rate (β) encapsulates the per-capita contact rate and the chance of infection, given a contact between susceptible and infectious individuals. The removal rate (γ) specifies the rate of transition $I \rightarrow R$. Deterministic models decribe these transitions using rates of change in the compartment size (frequency or density of individuals in a state) in the form of a system of differential equations.

Models for infections with differing natural history require relevant modifications. Additional compartments are added when more infection states are present such as SEIR models where the infected but not yet infectious

(E) state indicates that individuals do not become infectious immediately after infection. An SIS model is used when a recovered individual is immediately at risk (susceptible) to infection.

Stochastic models describe the probabilities (or intensities) for change in the size of compartments over time. The size of compartments define a discrete state-space and the transitions in the state-space is defined using counting processes (Andersen et al., 1993). In stochastic epidemic models, the counting processes are parameterised using mechanistic formulations.

One of the most explored epidemic models is the general stochastic epidemic model (GSEM). This is an SIR model where the transitions in the state-space $\{S, I\}$ over the epidemic duration $[0, T]$, $\{(S(t), I(t)), t \in [0, T]\}$, are Markovian in nature. Here, $\{(S(t), I(t))\}$ are the numbers of individuals in the model states S and I, respectively, at time t. Following Knock & Kypraios (2018), the conditional probabilities that infection and removal events occur during the time interval $[t, t + \delta t)$ given all observations up to t (history, $\mathcal{H}_t$) are

$$\begin{cases} P[S(t+\delta t) - S(t) = -1, I(t+\delta t) - I(t) = 1 | \mathcal{H}_t] = \beta S(t) I(t) \delta t + o(\delta t) \\ \\ P[S(t+\delta t) - S(t) = 0, I(t+\delta t) - I(t) = -1 | \mathcal{H}_t] = \gamma I(t) \delta t + o(\delta t). \end{cases} \tag{1.1}$$

The literature for stochastic modelling of infectious disease falls into two broad categories: (i) stochastic models characterising epidemic behaviour along with their analytical results and (ii) statistical inference from infectious disease data using stochastic epidemic models.

1.1.2 Stochastic models to study epidemic behaviour

Stochastic models are useful for understanding epidemic behaviour: insights can be obtained by analytical solutions when possible and by use of simulations. For example, for an epidemic model with given state transitions, the fraction (or number) of individuals infected at the end of outbreak (final size) is of interest. The distribution of final size and the associated probability of a major outbreak in the SIR model is described in Britton (2010).

A quantity of fundamental importance in infectious disease epidemiology is the basic reproduction number (R_0), defined as *"the expected number of secondary infections that result from a single infectious individual in an entirely susceptible population"* (Grassly & Fraser, 2008). A key result in infectious disease epidemiology is the threshold behaviour that relates R_0 to critical vaccination coverage (V_c) in a homogeneously mixing population: $V_c = 1 - (R_0)^{-1}$ (Grassly & Fraser, 2008). $R_0 > 1$ indicates a positive probability for occurrence of a major epidemic. Thus achieving a critical vaccination coverage would imply that a major outbreak will not occur.

The notions of final size and basic reproduction number can also be extended to a population of households. Household final size data present the distribution of number infected by household size. Household reproduction number is calculated based on how infectious an household is and for how long (Fraser, 2007).

1.1.3 Statistical inference for epidemic data

While stochastic epidemic models offer useful insights about the epidemic behaviour, having observed some data from an outbreak, appropriate statistical procedures are required to infer the underlying transmission dynamics. A key step in this process is to specify the appropriate likelihood function based on an underlying epidemic model.

For a completed SIR outbreak with n infection and n removal events (in a population of $N \geq n$ individuals), the data consist of the infection times $\mathbf{t^I} = \{t_1^I, \ldots, t_n^I\}$ and removal times $\mathbf{t^R} = \{t_1^R, \ldots, t_n^R\}$ occurring during a time interval $[0, T]$. Note that these are ordered times such that the time of the first infection is t_1^I and that the subsequent infection times are $\mathbf{t^I_{-1}} = \{t_2^I, \ldots, t_n^I\}$. Following O'Neill & Roberts (1999), the density of $(\mathbf{t^I_{-1}}, \mathbf{t^R})$ conditional on (β, γ, t_1^I) is given by

$$P(\mathbf{t^I_{-1}}, \mathbf{t^R}|\beta, \gamma, t^I_1) = \prod_{i=1}^{n} \gamma I(t^R_i -) \prod_{j=2}^{n} \beta I(t^I_j -) S(t^I_j -) \exp\Bigg\{ - \int_{t^I_1}^{T} \Big(\beta I(t) S(t) + \gamma I(t) \Big) dt \Bigg\}. \tag{1.2}$$

Note that both the infection and recovery processes are assumed to be Markovian as described in Section 1.1.1. Thus the labels (i, j) do not refer to specific individuals but the order of occurence of events.

Statistical inference becomes more challenging when the infection times are not observed, a common type of partial observation in epidemics of the SIR type. The observed data consist only of the removal times $\mathbf{t^R}$ and the expression (1.2) is required to be marginalised with respect to the unobserved infection times. This introduces multiple integrals in the expression which may render the likelihood analytically intractable.

Becker & Britton (1999) and the references therein provide likelihood-based procedures to infer from completely observed data and martingale based procedures for partially observed data. O'Neill & Roberts (1999) provide a procedure to fit GSEM to partially observed data in the Bayesian setting.

While most literature on statistical inference on infectious disease data use observed datasets to develop and / or demonstrate modelling and inference procedures, some are devoted to devoloping the necessary models, estimators and their asymptotic properties towards statistical inference (building the required theory for data analysis) without using any observed datasets (Becker & Hasofer, 1997 and Britton, 1998).

1.1.4 Some modelling considerations

Time

The GSEM is specified in continuous time. However, epidemic models and their associated data have also been considered using other time scales, for example, discrete time or generations defined by infectious periods. Assuming that the infectious period is fixed and that infections occur towards the end of infectious periods (Reed-Frost assumptions), epidemic models can be defined in generations and result in chain binomial models (Becker, 1989 and Britton, 2010). The chain binomial model description is useful even when infections are not observed by generation, for example, household final size data (non-temporal) can be described by marginalising over possible epidemic chains.

A modification to the GSEM is that the infectious period is not exponentially distributed making the removal process non-Markovian (Streftaris & Gibson, 2004). Such models require additional book-keeping in terms of time since infection and labelling of infected individuals (labels are not interchangeable as in the Markovian case).

Population structure

Assuming homogeneity in susceptibility, infectivity and mixing (contact process) across all individuals in the population may be suitable for some situations. When sub-groups in a population have varying levels of susceptibility and/or infectivity such as age-dependence, such heterogeneities have been accounted for using multi-type models (see Britton, 1998). Ignoring such heterogeneities may have strong implications.

Empirical studies have been conducted to understand mixing patterns (Mossong et al., 2008). Mixing at more than one level such as mixing at population level (global), household level and school or work-place levels have also been addressed in the epidemic modelling literature. The implications of such extensions on key parameters such as final size and R_0, have also been studied (Ball et al., 1997, Ball & Lyne 2002, 2007).

Observing epidemics in practise and their implications

How an epidemic is observed limits the statistical analysis and what can be possibly inferred from the data. While this is a vast topic spanning over multiple issues, some of the key issues that are considered towards statistical analyses are whether the data are temporal (event times) or non-temporal (final size), whether the entire population, certain sub-population or a sample is observed and presence of missing data and / or latent variables. Rhodes et al. (1996) provide a hierarchy of information levels in observed data and appropriate methods for analysing data from each level. These methods assume that data arise from observing counting processes.

1.2 Modelling epidemics in households

While homogeneous mixing may be appropriate in some situations such as describing outbreaks in boarding schools or military bases, they may not reflect realistic contact processes when the population contains natural grouping structures. From an epidemiological perspective, households are important units for transmission (House et al., 2012).

1.2.1 The role and importance of households in epidemics

When data are collected at the household level, whether merely observed or designed with respect to some intervention, the following considerations are approriate or even advantageous:

(i) Many infectious diseases require close contacts for transmission. For such infections, within-household transmission is identified as an important component of spread due to contacts occurring in close proximity and with prolonged durations (Kinyanjui et al, 2016). For example, Yang et al. (2009b) estimate the household secondary attack rate for Influenza A (H1N1) to be 27% (95% CI: 12.2% - 50.5%) which is a non-negligible contribution to the transmission process.

(ii) Household members are co-located and therefore they form an efficient sampling unit (House et al., 2012). Sampling households based on index cases is resource-efficient (as they form an exposure group); resource-intensive methods can be used to make observations more accurate on a smaller proportion of the population.

(iii) A related issue is that interventions may be targeted at the household level (House et al., 2012).

(iv) Durations that drive the transmission dynamics such as latent period, infectious period and generation time are estimable in household-based studies (Gough, 1977, Klinkenberg & Nishiura, 2011). In some household-based studies these quantities have been estimated along with transmission parameters (O'Neill et al., 2000, Cauchemez et al., 2004).

(v) Estimates of vaccine efficacy can be directly obtained from household based studies (Lau et al., 2015).

Depending on the research question and available data (along with information on study design and / or observation of the contact process), models can be specified using one level of mixing (only within-household) or two levels of mixing that include mixing outside the household (global contacts). From the perspective of epidemic behaviour, for epidemics with two levels of mixing, the final size distributions, basic reproduction numbers (individual and household), vaccination thresholds and associated vaccination strategies have been derived in Ball et al. (1997). A specific derivation of the household basic reproduction number is provided in Fraser (2007).

From the statistical inference perspective, data from household outbreaks have been of two types: temporal (observation of event times) and non-temporal (observation of final sizes in households). Analyses of both types of data have been described in the literature (Becker, 1989, O'Neill et al., 2000), especially in the context of transmission of respiratory infections and will be briefly reviewed in the following.

1.2.2 Final size data

Let k be the number of individuals in a household, all of whom are still susceptible before the start of the outbreak and let $j = 0, 1, \ldots, k$ be the number infected in the household by the end of the outbreak. Then j is referred to as the outbreak *final size* in a household of size k.

For a population of households, let n_{kj} be the number of households of size k with j infected individuals at the end of outbreak. The set of numbers $\{n_{kj}, k = 1, \ldots, K; j = 0, 1, \ldots, k\}$ constitute the final size data from an outbreak in a population of households. When only within-household transmission is of interest, chain binomial models based on the so called Reed-Frost assumption have been used to infer from household level final size data (Becker, 1989). O'Neill & Roberts (1999) present a Bayesian inference procedure for a measles outbreak final size data in households of size 3.

The Longini-Koopman model (Longini & Koopman, 1982) accounts for two levels of mixing: within-household and community transmission. The model is specified for the household-level final size data. They also presented estimation procedures for truncated data which would arise from *case-ascertained sampling*, i.e., sampling of households conditionally on household outbreak final size ≥ 1. O'Neill et al. (2000) present methods to infer the transmission parameters from final size data with two-levels of mixing using the Longini-Koopman model in the Bayesian setting. They also generalise the model and infer an additional parameter for probability of protection due to vaccination.

Demiris & O'Neill (2005) infer transmission rate parameters in a multi-type epidemic model with two levels of mixing using household final size data along with grouping individuals into strata (based on antibody titre levels). The problem of intractability of likelihood due to not observing the transmission events is dealt with using a data-augmentation procedure on random directed graph defining the contact structure at the population level. O'Neill (2009) modified this inference procedure to observations from sample data. Knock & O'Neill (2014) extended this approach further to include Bayesian model choice procedures for competing models.

1.2.3 Data containing event time information

Data containing information on observed event times for household members often include the times of onset of symptoms, removal or being tested positive for infection.

Rampey et al. (1992) present a discrete-time model for rhinovirus infection data in households and estimate within-household transmission (here, secondary attack rate) and global transmission (here, probability of infection from the community) parameters using likelihood-based procedures. The data consist of symptom times and viral culture test results. They use a multitype epidemic model with varying susceptibility and infectivity levels in an SEIR setting.

O'Neill et al. (2000) present Bayesian methods to infer only within-household transmission in an SEIR setting. The data consisted of interremoval times (time between two removals) in households of size 2. They provide two ways to handle the unobserved event times: (i) sample them jointly with parameters using a data-augmentation algorithm and (ii) marginalise over them using a simulation-based algorithm.

Cauchemez et al. (2004) used a continuous-time model for data on influenza infections in households and estimate within-household and global transmission parameters along with infectious periods. The data consist of symptom onset times and relevant laboratory test results. The study design is case-ascertained follow-up where households are sampled based on symptom times of an index case. A Bayesian MCMC procedure is used to estimate the model parameters and hypotheses tests were performed using Bayes factors.

Yang et al. (2006) used a discrete-time model on two household-randomised influenza trials data and estimate within-household and global transmission parameters along with vaccine efficacy parameters. The data consist of symptom onset times and relevant laboratory test results. The study design is case-ascertained follow-up where households are sampled based on symptom times of an index case. A conditional likelihood based approach is used towards statistical inference. Yang et al. (2009a) extends this analysis to fitting a (discretised) continuous-time model to include asymptomatic infections under a Bayesian framework.

1.3 Inferring direct transmission: Data and modelling

Understanding the transmission routes, especially, direct (person-to-person) transmission is of epidemiological importance towards management and control of infection. A particular case in point is influenza which has the potential to cross zoonotic barriers (of avian and swine origins) and result in a pandemic (Yang et al., 2007b).

1.3.1 Models to infer direct transmission from oubreak data

The literature described in Section 1.2.2 make use of household-level final size data to estimate within-household and global transmission parameters. However, those models and their inference procedures require further modifications to use them for testing the existence of direct transmission from household-level final size data. For data containing event time information, some procedures for such tests are available in the SIR setting.

Yang et al. (2007a) present an inference procedure to test the existence of direct transmission using a discrete-time model parameterised with escape probabilities for transmission. In particular, three parameters are used with respect to transmission from the community, from those within the household or close contacts and from those outside the household or casual contacts. The authors discuss non-standard conditions due to which likelihood-ratio type tests cannot be used, one of the relevant conditions to this thesis being that the hypothesised parameter for direct transmission falls at the boundary of the parameter space under a null hypothesis (i.e. absence of direct transmission). They develop a resampling-based test procedure to test for the existence of direct transmission. An underlying SIR epidemic model is assumed where the incubation period (duration from infection to onset of symptoms) is assumed to be the same as latent period (duration from infection to onset of infectiousness). All infected individuals are assumed to be symptomatic. Simulations were used to evaluate the performance of the developed inference procedures.

The procedure developed in Yang et al. (2007a) was used as the basis to test person-to-person transmission in avian influenza A (H5N1) (Yang et al., 2007b). However the procedure is modified to account for data collected using the case-ascertained design based on methods presented in Yang et al. (2006).

1.3.2 Transmissibility in Hepatitis A virus (HAV) infection

While there is some evidence for direct transmission in Hepatitis A virus (HAV) infections, most research involved so far do not use data from well-designed studies and / or rigorous statistical approach. Case reports suggest direct transmission in close contacts, especially within households (Sato, 1988 and Kumbang et al., 2012).

Victor et al. (2006) conducted a laboratory-based HAV surveillance to identify index cases (first symptomatic case within a household). They also enrolled the household and school/day-care contacts of index cases. The data consisted of symptom times and serological test results among the contacts of identified index cases. They used regression analysis that accounts for correlation among contacts within exposure groups. The results suggest that household contacts have higher odds of infection compared to school/day-care contacts.

Lima et al. (2015) used HAV genetic sequence data from index cases in households and their household contacts to investigate direct transmission within the household. Evidence towards direct transmission was evaluated based on sequence homology between index cases and their household contacts.

1.3.3 Models to infer direct transmission from HAV oubreaks

Zhang & Iacano (2018) estimate the reproduction numbers for HAV from an elementary school outbeak data. The data consist of symptom times, serological test results and social contact network of the children. They use likelihood methods to estimate the transmission trees, reproduction numbers and the serial interval. In addition, they fit an epidemic model to the outbreak data and estimate the model parameters. The results suggest person-to-person transmission may have played a key role in this elementary school outbreak and that there was no difference between symptomatic and asymptomatic infections in terms of transmissibility.

This thesis presents methods to estimate transmission parameters for person-to-person transmission from household outbreak data. A model comparison procedure to test the hypothesis of direct transmission is also developed. The natural history of infection used resembles that of HAV infection as in Zhang & Iacano (2018). The data that is assumed to be observed in this application also closely follows the data that would be typically observed from HAV outbreaks.

1.4 Bayesian Inference

Having specified a data generating process $P_{\mathbf{y}|\theta}(\mathbf{y}|\theta)$ for the observed data $(\mathbf{y})$, Bayesian inference allows to quantify the uncertainty in the underlying unknown quantities / model parameters θ. It uses the posterior probability $P_{\theta|\mathbf{y}}(\theta|\mathbf{y})$ of the parameters given the observed data $\mathbf{y}$ along with prior probability $P_\theta(\theta)$ of the parameters through the Bayes' rule

$$P_{\theta|\mathbf{y}}(\theta|\mathbf{y}) = \frac{P_{\mathbf{y}|\theta}(\mathbf{y}|\theta)P_\theta(\theta)}{\int_\Theta P_{\mathbf{y}|\theta}(\mathbf{y}|\theta)P_\theta(\theta)d\theta}. \tag{1.3}$$

The posterior $P_{\theta|\mathbf{y}}(\theta|\mathbf{y})$ contains all information necassary to infer about the parameters from the data. This representation follows from the Bayesian thinking: all sources of uncertainty about the parameters (data $\mathbf{y}$ and prior P_θ) are included in the posterior distribution and that all uncertainty about the parameters is expressed in terms of probabilities (here, the posterior probability $P_{\theta|\mathbf{y}}(\theta|\mathbf{y})$).

1.4.1 Why is Bayesian inference used in epidemic modelling?

Analysis of data arising from infectious disease outbreaks is a non-standard problem. In addition to inherently dependent and incomplete observation, factors such as the natural history of infection, population structure and observation design render unique settings for each problem.

Classical inference involves parameter estimation using estimators with known distributional properties. This requires the necessary theoretical results to be developed to arrive at uncertainty quantification for the estimators (O'Neill, 2002). This may involve deriving central limit theorems for estimators of parameters governing dependent processes, which is rather technical in nature (for example, see Becker & Hasofer, 1997).

When the likelihood is analytically tractable, deriving posterior quantities analytically under Bayesian inference may involve efforts similar to those described for classical inference. However, partial observation introduces additional integrals to the likelihood expression (cf. Section 1.1.3) which may render the likelihood analytically intractable. Computational techniques such as Markov chain Monte Carlo (MCMC) methods have made sampling from the posterior distribution of all model unknowns, including the missing data items, feasible. Using such techniques require developing algorithms to sample from the correct posterior density.

1.4.2 Bayesian computation

Development and implementation of algorithms to efficiently sample from the posterior distribution are at the core of Bayesian computing. MCMC algorithms such as Metropolis-Hastings and Gibbs samplers are used in this thesis and are briefly described in the following.

Metropolis-Hastings sampler
The Metropolis-Hastings (MH) sampler is useful for drawing samples from an unnormalised target density $P(\theta)$ by using a Markov chain $\{\theta^{(t)}\}$ that has the target density as its stationary distribution. A candidate value θ' is generated from the conditional density $k(\theta'|\theta^{(t)})$ (known as the *proposal density*). The candidate value is accepted (i.e., $\theta^{(t+1)} = \theta'$) with probability $A(\theta', \theta^{(t)})$ where

$$A(\theta', \theta^{(t)}) = \min\left\{1, \frac{P(\theta')k(\theta^{(t)}|\theta')}{P(\theta^{(t)})k(\theta'|\theta^{(t)})}\right\}. \tag{1.4}$$

Otherwise, $\theta^{(t+1)} = \theta^{(t)}$ (Robert & Casella, 2010).

Gibbs sampler
The Gibbs sampler is a special case of the MH algorithm when the full conditional distributions of a multi-dimensional random variable are available and can be used to sample.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ be a multi-dimensional random variable for which full conditional distributions of the form $P(\theta_k|\boldsymbol{\theta}_{-k}), k \in 1, \ldots, r$ are available, where $\boldsymbol{\theta}_{-k} = (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \theta_r)$. Note that each component θ_k

can be of one or more dimensions. Then the Gibbs sampler generates a Markov chain $\{\boldsymbol{\theta}^{(t)}\}$ from the joint distribution $P(\boldsymbol{\theta})$ by simulating from the full conditional distributions as follows (Robert & Casella, 2010):

$$\begin{cases} \theta_1^{(t+1)} \sim P(\theta_1|\theta_2^{(t)},\ldots,\theta_r^{(t)}), \\ \theta_2^{(t+1)} \sim P(\theta_2|\theta_1^{(t+1)},\theta_3^{(t)},\ldots,\theta_r^{(t)}), \\ \vdots \\ \theta_r^{(t+1)} \sim P(\theta_r|\theta_1^{(t+1)},\ldots,\theta_{r-1}^{(t+1)}). \end{cases} \tag{1.5}$$

Data augmentation

When the model contains latent variables ($\mathbf{z}$), the parameter space (θ) can be augmented to include the latent variables. If the joint posterior $P_{\theta,\mathbf{z}|\mathbf{y}}(\theta,\mathbf{z}|\mathbf{y})$ of parameters and latent variables can be hierarchically specified, sampling from the joint posterior can be performed using a Gibbs sampler approach (see Chapter 3 for further details).

1.5 Scope and structure of the thesis

1.5.1 What is addressed in this thesis

The main aim of this thesis is to develop methods to test the existence of direct transmission using data observed from outbreaks in households. Yang et al. (2007a) has addressed similar testing question for the existence of direct transmission using household outbreak data. This application, motivated by hepatitis A virus (HAV) infection, extends their work in the following ways: (i) the assumption that the incubation period is the same as the latent period is modified to allow the latent period to be shorter than the incubation period, i.e. an individual is infectious before being symptomatic. Consequently there are two unobserved times, namely, times of infection and onset of infectiousness, (ii) the presence of asymptomatic infections and individuals with false symptoms and (iii) addition of serological test results that are non-sensitive due to waning of antibodies.

The model state space follows the natural history of HAV infection and the transmission model is developed to account for transmission from two sources, from the contacts within households and from the environment. The observations are assumed to be imperfect, including both asymptomatic infections and false symptoms. The serology is also imperfectly observed due to antibody waning.

The likelihood is provided for the primary setting which is a completed outbreak in a population of households. The likelihood expression is further extended to an ongoing outbreak with censored observation and a case-ascertained follow-up design. The case-ascertained follow-up is common in practise as could be seen in Sections 1.2 and 1.3.

Models are specified in continuous time and parameterised using transmission hazards from both sources (from infectious household members and from the community) for the three described settings. The model for completed outbreak in a population of households is additionally specified in discrete time and is parameterised using escape probabilities per day from both sources. The inference and model comparison procedures are presented only for this setting (i.e. primary setting in discrete time).

The posterior of the transmission parameters are sampled by augmenting the unobserved infection times using a Gibbs sampler approach. The infection status is updated along with the infection times for imperfect observations. Deviance information criteria (DIC) is computed towards model comparison. DIC computation uses the Gibbs sampler routines for sampling from the joint posterior presented in Chapter 3.

1.5.2 Structure of the thesis

Chapter 2 presents the model for an HAV outbreak. The model is developed hierarchically with a model for transmission and a model for observation. The model is presented for imperfectly observed completed outbreak in a full cohort. It is further extended for two other observation designs.

Chapter 3 presents algorithms to sample from the joint posterior of the model parameters and latent variables using a data-augmented MCMC approach. Chapter 4 presents model comparison using DIC and procedures to implement DIC for the described model. Chapter 5 presents the results for evaluating the performance of parameter estimation and model comparison methods (presented in Chapters 3 and 4) based on simulated outbreak data. Chapter 6 discusses the results based on simulated data and provides directions on extending the modelling and inference to include more complex data.

Chapter 2

Modelling Outbreak Data

"It is not really difficult to construct a series of inferences, each dependent upon its predecessor and each simple in itself. If, after doing so, one simply knocks out all the central inferences and presents one's audience with the starting-point and the conclusion, one may produce a startling, though possibly a meretricious, effect."
- Sir Arthur Conan Doyle (Sherlock Holmes in The Dancing Men)

Statistical modelling aspects to describe the data observed from an outbreak in a population of households are presented in this chapter. A model for the underlying data generative process is first developed from a mechanistic interpretation. The observation process of the underlying model further describes the data that are generated. Some alternative observation designs are also presented in this chapter.

2.1 Outbreak setting and timelines

We assume a closed community of households in which an outbreak occurs. We assume that in a fully susceptible population, the outbreak is initiated by the environment becoming infectious and is further propagated by infectious individuals (or infectious household members in our setting). An individual can be infected either by the environment or by an infectious household member (person-to-person transmission). We additionally assume that person-to-person transmission between members of two different households is not possible, which make households independent units of observation.

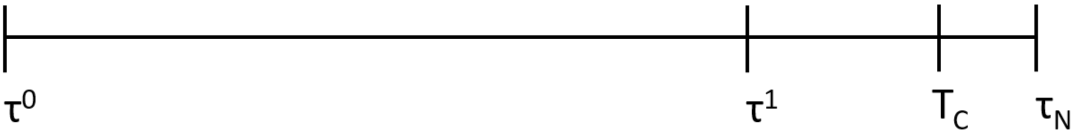
The outbreak is complete when the environment ceases to be infectious and there are no new infections in the population any more. Note that the outbreak could be complete earlier, if the entire population is already infected and recovered or immune i.e., there are no susceptible individuals present. In a perfectly observed setting, the duration of the outbreak is broadly characterised by three time points as shown in Figure 2.1: (i) τ^0, the time when environment becomes infectious (ii) τ^1 the time when the environment stops being infectious and (iii) τ_N the time when the last infectious individual has recovered.

For example, in a setting where the environment is treated, τ^1 is observed as the time of treatment. When all infections are treated or infected individuals are quarantined in a closed population, τ_N can be observed as the last treated or quarantined infection. In some other settings, it is possible to observe τ^0 as the time when environment becomes infectious. In practise, not all three time points are observed. In our setting, none of these time points are observed and we use time points derived from observations to approximate them.

2.2 State-space and transitions

From the viewpoint of epidemic model structure, the natural history of hepatitis A virus (HAV) infection follows an SEIR model (i.e, with susceptible, exposed, infectious and removed compartments). From the viewpoint of

Figure 2.1: Outbreak timelines. The three time points are τ^0 and τ^1 when the environment starts and stops to be infectious, respectively, and τ_N when the last infectious individual has recovered. T_c is an observed time point which is the last observed symptom time (see section 2.5.1).



observations, an infected individual becomes infectious before manifesting symptoms. This, in addition to the possibility of asymptomatic infections, leads to the following model specification.

2.2.1 State-space and transition times

The state-space of the model consists of the following five states: susceptible (U), infected but not yet infectious (E), infectious (I), infectious and symptomatic (S), and removed (N). In this context, the removed state refers only to recovery from infection, but in general, removal may refer to an individual becoming non-infectious due to recovery, treatment or immunisation. A susceptible individual, upon infection, passes through these model states. Thus the state-space for a single individual is $\mathcal{X} = \{U, E, I, S, N\}$.

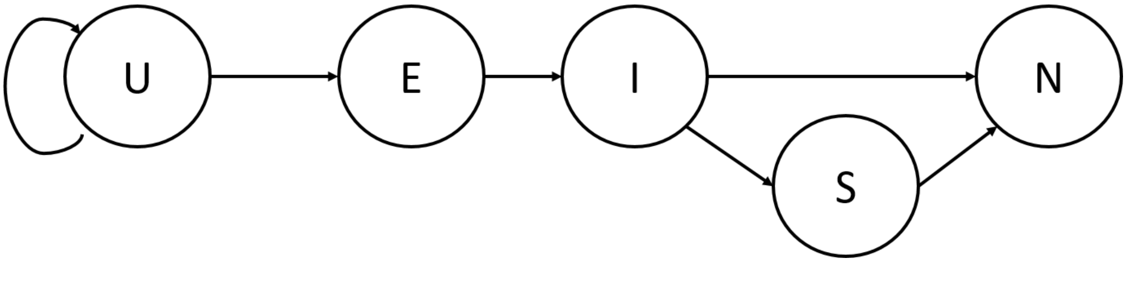
In this setting, all individuals are susceptible at time τ^0, i.e., $X_i(\tau^0) = U, i = 1, \ldots, N$. For an infected individual i who manifests symptoms, the evolution of the process $X_i(t)$ is specified in terms of the following four transition times, t_i^E, t_i^I, t_i^S and t_i^N:

$$X_i(t) = \begin{cases} U & \text{if} \quad t < t_i^E, \\ E & \text{if} \quad t_i^E \leq t < t_i^I, \\ I & \text{if} \quad t_i^I \leq t < t_i^S, \\ S & \text{if} \quad t_i^S \leq t < t_i^N, \\ N & \text{if} \quad t \geq t_i^N, \end{cases} \tag{2.1}$$

where t_i^E, t_i^I, t_i^S and t_i^N are the times of infection, onset of infectiousness, onset of symptoms and removal defined with respect to the time origin τ^0, the start of outbreak.

Let m_i be a binary mark associated with t_i^E, the infection time. When $m_i = 1$ the individual i is symptomatic with an observed time of onset of symptoms t_i^S. When $m_i = 0$ the individual i is asymptomatic. For an infected individual i who does not manifest symptoms, the state S is considered to be latent and the transition is made directly from state I to state N. The evolution of the process for symptomatic and asymptomatic individuals is depicted in Figure 2.2.

Figure 2.2: State space of the process defined by natural history of HAV. The natural history of HAV infection with the following states: susceptible (U), infected but not yet infectious (E), infectious (I), symptomatic (S) and removed (N). Asymptomatic individuals progress directly from state I to state N.



The process $X_i(\cdot)$ generates a history $\mathcal{F}_{it}^*$, a σ-algebra based on the visited states and corresponding transition times $\{t_i^E, m_i, t_i^I, t_i^S, t_i^N\}$ of the process over the interval $[\tau^0, t]$. The observed history $\mathcal{F}_{it}$ is limited to the observed states and corresponding transition times $\{t_i^E, m_i, t_i^S\}$. Note that although the times of infection t_i^E

are not part of observation, they are here included in $\mathcal{F}_{it}$ to obtain convenient likelihood expressions.

For an household H_i containing the individual i, $\mathcal{H}_{it} = \{\mathcal{F}_{ju}, j \in H_i, u \leq t\}$ denotes the history of all household members (including i) up to time t.

2.2.2 Sojourn time distributions

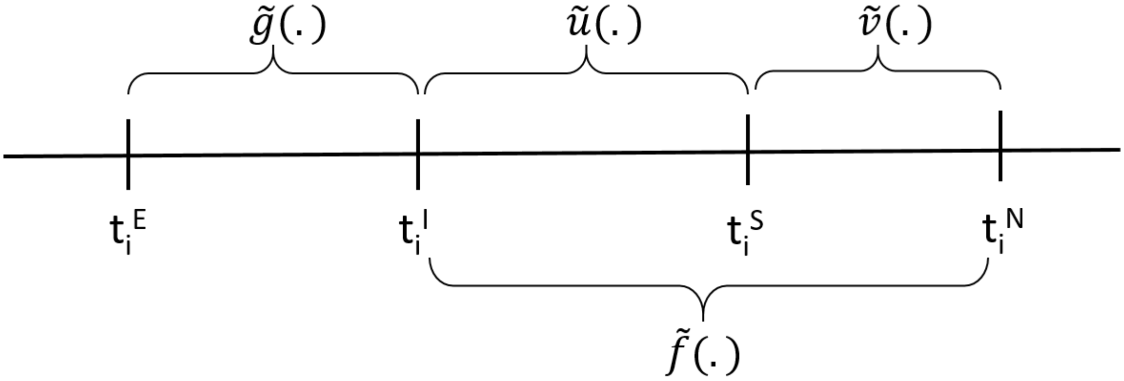
The model for infection time t_i^E defining the transition from state U to E is explained in Section 2.3. Upon infection, the natural history of infection is modelled in terms of three non-overlapping intervals: the latent period $(t_i^I - t_i^E)$, the duration between the onset of infectiousness and symptoms $(t_i^S - t_i^I)$, and the duration between onset of symptoms and removal (when the individual becomes non-infectious) $(t_i^N - t_i^S)$. The densities of these durations (sojourn times) are given by $\tilde{g}(\cdot)$, $\tilde{u}(\cdot)$ and $\tilde{v}(\cdot)$, respectively, as shown in Figure 2.3. The densities of the transition times are defined in terms of the corresponding densities of sojourn (waiting) times:

$$\begin{aligned} g(t_i^I|t_i^E) &= \tilde{g}(t_i^I - t_i^E), \\ u(t_i^S|t_i^I) &= \tilde{u}(t_i^S - t_i^I), \\ v(t_i^N|t_i^S) &= \tilde{v}(t_i^N - t_i^S). \end{aligned}$$

The ranges of these three sojourn times are given by (l_{min}, l_{max}), (s_{min}, s_{max}) and (n_{min}, n_{max}), respectively (their values are provided in section 3.1). When an infection is asymptomatic, the time of onset of symptom t_i^S is not observed. However, the duration of infectiousness of asymptomatic individuals is assumed to be the same as that of symptomatic individuals. The duration of infectiousness (i.e. the interval between t_i^I and t_i^N) is then sum of the two durations $t_i^S - t_i^I$ and $t_i^N - t_i^S$. Thus its corresponding density for sojourn time $\tilde{f}(t_i^N - t_i^I)$ and the density of transition time $f(t_i^N|t_i^I)$ are obtained by convolution:

$$f(t_i^N|t_i^I) = \int_{t_i^I}^{t_i^N} v(t_i^N|t_i^S)u(t_i^S|t_i^I)dt_i^S.$$

Figure 2.3: Sojourn times and their densities.



2.3 Transmission model

The probability that individual i is infectious at time t is denoted by $I_i(t|\mathcal{F}_{it-})$. For a symptomatic individual, recall that the history $\mathcal{F}_{it}$ of the process includes the transition times t_i^E and t_i^S to the states E and S, respectively. Let $I_i^{m=1}(t|\mathcal{F}_{it-})$ be the probability that individual i with symptomatic infection is infectious at time t, given the history $\mathcal{F}_{it}$:

$$I_i^{m=1}(t|\mathcal{F}_{it-}) = \begin{cases} 0 & \text{if} \quad X_i(t-) = U, \\ \int_{t_i^E}^t g(\tau|t_i^E)d\tau & \text{if} \quad X_i(t-) = E, \\ 1 - \int_{t_i^S}^t v(\tau|t_i^S)d\tau & \text{if} \quad X_i(t-) = S. \end{cases} \tag{2.2}$$

For an asymptomatic individual the history $\mathcal{F}_{it}$ of the process includes only the transition time t_i^E to state E. Let $I_i^{m=0}(t|\mathcal{F}_{it-})$ be the probability that individual i with asymptomatic infection is infectious at time t, given the history $\mathcal{F}_{it}$:

$$I_i^{m=0}(t|\mathcal{F}_{it-}) = \begin{cases} 0 & \text{if} \quad X_i(t-) = U, \\ \int\limits_{t_i^E}^{t} g(\tau|t_i^E)\big[1 - F(t|\tau)\big]d\tau & \text{if} \quad X_i(t-) = E, \end{cases} \tag{2.3}$$

where $F(t|\tau) = \int\limits_{\tau}^{t} f(u|\tau)du$. It is obvious that the probability of individual i being infectious before being infected (i.e., $X_i(t-) = U$) is $I_i(t|\mathcal{F}_{it-}) = 0$.

Let λ be the instantaneous hazard of infection of a still susceptible individual from an infectious household member. Then the hazard of a still susceptible household member i to be infected by a household member j that is infectious at time t be

$$\lambda_{ij}(t|\mathcal{F}_{it-}, \mathcal{F}_{jt-}) = \lambda I_j^m(t|\mathcal{F}_{jt-}). \tag{2.4}$$

For the sake of completeness, one can define $\lambda_{ij}(t|\lambda, \boldsymbol{\beta_i}, \boldsymbol{\beta_j}) = \lambda I_j^m(t|\mathcal{F'}_{jt-}, \boldsymbol{\beta_j})U_i(t|\mathcal{F'}_{it-}, \boldsymbol{\beta_i})$ using both infectiousness $I_j^m(t)$ of j and susceptibility $U_i(t)$ of i. Here $\mathcal{F'}_t$ is the history including relevant covariates up to time t and $(\boldsymbol{\beta_i}, \boldsymbol{\beta_j})$, parameters with respect to the covariates. See Rhodes et al. (1996) for a rigorous treatment of this formulation, Yang et al. (2009), for estimating the antiviral efficacies and Cauchemez and Ferguson (2011) for estimating transmission risk factors.

The infectiousness can additionally be defined using a duration-dependent function, especially if a model for viral shedding is available. In this case $\lambda_{ij}(t|\cdot) \propto \gamma(t - t_j^E)$ where $\gamma(\cdot)$ is a function defining the infectivity of j at time t. Although we do not use any information on covariates or infectivity in our model, it can be extended using the described formulations.

Let μ be the hazard for an individual being infected by the environment. It is assumed to be a constant as long as the environment is infectious. If the environment is infectious up to time τ^1, the infectiousness of the environment at any time t thus is

$$\mu(t|\tau^1) = 1_{\{t \leq \tau^1\}}\mu. \tag{2.5}$$

Finally, the model for infection time is defined. The probability density that individual i is infected on day t_i^E is given by

$$Z_i(t_i^E|\mathcal{H}_{it_i^E-}) = \Bigg\{\mu(t_i^E|\tau^1) + \sum_{\substack{j\in H_i,\\ j\neq i}} \lambda_{ij}(t_i^E|\mathcal{F}_{it_i^E-}, \mathcal{F}_{jt_i^E-})\Bigg\} \exp\Bigg[- \int\limits_{u=\tau^0}^{t_i^E} \Bigg(\mu(u|\tau^1) + \sum_{\substack{j\in H_i,\\ j\neq i}} \lambda_{ij}(u|\mathcal{F}_{iu-}, \mathcal{F}_{ju-})\Bigg)du\Bigg]. \tag{2.6}$$

The probability that an individual is not infected during the outbreak up to some time t is given by

$$Q_i(t|\mathcal{H}_{it-}) = \exp\Bigg\{ - \int\limits_{u=\tau^0}^{t} \Bigg(\mu(u|\tau^1) + \sum_{\substack{j\in H_i,\\ j\neq i}} \lambda_{ij}(u|\mathcal{F}_{iu-}, \mathcal{F}_{ju-})\Bigg)du\Bigg\} \tag{2.7}$$

The above model is specified in continuous time. However, the time scale of the observed data is generally in days. Yang et al. (2009) and Cauchemez & Ferguson (2011) use an underlying continuous time model but provide a discretised likelihood in days. Rampey et al. (1992) and Yang et al. (2006) use a discrete-time model parameterised using escape probabilities in comparison to the continuous time model using infection hazard rates (μ, λ).

A discrete-time version of the presented model is provided in Appendix A. The model is specified using escape probabilities per day (c, q) and follows in the spirit of Rampey et al. (1992).

2.4 Observation model

Of all events pertaining to the underlying transmission process and the natural history of disease, only a portion is observed. In particular, for each individual in the study population the following data are observed: the time of onset of symptoms t_i^S and whether or not an individual tests serologically positive for infection at the end of outbreak.

Symptoms:
A certain proportion of infected individuals go on to manifest symptoms. The sensitivity parameter (η) is the proportion of symptomatic cases among all infections $\eta = P(m_i = 1|X_i(\tau_N) \neq U)$.

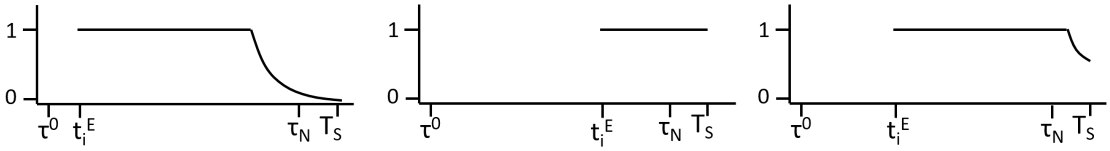
A certain proportion of non-infected indivduals (are reported to) manifest symptoms and are called *false symptoms*. This proportion is given by $(1 - \psi)$ where $\psi = P(m_i = 0|X_i(\tau_N) = U)$ and is called the specificity parameter. There may be more than one reason for false symptoms (parent's incorrect recalling or bias given an already infected child in household or symptoms that are not differential to an underlying infection).

Serology:
Serological tests are performed on sera samples collected at time T_s (some time after the outbreak is complete: $T_s > \tau_N$) from all individuals in the population. Let a_i be an indicator for the serological test result being positive. For an individual i infected at time t_i^E, the probability of being tested positive for IgM antibodies at time T_s is given by

$$P(a_i = 1) = R(T_s - t_i^E|\alpha, t_w) = 1_{[0,t_w]}(T_s - t_i^E) + (1 - 1_{[0,t_w]}(T_s - t_i^E))e^{-\alpha(T_s - t_i^E - t_w)} \tag{2.8}$$

where, t_w is the threshold duration from time of infection t_i^E after which the IgM antibodies begin to wane. The parameter α is the rate of decrease in the detectability of antibodies and is related to the waning rate of antibodies. Figure 2.4 illustrates the observation model for serology.

Figure 2.4: Model for observation of serological test results using IgM antibodies. The probability of being tested positive (y-axis) is provided for three different infection times (t_i^E) with respect to the outbreak timelines : an infection close to the start of outbreak, an infection close to the end of outbreak and an infection somewhere in between them.



IgM antibodies are a short term indicator of infection. They can be observed immediately after infection but wane over a short duration defined by t_w and α. IgG antibodies, on the other hand, offer a long term indication of infection. They can be detected after a certain period (t_x) from infection and do not wane over time. In this setting, we assume that IgM antibodies were tested for. If the serological test involves IgG antibodies, then the observation can be modelled as

$$P(a_i = 1) = R(T_s - t_i^E|\beta, t_x) = 1 - 1_{[0,t_x]}(T_s - t_i^E) + (1_{[t_x,\infty]}(T_s - t_i^E))\phi(\beta, t) \tag{2.9}$$

where t_x is the threshold duration from infection time after which the IgG antibodies begin to develop. The parameter β is the rate of increase in the detectability of antibodies and is related to the rate of development of antibodies. For example, $\phi(\beta, t)$ can be a logistic function with $t = T_s - t_i^E - t_x$ and has values in the interval $[0, 1]$.

2.5 Likelihood

In this section likelihood expressions are provided for three different observation settings or designs. The data that are assumed to be observed consist of symptom times and serological test results $\mathbf{y} = \{(m_i, t_i^S, a_i), i = 1, \ldots, N\}$. The missing data $\mathbf{z} = \{t_i^E, i = 1, \ldots, N\}$ are the infection times. Together, they form the complete-data $\mathbf{C} = (\mathbf{z}, \mathbf{y}) = \{(t_i^E, m_i, t_i^S, a_i), i = 1, \ldots, N\}$ and $\mathcal{F}_{it}$ is the history based on $\mathbf{C}$ (see Section 2.2.1).

The likelihood expressions are of the *complete-data likelihood* form ($P(\mathbf{z}, \mathbf{y}|\theta) \propto L(\theta; \mathbf{z}, \mathbf{y})$) based on the complete data $\mathbf{C} = (\mathbf{z}, \mathbf{y})$. It includes the components of transmission, natural history of infection and observation as described in the preceding sections; their notations are collected in Table 2.1 for quick reference.

2.5.1 Completed outbreak in a full cohort

In this setting, the likelihood is provided when: (i) all individuals in the population are observed and (ii) the outbreak is complete i.e., there are no new infections.

Outbreak timelines:
The completed outbreak is defined in Section 2.1 and is characterised by timepoints τ^0, τ^1 and τ_N (see Figure 2.1). In practice, however, these timepoints are not observed. The time point τ^0 is assumed to be known (see Section 5.3 for further explanation on this). Let T_c be the last observed symptom time after which there are no new symptomatic infections. The observed timepoint T_c is used as an approximation for both τ^1 and τ_N. Having observed that the outbreak is complete at T_c, serology samples are collected at $T_s (> T_c)$.

The complete data likelihood.
The complete data contribution from individual i, $C_i = (t_i^E, m_i, t_i^S, a_i)$, contains the infection time, whether symptoms were observed along with the time of symptom onset and the serological test result. Note that for an asymptomatic individual i, $m_i = 0$ and $t_i^S = \infty$.

The parameters of interest are μ and λ. Given the history $\mathcal{H}_{iT_c}$ of all household members of individual i up to time T_c for the complete data C_i, the likelihood contribution from an individual i, $L_i(\mu, \lambda; \mathcal{H}_{iT_c})$ is

$$= \begin{cases} \eta R(T_c - t_i^E) \int\limits_{t_i^I = t_i^E + l_{min}}^{min(t_i^S, t_i^E + l_{max})} u(t_i^S|t_i^I) g(t_i^I|t_i^E) Z_i(t_i^E|\mathcal{H}_{it_i^E-}) dt_i^I, & \text{if } C_i{=}(t_i^E{<}T_c, m_i{=}1, t_i^S{<}T_c, a_i{=}1), \\ (1-\eta) R(T_c - t_i^E) Z_i(t_i^E|\mathcal{H}_{it_i^E-}), & \text{if } C_i{=}(t_i^E{<}T_c, m_i{=}0, t_i^S{=}\infty, a_i{=}1), \\ \eta(1 - R(T_c - t_i^E)) \int\limits_{t_i^I = t_i^E + l_{min}}^{min(t_i^S, t_i^E + l_{max})} u(t_i^S|t_i^I) g(t_i^I|t_i^E) Z_i(t_i^E|\mathcal{H}_{it_i^E-}) dt_i^I, & \text{if } C_i{=}(t_i^E{<}T_c, m_i{=}1, t_i^S{<}T_c, a_i{=}0), \\ (1-\eta)(1 - R(T_c - t_i^E)) Z_i(t_i^E|\mathcal{H}_{it_i^E-}), & \text{if } C_i{=}(t_i^E{<}T_c, m_i{=}0, t_i^S{=}\infty, a_i{=}0), \\ (1-\psi) Q_i(T_c|\mathcal{H}_{iT_c-}), & \text{if } C_i{=}(t_i^E{>}T_c, m_i{=}1, t_i^S{<}T_c, a_i{=}0), \\ \psi Q_i(T_c|\mathcal{H}_{iT_c-}), & \text{if } C_i{=}(t_i^E{>}T_c, m_i{=}0, t_i^S{>}T_c, a_i{=}0). \end{cases} \tag{2.10}$$

Note that the complete-data likelihood in (2.10) is marginalised with respect to the time of onset of infectiousness t_i^I which is also an unobserved time point. A discrete-time version of the complete-data likelihood (equivalent to (2.10)) is provided in Appendix A.

In the special case of perfect observation, the observation model is deterministic with $\eta = 1$ and $\psi = 1$ (i.e., all infected individuals manifest symptoms and there are no false symptoms). Then the likelihood L_i in equation (2.10) reduces to

$$L_i = \begin{cases} Z_i(t_i^E|\mathcal{H}_{it_i^E-}) \int\limits_{t_i^I=t_i^E+l_{min}}^{min(t_i^S,t_i^E+l_{max})} u(t_i^S|t_i^I)g(t_i^I|t_i^E)dt_i^I & \text{if} \quad (t_i^E < t_i^S \leq T_c), \\ Q_i(T_c|\mathcal{H}_{iT_c}), & \text{if} \quad t_i^E > T_c. \end{cases} \tag{2.11}$$

This thesis examines the current setting of completed outbreak in the whole population. However, under the assumptions $\eta \approx 1$ and $\psi \approx 1$, the following two alternative observation designs may often be encounterd in practise. The likelihood under these designs are provided in the following subsections.

2.5.2 Outbreak in progress

In this section, the likelihood is provided when: (i) all individuals in the population are observed (ii) the outbreak is in progress and is observed up to some time T_c' from the start of outbreak.

Outbreak timelines:
Let T_c' be the time up to which the outbreak is observed and that the outbreak is still in progress. In addition, it is assumed that $T_c' < \tau^1$, i.e., the common source (environment) is still infectious. Otherwise $T_c' \approx T_c$ and the observations can be analysed as if the outbreak is (almost) complete.

Under the stated assumptions ($\eta \approx 1$ and $\psi \approx 1$), serological tests are only performed on symptomatic cases that occur during the follow-up period in order to confirm the infection. In this case it can be assumed that samples for serological tests are obtained at t_i^S for symptomatic cases i.e., as and when they occur. Serological data on the non-symptomatic individuals at T_c' may be feasible under this observation design but are deemed redundant when $\eta \approx 1$.

The complete data likelihood.
Data under this design is said to arise from a censored process of transmission and observation. The complete data $C_i = (t_i^E, t_i^S)$ contain the infection and symptom onset times. The complete data likelihood $L_i(\mu, \lambda; \mathcal{H}_{iT_c})$ for an individual i is given as

$$L_i = \begin{cases} \int\limits_{t_i^I=t_i^E+l_{min}}^{min(t_i^S,t_i^E+l_{max})} u(t_i^S|t_i^I)g(t_i^I|t_i^E)Z_i(t_i^E|\mathcal{H}_{it_i^E-})dt_i^I & \text{if} \quad (t_i^E < t_i^S \leq T_c'), \\ Z_i(t_i^E|\mathcal{H}_{it_i^E-})\left(1 - \int\limits_{t_i^S=t_i^I+s_{min}}^{T_c'} \int\limits_{t_i^I=t_i^E+l_{min}}^{min(T_c',t_i^E+l_{max})} u(t_i^S|t_i^I)g(t_i^I|t_i^E)dt_i^I dt_i^S\right) & \text{if} \quad (t_i^E \leq T_c', t_i^S > T_c'), \\ Q_i(T_c'|\mathcal{H}_{iT_c'}), & \text{if} \quad t_i^E > T_c'. \end{cases} \tag{2.12}$$

Those infected during $[T_c' - (l_{max} + s_{max}), T_c' - (l_{min} + s_{min}))$ have some probability of not being symptomatic until T_c' whereas those infected during $[T_c' - (l_{min} + s_{min}), T_c']$ are certainly not symptomatic at or before T_c'.

2.5.3 Case-ascertained sampling

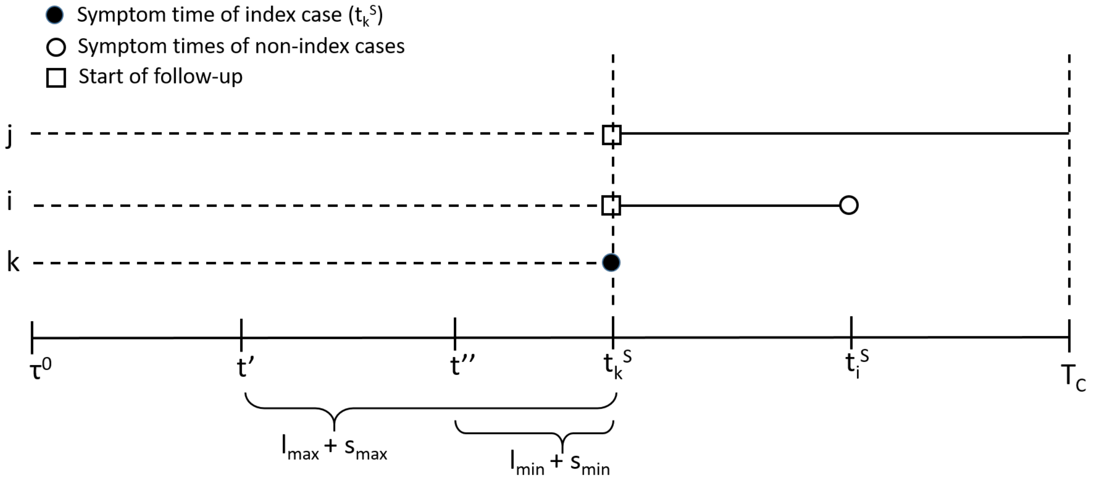
In this setting, the likelihood is provided when: (i) households are sampled on identifying a symptomatic case within the household and (ii) all members of the sampled households are followed-up until the outbreak is complete and there are no new infections.

This observation design is called *case-ascertained follow-up* and is resource-efficient because households with no symptomatic infections are excluded in the sample. However, this design is only useful under the stated assumptions where infected individuals are almost always symptomatic i.e., $\eta \approx 1$ and non-infected individuals do not manifest false symptoms $\psi = 1$. The treatment in this sub-section closely follows Yang et al. (2006) which uses a conditional likelihood approach in the SIR-type epidemic setting.

Outbreak timelines:
Let $H_k, k = 1, 2, \ldots, K$ be the sampled households. Let t_k^S be the symptom onset time of the index case k of

the household H_k. t_k^S is also the time at which the household H_k enters the sample. By assumption of our model, $t_k^S < \tau^1$ as this is the only way an household gets infected. The members of the households in the sample are followed-up until T_c (T_c as defined in 2.5.1). It is also possible to define the times T_c^k up to which the members of each household H_k are followed-up, but we use T_c for convenience such that $\max(T_c^k) = T_c$.

Figure 2.5: Timelines for an household sampled at t_k^S under case-ascertained follow-up. All the household members enter the sample at t_k^S, the symptom onset time of the index case in the household. The earliest and latest possible infection times for the index case are denoted by t' and t''.



● Symptom time of index case (t_k^S)
○ Symptom times of non-index cases
□ Start of follow-up
j
i
k
τ^0
t'
t''
t_k^S
t_i^S
T_C
l_max + s_max
l_min + s_min

The marginal, full and conditional likelihoods.
Let $t' = t_k^S - (l_{max} + s_{max})$ and $t'' = t_k^S - (l_{min} + s_{min})$ be the earliest and latest possible times of infection of the index case. Figure 2.5 illustrates the timelines for an household sampled at t_k^S. For an individual $i, i \in H_k, i \neq k$ the marginal likelihood $L_i^m = P(t_i^S > t_k^S)$ i.e., the probability that an household member i does not manifest symptoms before t_k^S, the symptom onset time of the index case. From the Figure 2.5 it is clear that for any household member i other than the index case to manifest symptoms by t_k^S must be infected earliest at t' and therefore would escape infection up to t'. For $t_i^S > t_k^S$ to hold, having escaped infection up to t', i would either be infected during $(t', t'']$ but manifest symptoms after t_k^S or be infected after t'' (which includes not being infected during the outbreak). Hereafter, the short-hand notations $\lambda_{ij}(u|\cdot) = \lambda_{ij}(u|\mathcal{F}_{iu-}, \mathcal{F}_{ju-})$ and $\mu(u|\cdot) = \mu(u|T^1)$ are used.

$$\mathrm{L}_i^m = \exp\Big\{ - \int\limits_{u=\tau^0}^{t'} \mu(u|\cdot)du \Big\} \Bigg[\int\limits_{t_i^E=t'}^{t''} \exp\Big\{ - \int\limits_{u=t'}^{t_i^E} \big(\mu(u|\cdot) + \textstyle\sum_{\substack{j\in H_k,\\ j\neq i}} \lambda_{ij}(u|\cdot)\big)du \Big\} \Big\{ \mu(t_i^E|\cdot) + \textstyle\sum_{\substack{j\in H_k,\\ j\neq i}} \lambda_{ij}(t_i^E|\cdot) \Big\}$$

$$\times \Bigg(1 - \int\limits_{t_i^S=t_i^I+s_{min}}^{t_k^S} \int\limits_{t_i^I=t_i^E+l_{min}}^{min(t_k^S,t_i^S,t_i^E+l_{max})} u(t_i^S|t_i^I)g(t_i^I|t_i^E)dt_i^I dt_i^S \Bigg) dt_i^E + \exp\Big\{ - \int\limits_{u=t'}^{t''} \big(\mu(u|\cdot) + \textstyle\sum_{\substack{j\in H_k,\\ j\neq i}} \lambda_{ij}(u|\cdot)\big)du \Big\} \Bigg]$$

The marginal likelihood L_i^m is of the form $Q_i(t')\mathcal{A}_i$ where $Q_i(t') = \exp\Big\{ - \int\limits_{u=\tau^0}^{t'} \big(\mu(u|\cdot) + \sum_{\substack{j\in H_k,\\ j\neq i}} \lambda_{ij}(u|\cdot)\big)du \Big\} =$ $\exp\Big\{ - \int\limits_{u=\tau^0}^{t'} \mu(u|\cdot)du \Big\}$ because there are no infected members within the household up to t'. On using $Z_i(\cdot)$ and $Q_i(\cdot)$ from equations 2.6 and 2.7, $\mathcal{A}_i$ is simply written as

$$\mathcal{A}_i = \int\limits_{t_i^E=t'}^{t''} Z_i(t_i^E|\mathcal{H}_{it_i^E-}) \Bigg(1 - \int\limits_{t_i^S=t_i^I+s_{min}}^{t_k^S} \int\limits_{t_i^I=t_i^E+l_{min}}^{min(t_k^S,t_i^S,t_i^E+l_{max})} u(t_i^S|t_i^I)g(t_i^I|t_i^E)dt_i^I dt_i^S \Bigg) dt_i^E + Q_i(t''|\mathcal{H}_{it''-}) \qquad (2.13)$$

The likelihood L_i is that given in equation (2.11) which, in this case, is called the *full likelihood.* This is due to L_i being written as if the individual i is followed-up from the start of outbreak without accounting for the sampling mechanism. For any non-index case household member $i, i \neq k$, we have $t_i^S > t_k^S$ with no infected household member up to t'. Therefore the full likelihood L_i can be factorised as $Q_i(t')\mathcal{B}_i$ where $\mathcal{B}_i$ is given as

$$\begin{cases} \int\limits_{t_i^I=t_i^E+l_{min}}^{min(t_i^S,t_i^E+l_{max})} u(t_i^S|t_i^I)g(t_i^I|t_i^E)\exp\Big\{-\int\limits_{u=t'}^{t_i^E}\big(\mu(u|\cdot)+\sum\limits_{\substack{j\in H_k,\\ j\neq i}}\lambda_{ij}(u|\cdot)\big)du\Big\}\Big\{\mu(t_i^E|\cdot)+\sum\limits_{\substack{j\in H_k,\\ j\neq i}}\lambda_{ij}(t_i^E|\cdot)\Big\}dt_i^I, & t_k^S<t_i^S\leq T_c, \\ \exp\Big\{-\int\limits_{u=t'}^{T_c}\big(\mu(u|\cdot)+\sum\limits_{\substack{j\in H_k,\\ j\neq i}}\lambda_{ij}(u|\cdot)\big)du\Big\}, & t_i^E>T_c. \end{cases} \tag{2.14}$$

Then the conditional likelihood for i is $L_i^c = L_i/L_i^m = \mathcal{B}_i/\mathcal{A}_i$ as $Q_i(t')$ cancels out. Further, the full and the marginal likelihoods for the index case are the same and is cancelled by conditioning. Thus the index-cases within the household do not have a likelihood contribution (Yang et al. (2006), Gordon et al (2018). This reduces the household level contribution of the likelihood to only non-index members.

2.5.4 Joint likelihood for all individuals under observation

The households are assumed to be independent under all the described observation designs. The complete-data likelihood is then the product of household-level likelihood contributions: $\prod_{k=1}^{K} L_{H_k}$.

Furthermore, in the two settings in Sections 2.5.1 and 2.5.2, each individual's contribution to the likelihood is conditioned on the history of all other household members in addition to the individual's own history. This allows the household-level contribution to likelihood to be re-organised as a product of individual likelihoods given the history of all other household members for each day (see §2.7 of Andersen et al. (1993) for a formal treatment and §6.5 of Davison (2003) for a heuristic treatment). Thus the likelihood is the product of individual-level contributions to the likelihood over the observation period:

$$L(\mu,\lambda;\mathcal{F}_{iT_c}, i=1,\ldots,N) = \prod_{i=1}^{N} L_i(\mu,\lambda;\mathcal{H}_{iT_c}). \tag{2.15}$$

In the case-ascertained follow-up design setting in Section 2.5.3, however, within the households, each individual's contribution to the likelihood is additionally conditioned on the index case, in particular, the earliest possible time of infection of the index case. Besides, the index cases in the sampled households do not contribute to the likelihood. These require additional book-keeping to be re-organised in the form equivalent to equation (2.15). Thus the likelihood for the sample of households $H_k, k = 1, 2, \ldots, K$, is simply given as the product of household-level contributions to the likelihoods $\prod_{k=1}^{K} L_{H_k}^c$, where

$$L_{H_k}^c = \prod_{i\in H_k} L_i^c = \prod_{i\in H_k} \frac{\mathcal{B}_i}{\mathcal{A}_i}. \tag{2.16}$$

2.6 Discrete-time model

Hitherto, the model has been defined in continuous time with two transmission hazards $\theta = (\mu, \lambda)$ as its parameters. In infinitesimal time intervals transmission hazards from multiple sources towards a susceptible individual are additive. Heuristically, this would mean that in a small time interval δt, the chance of more than one event occuring simultaneously is negligible i.e., $\lim_{\delta t\to 0} o(\delta t)/\delta t = 0$. This implies that the process is *orderly* (§6.5 of Davison, 2003).

In general, the time scale of the observed temporal data are in days. This, in addition to the computational ease of inference methods, leads towards specifying the model in discrete time. The discete-time model is governed by two transmission parameters $\theta = (c, q)$, which are the escape probabilities per day for a still susceptible individual from the environment and an infectious household member, respectively. Assuming that a still susceptible individual avoids infection from multiple sources independently each day, the escape probabilities are multiplicative. The discrete-time version of the model in Section 2.5.1 is provided in *Appendix A*.

Henceforth, the discrete-time model will be used for presenting the inference procedures, model comparison, simulation and numerical results. Some differences in implementation between the continuous-time and discrete-time models will mentioned later when necessary.

Table 2.1: Table of notations and their description

Component	Notation	Description
Outbreak timepoints		
	τ^0	the time at which the environment becomes infectious
	τ^1	the time up to which the environment is infectious
	τ_N	the time of removal of the last infectious individual
	T_c	observed symptom onset time of the last symptomatic case in completed outbreak
	T_s	serological sample collection time in completed outbreak
	T_c'	censoring time for an outbreak in progress
Natural history		
	t_i^E	time of infection of a susceptible individual i
	t_i^I	time of onset of infectiousness of an infected individual i
	t_i^S	time of onset of symptoms of an infected individual i
	t_i^N	time an infected individual i becoming non-infectious
	$\mathcal{F}_{it}$	history of the process for individual i up to time t
	$\mathcal{H}_{it}$	history of the process for an household with individual i (including i) up to time t
	$g(t_i^I \vert t_i^E)$	transition density for the time of onset of infectiousness (t_i^I) given the time of infection (t_i^E) with support (l_{min}, l_{max})
	$u(t_i^S \vert t_i^I)$	transition density for the time of onset of symptoms (t_i^S) given the time of onset of infectiousness (t_i^I) with support (s_{min}, s_{max})
	$v(t_i^N \vert t_i^S)$	transition density for the time of removal (t_i^N) given the time of onset of symptoms (t_i^S) with support (n_{min}, n_{max})
	$f(t_i^N \vert t_i^I)$	transition density for the time of removal (t_i^N) given the time of onset of infectiousness (t_i^I)
Transmission process		
- in continuous time	μ	infection hazard from the environment to a susceptible individual
	λ	transmission hazard from an infectious individual to a susceptible household member
- in discrete time	c	escape probability per day for a susceptible individual from the environment
	q	escape probability per day for a susceptible individual from an infectious household member
Observation process		
	η	sensitivity of symptoms for an infected individual
	ψ	specificity of symptoms for a non-infected individual
	R_{α,t_w}	probability of observing a positive serology given infection on a certain day
	α	rate of decrease in the detectability of IgM antibodies
	t_w	threshold duration from infection to onset of antibody waning

Chapter 3

Inference on Direct Transmission

> *"When the facts change, I change my opinion. What do you do, sir?"*
> \- John Maynard Keynes

This chapter outlines the procedures to estimate the transmission parameters under Bayesian inference for the discrete-time version of the model presented in Section 2.5.1 (see also Appendix A). The associated observed and complete data are reviewed. Some aspects of the likelihood such as the appropriate representation of data and missingness are discussed with reference to inference procedures. This chapter also explains in detail the estimation procedure based on Markov chain Monte Carlo methods.

3.1 Outbreak data

In this study, observed data are assumed to consist of symptom times and serological test results $\mathbf{y} = \{(m_i, t_i^S, a_i), i = 1, \ldots, N\}$. The missing data $\mathbf{z} = \{t_i^E, i = 1, \ldots, N\}$ are the infection times. Together, they form the complete data $\mathbf{C} = (\mathbf{z}, \mathbf{y}) = \{(t_i^E, m_i, t_i^S, a_i), i = 1, \ldots, N\}$.

Based on the four possible combinations of observed data (presence of symptoms (m_i) and serological test results (a_i)), six types of case status are defined. These case status and their corresponding complete data are presented in Table 3.1.

Table 3.1: Table presenting the four possible combinations of symptoms and serological data (observed) that define six types of case status. The complete data (C_i) associated with each case status are indicated. The symptom data (m_i, t_i^S) indicate presence of symptoms and symptom times; the serological data (a_i) indicate the serological test result at time T_s.

Symptoms (m_i, t_i^S)	Serology (a_i)	Case Status	Complete data (C_i)
(i) yes	positive	(1) symptomatic case	$(t_i^E < T_c, m_i = 1, t_i^S \leq T_c, a_i = 1)$
(ii) no	positive	(2) asymptomatic case	$(t_i^E < T_c, m_i = 0, t_i^S = \infty, a_i = 1)$
(iii) yes	negative	(3) symptomatic case (or)	$(t_i^E < T_c, m_i = 1, t_i^S \leq T_c, a_i = 0)$
		(4) non-case with false symptoms	$(t_i^E > T_c, m_i = 1, t_i^S \leq T_c, a_i = 0)$
(iv) no	negative	(5) asymptomatic case (or)	$(t_i^E < T_c, m_i = 0, t_i^S = \infty, a_i = 0)$
		(6) non-case	$(t_i^E > T_c, m_i = 0, t_i^S > T_c, a_i = 0)$

For the transmission parameters $\theta = (c, q)$ to be identifiable the following are assumed to be known: (i) the parameters that govern the observation model $(\eta, \psi, \alpha, t_w)$ and (ii) the sojourn time distributions ($\tilde{g}(t_i^I - t_i^E)$, $\tilde{u}(t_i^S - t_i^I)$, $\tilde{v}(t_i^N - t_i^S)$). The support for the sojourn time distributions (l_{min}, l_{max}), (s_{min}, s_{max}) and (n_{min}, n_{max}) have not been specified up to this point. The values pertaining to the natural history of hepatitis A virus infection from existing literature will be used. For the discrete-time model, we use discrete uniform distributions to model the sojourn time distributions.

3.2 Likelihood revisited

This section presents some general aspects of the likelihood representations corresponding to partially observed outbreak data and their relevance for inference procedures.

3.2.1 Structure of the likelihood

The underlying model for event times corresponding to partially observed outbreak is an *incompletely observed multi-state process*. When the process is Markovian, they are also known as *partially observed Markov processes* (or POMPs).

Equations (2.10) to (2.14) present the complete-data likelihood. It is of the form $P(\mathbf{z}, \mathbf{y}|\theta)$, where $\mathbf{y} = \{(m_i, t_i^S, a_i), i = 1, \ldots, N\}$, the symptom times and serology, are the observed data and $\mathbf{z} = \{t_i^E, i = 1, \ldots, N\}$, the infection times are the missing data items.

Three reprentations of the complete-data model can be formulated for the book-keeping of the evolving process as observed in either continuous or discrete time as follows.

(i) The population-level histories are described using number of individuals in each model state $\{\mathcal{X}(t), \tau^0 \leq t \leq T_c\}$ at a time point t during the outbreak, where $\mathcal{X}(t) = \{U(t), E(t), I(t), S(t), N(t)\}$. This representation is more straight-forward for state-transitions governed by Markov processes. This is because, semi-Markov processes (with duration-dependent transitions) would require to keep track of the age-of-infection data. Fintzi et al. (2017) call these compartment-wise trajectories as *lumped processes*.

Moreover, it is tedious to include any structures in the population such as households in this representation. The two following representations are amenable to including household structure within a population.

(ii) The household-level histories are described using $P_{u,e,i,s,n}(t)$, the proportion of households at time t with the numbers of individuals in each model state within the household as in Kinyanjui et al. (2016) and House & Keeling (2008).

(iii) The individual-level histories are described for each individual i in the population progressing through the model state space $\{X_i(t), \tau^0 \leq t \leq T_c\}$ at a time point t during the outbreak. This representation is used in the application presented in this thesis (see Section 2.2.1).

In general, the representation also concerns the choice of the latent variables that is introduced towards data-augmentation.

3.2.2 Likelihood in presence of missing data

When missing data (or latent variables) are present, the likelihood can be written in one of the following two forms based on the type of inference procedure and how missing values are treated. These descriptions closely follow Celeux et al. (2006).

Complete-data likelihood
When inference on the model parameters is easier if both the observed and missing data would be known, the joint likelihood of observed and missing data $P(\mathbf{z}, \mathbf{y}|\theta)$, also called the *complete-data likelihood* is specified. This is particularly suitable for methods like data-augmented MCMC and EM algorithms.

For the application presented in this thesis the following factorisation further clarifies the hierarchical dependence structure:

$$P(\mathbf{z}, \mathbf{y}|\theta, \gamma, \sigma) = P(\mathbf{y}|\mathbf{z}, \gamma, \sigma)P(\mathbf{z}|\theta, \sigma), \tag{3.1}$$

where $\theta = (c, q)$ is the set of parameters for the transmission model, $\gamma = (\eta, \psi, \alpha, t_w)$ is the set of parameters governing observation of symptoms and serology, and σ is the set of parameters governing the sojourn time distributions. Henceforward, γ and σ are assumed to be known and therefore dropped from the notation. The shorter notation $P(\mathbf{z}, \mathbf{y}|\theta) = P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\theta)$ will be used.

Observed-data likelihood
The missing data $\mathbf{z}$ may neither be useful for specifying and/or computing the likelihood nor be the focus of inference. In this case, the likelihood can be integrated over $\mathbf{z}$ (i.e., missing data) as $P(\mathbf{y}|\theta) = \int_{\mathbf{z}} P(\mathbf{z}, \mathbf{y}|\theta)d\mathbf{z}$. This form of the likelihood is called the *observed-data likelihood* or *integrated likelihood.* It is particularly suitable for simulation-based methods such as pseudo-marginal methods or particle filtering.

Celeux et al. (2006) describe a third form of likelihood called the *conditional likelihood* which is used when inference on the missing data is required and therefore $\mathbf{z}$ is considered as an additional unknown quantity. As this form is not relevant for the application in this thesis, it will not be discussed further.

The likelihood form, whether it is based on complete-data or observed-data model, has implications on the procedures of sampling from the posterior and model selection.

3.3 Inference in presence of latent variables

This section describes some general aspects of sampling-based procedures that are useful towards doing probabilistic inference in models with missing data or latent variables.

3.3.1 Posterior distribution

Having specified the likelihood, the next step under Bayesian inference is to obtain the posterior distribution of the model parameters θ given that the data $\mathbf{y}$ have been observed (see equation (1.3) and its description). In presence of latent variables or missing data $\mathbf{z}$, the likelihood is specified using either observed-data or complete-data form as described in Section 3.2.2. Then the posterior distribution is obtained in accordance with the likelihood specification.

When using the observed-data likelihood $P_{\mathbf{y}|\theta}(\mathbf{y}|\theta)$, the marginal posterior can be directly obtained as $P_{\theta|\mathbf{y}}(\theta|\mathbf{y}) \propto P_\theta(\theta)P_{\mathbf{y}|\theta}(\mathbf{y}|\theta)$. However, when the observed-data likelihood is not available in closed form, sampling from the marginal posterior is not straightforword. On the other hand, when using the complete-data likelihood $P_{\mathbf{z},\mathbf{y}|\theta}(\mathbf{z}, \mathbf{y}|\theta)$, the joint posterior $P_{\theta,\mathbf{z}|\mathbf{y}}(\theta, \mathbf{z}|\mathbf{y})$ can be obtained by sampling iteratively from its corresponding full conditional distributions (see Section 3.3.2). Henceforth the subscripts in the probability notation are dropped and will be used only when needed for clarity.

3.3.2 Data-augmented Markov chain Monte Carlo (DA-MCMC) method

Augmenting the posterior parameter space to include the missing data ($\mathbf{z}$) and simultaneously explore the joint posterior $P(\theta, \mathbf{z}|\mathbf{y})$ has been a popular choice due to the flexibility of its implementation (McKinley et al, 2014). A component-wise MCMC sampler such as the Gibbs sampler can be used when the full conditional distributions $P(\mathbf{z}|\mathbf{y}, \theta)$ and $P(\theta|\mathbf{y}, \mathbf{z})$ are available.

$P(\theta, \mathbf{z}|\mathbf{y})$ can be approximated by iteratively sampling $\mathbf{z}$ and θ from the following (Robert & Casella, 2000, Tanner & Wong, 2010):

$$\begin{array}{ll} \texttt{simulate} & \mathbf{z}^{(t+1)} \sim P(\mathbf{z}|\mathbf{y}, \theta^{(t)}) \ , \\ \texttt{simulate} & \theta^{(t+1)} \sim P(\theta|\mathbf{y}, \mathbf{z}^{(t+1)}). \end{array}$$

This procedure makes use of the complete-data likelihood form and can be seen, for example, from the full conditional distribution for updating θ: $P(\theta|\mathbf{y}, \mathbf{z}) \propto P(\theta)P(\mathbf{y}, \mathbf{z}|\theta)$.

The DA-MCMC method is implemented for the application in this thesis towards estimating the transmission parameters. The sampling procedures required for its implementation are elaborated in Section 3.4.

3.3.3 Pseudo-marginal methods

Also pseudo-marginal methods (Andrieu and Roberts, 2009) may be useful in the presence of latent variables. As opposed to sampling from the joint posterior $P(\theta, \mathbf{z}|\mathbf{y})$ using DA-MCMC, pseudo-marginal methods sample from the marginal posterior $\hat{P}(\theta|\mathbf{y})$ using a simulation-based approximation. McKinley et al. (2014) compare two such algorithms in the context of epidemic models: Monte Carlo within Metropolis algorithm (MCWM) by O'Neill et al. (2000) and grouped independence Metropolis-Hastings algorithm (GIMH) by Beaumont (2003).

Both algorithms (MCWM and GIMH) use a Metropolis-Hastings (M-H) sampler to draw samples from the marginal posterior $\hat{P}(\theta|\mathbf{y})$. Here the required (marginalised) likelihood ratio $\hat{R} = \hat{L}(\theta';\mathbf{y})/\hat{L}(\theta;\mathbf{y})$ within the M-H ratio is obtained by a Monte Carlo (MC) estimate using an importance sampling procedure:

$$\hat{L}(\theta;\mathbf{y}) = \frac{1}{M}\sum_{g=1}^{M}\frac{P(\mathbf{z}^{(g)},\mathbf{y}|\theta)}{P_{IS}(\mathbf{z}^{(g)},\mathbf{y}|\theta)}, \tag{3.2}$$

where $P_{IS}(\cdot)$ is the importance-sampling distribution and $\mathbf{z}^{(g)}$ is the gth random sample from $P_{IS}(\cdot)$. While MCWM computes $\hat{R}$ at every iteration of the M-H algorithm, GIMH re-uses $\hat{L}(\theta;\mathbf{y})$ from the previous iteration and only computes $\hat{L}(\theta';\mathbf{y})$ for the proposed value θ' at the current iteration. The pseudo-marginal methods thus use the observed-data likelihood form.

Other simulation-based methods such as sequential Monte Carlo (SMC) or iterated-filtering (IF) can be used in alternative model representations, for example, when the number of individuals in each state are modelled over the duration of the outbreak (see Section 3.2.1). All of the described methods require likelihood specification. In absence of a specified likelihood for the observed data, approximate Bayesian computation (ABC) methods are quite useful (Kypraios et al., 2017).

3.4 Sampling from the joint posterior

This section presents a sampling scheme based on DA-MCMC for the application presented in this thesis. A *component-wise Gibbs sampler* is used for the joint exploration of $P(\theta, \mathbf{z}|\mathbf{y})$ where the components $P(\mathbf{z}|\mathbf{y}, \theta)$ and $P(\theta|\mathbf{y}, \mathbf{z})$ are sampled using either a Gibbs step or a Metropolis-Hastings step. As mentioned in Section 2.6, the implementation of the inference scheme is for the discrete-time model with $\theta = (c, q)$ being escape probabilities per day.

The observed data are symptom times and serology $\mathbf{y} = \{(m_i, t_i^S, a_i), i = 1, \dots, N\}$. The missing data are infection times $\mathbf{z} = \{t_i^E, i = 1, \dots, N\}$. Together they form the complete data $(\mathbf{z}, \mathbf{y}) = \{(t_i^E, m_i, t_i^S, a_i), i = 1, \dots, N\}$.

3.4.1 Gibbs sampler for updating infection times

The generic set-up of updating the infection times

For an individual i infected during the outbreak with an observed positive serological test result ($a_i = 1$), the latent (i.e., missing) infection time t_i^E is sampled from $\mathcal{T}$, the set of candidate infection times (based on the observed symptom data), using the conditional probability

$$P(t_i^E = t|\mathcal{H}'_{it}, \theta) = \frac{\prod\limits_{\substack{j=1 \\ j\in H_i, j\neq i}}^{N} L_j(\theta; \mathcal{H}'_{iT_c}, t) L_i(\theta; \mathcal{H}'_{iT_c}, t)}{\sum\limits_{t_i^E\in\mathcal{T}} \prod\limits_{\substack{j=1 \\ j\in H_i, j\neq i}}^{N} L_j(\theta; \mathcal{H}'_{iT_c}, t_i^E) L_i(\theta; \mathcal{H}'_{iT_c}, t_i^E)}, \tag{3.3}$$

where L_i is the discrete-time version of the likelihood function provided in equation 2.10.

A note on notations: $\mathcal{H}_{it}$ is the generated history based on complete-data for all household members of individual i, including i, $\{(t_j^E, m_j, t_j^S, a_j), j \in H_i\}$ (see Chapter 2). Let $\mathcal{H}'_{it}$ denote the history based on

$\{(t_j^E, m_j, t_j^S, a_j), j \in H_i\} \setminus \{t_i^E\}$. Thus at each update, t_i^E is sampled from $\mathcal{T}$ according to the probability on the right hand side of (3.3). Also, in continuous time description, in absence of symptoms, the corresponding time of symptom onset is defined as $t_i^S = \infty$ (see Table 3.1). For practical purpose, in discrete time, this is defined as $t_i^S = T_c + 1$.

When the observed serological test result is negative ($a_i = 0$), the latent infection time is drawn from a 2-component mixture $X + (1 - X)P(t_i^E = t|\mathcal{H}'_{it}, \theta)$, where $X = P(t_i^E > T_c|\mathcal{H}'_{it}, \theta)$ is the probability that the individual was not infected during the outbreak. For individuals who were not infected during the outbreak, the infection times are set to $t_i^E > T_c$.

The set of candidate infection times ($\mathcal{T}$) for an individual i infected during the outbreak is defined by the observed symptom data (m_i, t_i^S). When the individual is symptomatic ($m_i = 1, t_i^S \leq T_c$), $\mathcal{T}$ is defined based on t_i^S. When the individual is asymptomatic ($m_i = 0, t_i^S = T_c+1$), $\mathcal{T}$ is defined by the entire duration of the outbreak.

In the discrete-time model, the number of candidate infection times is finite, i.e., $\mathcal{T}$ is a finite set. Thus the denominator (normalising constant) in (3.3) is a finite sum and the probabilites for all possible infection times ($\mathcal{T}$), given the parameters and observed values is computed. This allows a Gibbs update through direct sampling from the full conditional distribution.

In the simple case when $q = 1$, i.e., in the absence of person-to-person transmission, (3.3) reduces to

$$P(t_i^E = t|\mathcal{H}'_{it}, q, c) = \frac{L_i(\theta; \mathcal{H}'_{iT_c}, t)}{\sum\limits_{t_i^E \in \mathcal{T}} L_i(\theta; \mathcal{H}'_{iT_c}, t_i^E)}. \tag{3.4}$$

Updating the infection times for each combination of observed data
For each of the four combinations of observed data (see Table 3.1), the infection status and time is sampled using equation (3.3) as follows:

(i) When $y_i = (m_i = 1, t_i^S \leq T_c, a_i = 1)$, i.e, the time of symptom onset (t_i^S) is observed and the serological test is positive, t_i^E is drawn from the set candidate infection times $\mathcal{T} = \{t_i^S - (l_{max} + s_{max}), \ldots, t_i^S - (l_{min} + s_{min})\} = \{t'_i, ..., t''_i\}$ and (3.3) becomes

$$P(t_i^E = t|\mathcal{H}'_{it}, \theta) = \frac{\prod\limits_{\substack{j=1 \\ j \in H_i, j \neq i}}^{N} L_j(\theta; \mathcal{H}'_{iT_c}, t)\{\sum\limits_{t_i^I = t + l_{min}}^{t + l_{max}} \eta R(T_c - t) u(t_i^S|t_i^I) g(t_i^I|t) Z_i(t|\mathcal{H}'_{it-1})\}}{\sum\limits_{t_i^E = t'_i}^{t''_i} \prod\limits_{\substack{j=1 \\ j \in H_i, j \neq i}}^{N} L_j(\theta; \mathcal{H}'_{iT_c}, t_i^E)\{\sum\limits_{t_i^I = t_i^E + l_{min}}^{t_i^E + l_{max}} \eta R(T_c - t_i^E) u(t_i^S|t_i^I) g(t_i^I|t_i^E) Z_i(t_i^E|\mathcal{H}'_{it_i^E - 1})\}}. \tag{3.5}$$

(ii) When $y_i = (m_i = 0, t_i^S = T_c + 1, a_i = 1)$, i.e., the symptoms are not observed but the serological test is positive, then t_i^E is drawn from the set of candidate infection times $\mathcal{T} = 1, 2, \ldots, T_c$ and (3.3) becomes

$$P(t_i^E = t|\mathcal{H}'_{it}, \theta) = \frac{\prod\limits_{\substack{j=1 \\ j \in H_i, j \neq i}}^{N} L_j(\theta; \mathcal{H}'_{iT_c}, t)(1 - \eta) R(T_c - t) Z_i(t|\mathcal{H}'_{it-1})}{\sum\limits_{t_i^E = 0}^{T_c} \prod\limits_{\substack{j=1 \\ j \in H_i, j \neq i}}^{N} L_j(\theta; \mathcal{H}'_{iT_c}, t_i^E)(1 - \eta) R(T_c - t_i^E) Z_i(t_i^E|\mathcal{H}'_{it_i^E - 1})}. \tag{3.6}$$

(iii) When $y_i = (m_i = 1, t_i^S \leq T_c, a_i = 0)$, i.e, the time of symptom onset (t_i^S) is observed and the serological test is negative, the individual i could be either a symptomatic case or a non-case. If was infected during the outbreak and was symptomatic, the set of candidate infection times are $\mathcal{T} = \{t_i^S - (l_{max} + s_{max}), \ldots, t_i^S - (l_{min} + s_{min})\} = \{t'_i, ..., t''_i\}$. Thus the probability that individual i was infected and manifested symptoms at t_i^S but antibodies were not detectable at T_c is given by $x = x_1/(x_1 + x_2)$, where

$x_1 = \sum\limits_{t_i^E=t_i'}^{t_i''} \sum\limits_{t_i^I=t_i^E+l_{min}}^{t_i^E+l_{max}} \eta(1-R(T_c-t_i^E))u(t_i^S|t_i^I)g(t_i^I|t_i^E)Z_i(t_i^E|\mathcal{H}'_{it_i^E-1})$ and $x_2 = (1-\psi)\prod\limits_{u=1}^{T_c} e_i(u|\mathcal{H}_{iu-1})$.

A random variable $X \sim Bin(1,x)$ is first drawn. If $X=1$, t_i^E is drawn from $\mathcal{T}$ with the following conditional probability:

$$P(t_i^E = t|\mathcal{H}'_{it},\theta) = \frac{\prod\limits_{\substack{j=1\\ j\in H_i, j\neq i}}^{N} L_j(\theta;\mathcal{H}'_{iT_c},t)\{\sum\limits_{t_i^I=t+l_{min}}^{t+l_{max}} \eta(1-R(T_c-t))u(t_i^S|t_i^I)g(t_i^I|t)Z_i(t|\mathcal{H}'_{it-1})\}}{\sum\limits_{t_i^E=t_i'}^{t_i''}\prod\limits_{\substack{j=1\\ j\in H_i, j\neq i}}^{N} L_j(\theta;\mathcal{H}'_{iT_c},t_i^E)\{\sum\limits_{t_i^I=t_i^E+l_{min}}^{t_i^E+l_{max}} \eta(1-R(T_c-t_i^E))u(t_i^S|t_i^I)g(t_i^I|t_i^E)Z_i(t_i^E|\mathcal{H}'_{it_i^E-1})\}}. \tag{3.7}$$

If $X=0$, then, the individual was not infected during the outbreak and $t_i^E > T_c$.

(iv) When $y_i = (m_i = 0, t_i^S = T_c+1, a_i = 0)$, i.e, when the symptoms are not observed and the serological test is negative, the individual i could be either an asymptomatic case or a non-case. If i was infected during the outbreak and was asymptomatic, the set of candidate infection times are $\mathcal{T} = 1,2,\ldots,T_c$. Thus the probability that individual i was infected, but was neither symptomatic nor antibodies for infection were detected at T_c is given by $x = x_1/(x_1+x_2)$, where $x_1 = \sum\limits_{t_i^E=1}^{T_c}(1-\eta)(1-R(T_c-t_i^E))Z_i(t_i^E|\mathcal{H}'_{it_i^E-1})$ and $x_2 = \psi\prod\limits_{u=1}^{T_c} e_i(u|\mathcal{H}_{iu-1})$.

A random variable $X \sim Bin(1,x)$ is first drawn. If $X=1$, t_i^E is drawn from $\mathcal{T}$ with the following conditional probability:

$$P(t_i^E = t|\mathcal{H}'_{it},\theta) = \frac{\prod\limits_{\substack{j=1\\ j\in H_i}}^{N} L_j(\theta;\mathcal{H}'_{iT_c},t)(1-\eta)(1-R(Tc-t))Z_i(t|\mathcal{H}'_{it-1})}{\sum\limits_{t_i^E=0}^{T_c}\prod\limits_{\substack{j=1\\ j\in H_i}}^{N} L_j(\theta;\mathcal{H}'_{iT_c},t_i^E)(1-\eta)(1-R(T_c-t_i^E))Z_i(t_i^E|\mathcal{H}'_{it_i^E-1})}. \tag{3.8}$$

If $X=0$, then, the individual was not infected during the outbreak and $t_i^E > T_c$.

3.4.2 Metropolis-Hastings sampler for updating transmission parameters

Parameters ($\theta = (c,q)$) are sampled using a random-walk Metropolis-Hastings algorithm. The proposed value θ' is accepted with probability $A(\cdot)$ such that

$$A(\theta',\theta) = \min\left\{1, \frac{P_\theta(\theta')L(\theta';\mathcal{H}_{T_c})k(\theta|\theta')}{P_\theta(\theta)L(\theta;\mathcal{H}_{T_c})k(\theta'|\theta)}\right\} \tag{3.9}$$

where $k(\cdot)$ is the proposal distribution, $P_\theta(\cdot)$ is the prior distribution for θ and $L(\theta;\mathcal{H}_{T_c})$ is the complete-data likelihood for the population which is the discrete-time version of (2.15). The generated history based on complete-data for the population is $\mathcal{H}_{T_c} = \bigcup\limits_{i=1}^{N}\mathcal{H}_{iT_c}$.

3.4.3 Prior distributions

In the Bayesian inference framework, a prior density $P_\theta(\theta)$ must be specified for the transmission parameters $\theta = (c,q)$ and will be included in the Metropolis-Hastings sampler as indicated in equation (3.9). As c and q are probabilities, their marginal prior densities can be specified in the form of beta distributions. For inference from an observed outbreak data, informative priors can be specified through providing values for hyperparameters of suitable beta distributions for c and q.

Note that the beta distributions for marginal prior densities of the transmission parameters are only a suggestion as they do not offer any advantage with respect to the functional form of the complete-data likelihood for the population in discrete-time (see Section 3.4.5). Other distributions can also be used, especially in conjunction with any transformation when sampling θ (such as the logit transformation for sampling probability parameters in O'Neill et al., 2000).

3.4.4 Additional remarks

Updating infection times involves changing, removing and adding infection times. For the combinations of observed data (i) and (ii) in Table 3.1, the corresponding sampling schemes to update infection times in Section 3.4.1 will result only in change of infection times in consecutive MCMC samples. However, for the combinations of observed data (iii) and (iv), updating an infection time may result in adding or removing an infection time (an non-case becoming an infected individual or vice-versa) in consecutive MCMC samples. These moves results in a trans-dimensional MCMC. However, in the discrete-time model, the state-space is finite allowing a Gibbs update through direct sampling from the full conditional distribution.

When $\eta = \psi = 1$, the infection status of all individuals are fully known from the observed data, which results in fixed dimensionality of the MCMC procedure. In this case, devising a household-wise update scheme for infection times is straightforword. A blocked Gibbs update can be constructed similar to (3.3) by jointly sampling the infection times for all household members. For the J household members belonging to an household H_k, the conditional probability infection times $t_j^E, j \in H_k$ is

$$P(\{t_j^E, j \in H_k\}|\{y_j, j \in H_k\}, \theta) = \frac{\prod_{j=1}^{J} L_j(\theta; \mathcal{H}_{jT_c})}{\sum_{t_1^E \in \mathcal{T}_1} \cdots \sum_{t_J^E \in \mathcal{T}_J} \prod_{j=1}^{J} L_j(\theta; \mathcal{H}_{jT_c})}. \tag{3.10}$$

However, it should be noted that the expressions for larger household sizes are more cumbersome and computationally less efficient as compared to updating individual infection times.

3.4.5 Sampling from the joint posterior in the continuous-time model

Gibbs sampler for updating parameters
When the model is specified in continuous time, the parameters $\theta = (\mu, \lambda)$ can be sampled using a Gibbs sampler. This is because the parameters govern Poisson processes and can be provided with suitable gamma priors. Such conditional densities can be directly sampled from the appropriate gamma density by exploiting the Poisson-gamma conjugacy (O'Neill & Roberts, 1999, Streftaris & Gibson, 2004).

Metropolis-Hastings sampler for updating infection times
In discrete-time model the probabilities for sampling from candidate infection times are well-defined. In continuous-time model the integral in the denominator which is the normalising constant would be tedious to compute exactly. Thus a M-H sampler can be constructed for drawing infection times from the unnormalised density.

As pointed out in Section 3.4.4, updating infection times involves changing, removing and adding infection times resulting in a trans-dimensional MCMC. Thus, in the continuous-time model, Jacobians associated with trans-dimensional moves must be incorporated into the computation of M-H ratio (Forrester et al., 2007, O'Neill & Roberts, 1999). These computations are more tedious and therefore the dicrete-time model is better suited towards computational efficiency.

3.5 Computation

3.5.1 Numerical implementation

The following pseudocode is provided to implement the DA-MCMC procedure for an observed (or simulated) dataset using a computing program such as `R`. In order to implement (handle memory storage and computing time) efficiently, one can write the demanding parts of code in a language such as `C++`.

Table 3.2: Pseudocode for sampling from the joint posterior using DA-MCMC

Sampling from the joint posterior

```
Initialise  $\theta^{(1)} = (q^{(1)}, c^{(1)})$;
Initialise  $\{t_i^{E(1)}, i = 1, \ldots, N\}$
for m in 2 to M (the number of MCMC samples)
  for i in 1 to N
   sample $t_i^{E(m)} \sim P(T_i^{E(m)}|\theta^{(m-1)}, \mathbf{t}_{-\mathbf{i}}^{\mathbf{E(m_i)}}, m_i, t_i^S, a_i)$
   where $\mathbf{t}_{-\mathbf{i}}^{\mathbf{E(m_i)}} = \{t_1^{E(m)}, \ldots, t_{i-1}^{E(m)}, t_{i+1}^{E(m-1)}, \ldots, t_N^{E(m-1)}, \}$
 end
  sample $\theta' \sim k(\theta'|\theta^{(m-1)})$ from the proposal
  accept $\theta'$ with probability $A(\theta', \theta^{(m-1)}) = \min\left\{1, \frac{P_\theta(\theta')L(\theta';\cdot)k(\theta^{(m-1)}|\theta')}{P_\theta(\theta^{(m-1)})L(\theta^{(m-1)};\cdot)k(\theta'|\theta^{(m-1)})}\right\}$ and $\{\theta^{(m)} \leftarrow \theta'\}$
  otherwise $\{\theta^{(m)} \leftarrow \theta^{(m-1)}\}$
 end
Compute the marginal posterior estimates $\hat{\theta}(\mathbf{y}) = E[\theta|\mathbf{y}]$  using $\{\theta^{(m)}, m = 1, \ldots, M\}$.
```

3.5.2 Estimating the transmission parameters

Having obtained the joint posterior samples $\{(\theta^{(m)}, \mathbf{z}^{(m)}), m = 1, \ldots, M\}$, the required point and interval estimates $\hat{\theta}$ and $h(\hat{\theta})$ for the parameter θ can be directly computed. For a unimodal marginal posterior of the parameters $P(\theta|\mathbf{y})$, the maximum a posteriori MAP(θ) = arg max$_\theta P(\theta|\mathbf{y})$ can be computed. Similarly, the credible intervals $h(\hat{\theta})$ such as the percentile interval or the highest posterior density (HPD) interval can be computed using the posterior samples.

The procedures described up to this point are sufficient for estimating the transmission parameters $\theta = (c, q)$ when the parameters governing the observation model and the sojourn time distributions are known (see Section 3.1). When any of these quantities are unavailable and are required to be jointly estimated with the transmission parameters, additional procedures need to be developed (see Sections 5.3 to 5.5 for further discussion).

Chapter 4

Model Comparison

"We balance probabilities and choose the most likely. It is the scientific use of the imagination."
- Sir Arthur Conan Doyle (Sherlock Holmes in The Hound of Baskervilles)

Chapter 3 presents methods for Bayesian estimation of the transmission parameters through sampling from the joint posterior of the parameters and missing data. However, to answer the question "is there person-to-person transmission?" from some observed outbreak data using the model of Section 2.5.1 may require more formal methods of hypothesis testing. In this chapter, this problem of hypothesis testing is cast as a Bayesian model comparison problem and appropriate computational methods are developed. These methods utilise the sampling algorithms from Section 3.4.

The null hypothesis is the absence of person-to-person transmission i.e., that an infected individual cannot infect others. This corresponds to a null model where the transmission hazard (λ) of a susceptible individual from an infectious household member is zero or the avoidance probability (q) is one. The alternative hypothesis is the presence of person-to-person transmission and corresponds to a full model where the transmission hazard (λ) of a susceptible individual from an infectious household member is positive or the avoidance probability (q) is less than one. It is to be noted that the parameter for person-to-person transmission lies in the boundary of the parameter space for the null hypothesis.

4.1 Model comparison for partially observed stochastic epidemic models

Three approaches to compare stochastic epidemic models with partially observed data have been used in the Bayesian setting: information criteria, Bayes factors and methods based on the posterior predictive distribution (Alharthi et al, 2018). This section presents, in general, some procedures for the first two approaches.

Very often in settings like the issue dealt with in this thesis, the missing data $\mathbf{z}$ include unobserved infection times and events. In some cases, the number of infections may not be exactly known based on the observed data (Table 3.1 provides six case status from four combinations of observed data). When the parameter space is augmented to include the missing data as in DA-MCMC, the unknown number of infections imply that the number of parameters is not well-defined.

Model comparison methods need to account for the number of parameters whether information criteria (which include concepts of model fit and complexity) is used or model comparison is directly based on the posterior probability such as Bayes factors.

4.1.1 Model comparison using information criteria

In general, for competing parametric models $P(\mathbf{y}|\theta)$, model selection using information criteria are based on deviance $D(\theta) = -2\log P(\mathbf{y}|\theta) + 2\log h(\mathbf{y})$. Here $h(\mathbf{y})$ is a function of data alone (a standardising term) that can be set as $h(\mathbf{y}) = 1$ for the purpose of model comparison (Celeux et al., 2006).

Information based criteria provide information loss on using a certain model $P(\mathbf{y}|\theta)$ with reference to an ideal (or a theoretically true) model. The measure is directly based on deviance along with a penalty for the number of parameters in the model. Thus both model fit and model complexity are addressed.

Akaike information criteria (AIC) is given by $-2\log P(\mathbf{y}|\hat{\theta}) + 2k = D(\hat{\theta}) + 2k$ where k is the number of parameters in the model and $\hat{\theta}$ is a posterior point estimate, for example, $\hat{\theta} = \text{MAP}(\theta)$. The difference in AIC quantifies the relative information loss between the compared models. When the number of parameters is not exactly known, as in the case where the number of infections is not exactly known when the epidemic is partially observed, AIC cannot be used.

Deviance information criteria (DIC) overcomes the problem of unknown number of parameters by calculating an effective number of parameters (p_D). Following Spiegelhalter et al. (2002) and Celeux et al. (2006), the effective number of parameters is defined as the difference between the posterior mean deviance ($\overline{D(\theta)} = E_\theta[-2\log P(\mathbf{y}|\theta)|\mathbf{y}]$) and the deviance at the posterior mean ($D(\hat{\theta})$) i.e., $p_D = \overline{D(\theta)} - D(\hat{\theta})$. Here, $\hat{\theta}$ is the posterior mean ($E[\theta|\mathbf{y}]$), but other estimates such as MAP or posterior median can also be used in practise.

Then in a similar vein to AIC, the DIC can then be given as a measure of fit and complexity using the deviance at posterior mean and effective number of parameters as follows:

$$\text{DIC} = D(\hat{\theta}) + 2p_D = 2\overline{D(\theta)} - D(\hat{\theta}). \tag{4.1}$$

DIC in the presence of missing data is presented in Section 4.2.

4.1.2 Bayes factors

Bayes factors (BF) have been used in epidemic modelling for model comparison (Cauchemez et al., 2004, Knock & O'Neill, 2014, Alharthi et al, 2018). For K competing models $\{\mathcal{M}_k, k = 1, \ldots, K\}$ with associated parameters θ_k, computing the marginal likelihood $p(y|\mathcal{M}_k) = \int p(\theta_k)p(\mathbf{y}|\theta_k)d\theta_k$ is crucial for calculating BF. This computation is further complicated by partially observed data.

The most common method to calculate Bayes factors is by using the reversible jump MCMC (RJ-MCMC) approach (Alharthi et al., 2018). When computing the required marginal likelihoods, RJ-MCMC explores the union of parameter spaces defined by the K competing models (Gibson et al., 2018). However, as pointed out in Section 3.4, RJ-MCMC approach is already required for partially observed data with unknown number of infections even for a single model. Adding a layer of models will increase the number of trans-dimensional moves across competing models and may render such computation infeasible (Gibson et al., 2018). Alharthi et al. (2018) provide a power posterior approach to BF computation in the SIR model setting with partial observation.

4.2 DIC for partially observed stochastic epidemic models

This section considers some general aspects of DIC for missing data models with refernce to epidemic modelling. Multiple implementations of DIC for missing data exist based on observed and complete-data likelihood forms (Celeux et al., 2006). In epidemic modelling with partial observations, DIC based on complete-data likelihood is generally used because the observed data likelihood is not available in closed form (Gibson et al., 2018). This is due to multiple integrals being involved in the observed data likelihood thereby rendering it analytically intractable. Furthermore, the choice depends on where the focus of inference is placed and how the missing data are treated.

Three implementations of DIC for complete-data likelihood (DIC$_4$, DIC$_5$ and DIC$_6$) are presented in Celeux et al. (2006), two of which (DIC$_4$ and DIC$_6$) have been applied towards MCMC-based inference in epidemic modelling (Gibson et al., 2018). On computing DIC using the expression $2\overline{D(\theta)} - D(\hat{\theta})$, the posterior mean deviance $\overline{D(\theta)} = -2E_{\theta,\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\theta)|\mathbf{y})]$ is identical in all threee implementations. But the deviance at the posterior mean $D(\hat{\theta})$ is interpreted based on the focus of inference, i.e., how the missing data $\mathbf{z}$ are treated.

While in DIC$_5$, the missing data $\mathbf{z}$ are treated as additional parameters, in DIC$_4$ and DIC$_6$ they are treated as nuisance quantities. Thus they are marginalised by taking expectations of the log likelihood with respect to $\mathbf{z}$ in the calculation of $D(\hat{\theta})$. Table 4.1 presents the computations involved in $D(\hat{\theta})$ for DIC$_4$, DIC$_5$ and DIC$_6$.

In DIC$_4$, the deviance at the posterior mean is calculated as $D(\hat{\theta}) = -2E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|E_\theta[\theta|\mathbf{y}, \mathbf{z}])|\mathbf{y})]$. Here the posterior expectation $E_\theta[\theta|\mathbf{y}, \mathbf{z}]$ is a complete data estimator and is computed for each value of $\mathbf{z}$ (Celeux et al., 2006). The information provided by $\mathbf{z}$ towards θ is utilised through the use of a complete-data estimator here.

In DIC$_6$, $D(\hat{\theta}) = -2E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y}))]$ where $\hat{\theta}(\mathbf{y})$ is based on the marginal MAP estimator, thus ignoring the additional information provided by $\mathbf{z}$ (Gibson et al., 2018). Thus the missing data $\mathbf{z}$ is not the main focus of inference in DIC$_6$ (Alharthi et al., 2018). The computation of DIC$_6$ for the application in this thesis is provided in detail in Section 4.3.

In DIC$_5$, the deviance at the posterior mean is calculated as $D(\hat{\theta}) = -2\log P(\mathbf{y}, \hat{\mathbf{z}}(\mathbf{y})|\hat{\theta}(\mathbf{y}))$. Here the posterior estimates $\big(\hat{\theta}(\mathbf{y}), \hat{\mathbf{z}}(\mathbf{y})\big)$ are directly plugged-in from the joint estimator (Celeux et al., 2006).

Table 4.1: The deviance information criteria (DIC) based on complete data likelihood: implementation of the deviance at posterior mean based on focus of inference and their computations

DIC	Estimator of posterior mean	$D(\hat{\theta})$: Deviance at posterior mean	Calculation of $D(\hat{\theta})$
DIC$_4$	$\hat{\theta} = E_\theta[\theta\|\mathbf{y}, \mathbf{z}]$ complete data estimator	$-2E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}\|E_\theta[\theta\|\mathbf{y}, \mathbf{z}])\|\mathbf{y})]$	Samples of $\mathbf{z}$ are used from the joint posterior $\{\mathbf{z}^{(m)} \sim P(\mathbf{z}\|\theta, \mathbf{y}), m = 1, \ldots, M\}$. For each sampled value $\mathbf{z}^{(m)}$, the complete-data estimates $E_\theta[\theta\|\mathbf{y}, \mathbf{z}^{(m)}]$ are computed either exactly (if available) or by drawing further samples for $\{\theta^{(l)} \sim P(\theta\|\mathbf{z}^{(m)}, \mathbf{y}), l = 1, \ldots, L\}$ towards the calculation of $D(\hat{\theta})$. [1]
DIC$_6$	$\hat{\theta} = E_\theta[\theta\|\mathbf{y}]$ marginal estimator	$-2E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}\|\hat{\theta}(\mathbf{y}))\|\mathbf{y}, \hat{\theta}(\mathbf{y}))]$	Using the marginal estimate $\hat{\theta}$ from the joint posterior, samples for $\mathbf{z} \sim P(\mathbf{z}\|\mathbf{y}, \hat{\theta})$ are drawn towards the calculation of $D(\hat{\theta})$. [1, 2]
DIC$_5$	$\big(\hat{\theta}(\mathbf{y}), \hat{\mathbf{z}}(\mathbf{y})\big)$ joint estimator	$-2\log P(\mathbf{y}, \hat{\mathbf{z}}(\mathbf{y})\|\hat{\theta}(\mathbf{y}))$	The joint estimates $\big(\hat{\theta}(\mathbf{y}), \hat{\mathbf{z}}(\mathbf{y})\big)$ from the joint posterior samples are plugged-in and no further computations are required towards the calculation of $D(\hat{\theta})$. [1]

[1] - Celeux et al. (2006), [2] - Alharthi et al. (2018)

4.3 Model comparison for the household transmission parameter

This section presents the implementation of model comparison using DIC$_6$ for the application in this thesis.

4.3.1 Hypotheses and models

The two models are based on the hypotheses $H_0 : q = 1$ vs. $H_1 : q < 1$. In the continuous-time model, these hypotheses are equivalent to $H_0 : \lambda = 0$ vs. $H_1 : \lambda > 0$.

The model for infection $P(\mathbf{z}|\theta)$ pertaining to the null model is defined in discrete-time as

$$P(t_i^E|\theta) = \begin{cases} Z_i(t_i^E|\cdot) = (1 - c(t_i^E|\tau^1)) \prod_{u=1}^{t_i^E - 1} c(u|\tau^1) & \text{if } i \text{ is infected,} \\ Q(T_c) = \prod_{u=1}^{T_c} c(u|\tau^1) & \text{if } i \text{ is not infected.} \end{cases} \tag{4.2}$$

and in continuous time as

$$P(t_i^E|\theta) = \begin{cases} Z_i(t_i^E|\cdot) = \mu(t_i^E|\tau^1)\exp\left\{ -\int\limits_{u=0}^{t_i^E} \mu(u|\tau^1)du \right\} & \text{if } i \text{ is infected,} \\ Q(T_c) = \exp\left\{ -\int\limits_{u=0}^{T_c} \mu(u|\tau^1)t_i^E du \right\} & \text{if } i \text{ is not infected.} \end{cases} \tag{4.3}$$

The null models are of the complete-data likelihood form $P(\mathbf{z}, \mathbf{y}|\theta) = P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\theta)$, where $P(\mathbf{z}|\theta)$ is defined as in (4.2) or (4.3) and $P(\mathbf{y}|\mathbf{z})$ pertains to the observation model (see equation (3.1) for a short hand notation and also equations (2.10) and (A.8) for the complete-data likelihood corresponding to the full model in continuous time and in discrete time, respectively).

4.3.2 Implementing DIC$_6$

For the application in this thesis, due to the non-availability of observed data likelihood in closed form, DIC based on the complete data likelihood $P(\mathbf{z}, \mathbf{y}|\theta)$ is used for model comparison. Furthermore, as the focus of inference is only on the transmission parameters $\theta = (c, q)$ and not on the missing data $\mathbf{z}$, DIC$_6$ is considered most appropriate here. It is also easier to compute than DIC$_4$ as discussed earlier. Based on Celeux et al. (2006), DIC$_6$ is computed as follows:

$$\text{DIC}_6 = 2\overline{D(\theta)} - D(\hat{\theta}) = -4E_{\theta,\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\theta)|\mathbf{y})] + 2E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y}))]. \tag{4.4}$$

The expectation in the first term, $E_{\theta,\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\theta)|\mathbf{y})]$ is obtained directly from the DA-MCMC algorithm presented in Section 3.4. Additionally, $\hat{\theta}(\mathbf{y}) = E_{\theta|\mathbf{y}}(\theta|\mathbf{y})$ is evaluated using the same DA-MCMC run (Celeux et al., 2006, Alharthi et al., 2018).

The expectation in the second term $E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y}))]$ is computed by sampling $\mathbf{z}$ from $P(\mathbf{z}|\mathbf{y}, \theta)$ by setting $\theta = \hat{\theta}(\mathbf{y})$. The following describes the computation of this expectation.

Note that for small household sizes and in case of fully symptomatic infections (where the number of candidate infection times is small) the required expectation can be computed analytically. For an household H_k with J members we have

$$E_{\mathbf{z}|\mathbf{y},\hat{\theta}}[\log(P_{H_k}(\mathbf{y}, \mathbf{z}|\hat{\theta}))] = \int \log\left[P_{H_k}(\mathbf{y}, \mathbf{z}|\hat{\theta})\right] P_{H_k}(\mathbf{z}|\mathbf{y}, \hat{\theta})d\mathbf{z} = \sum_{t_1^E \in \mathcal{T}_1} \cdots \sum_{t_J^E \in \mathcal{T}_J} \log\left[\prod_{j=1}^{J} L_j(\hat{\theta}; \mathcal{H}_{jT_c})\right] P_{H_k}(\mathbf{z}|\mathbf{y}, \hat{\theta}), \tag{4.5}$$

where $P_{H_k}(\mathbf{z}|\mathbf{y}, \hat{\theta}) = P(\mathbf{z} = \{t_j^E, j \in H_k\}|\{y_j, j \in H_k\}, \hat{\theta})$ is calculated using equation (3.10). The expectation for the population of K households is given by

$$E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y}))] = \sum_{k=1}^{K} E_{\mathbf{z}|\mathbf{y},\hat{\theta}}[\log(P_{H_k}(\mathbf{y}, \mathbf{z}|\hat{\theta}))]. \tag{4.6}$$

For large household sizes and/or in the presence of asymptomatic infections such analytical integration is tedious to compute. In such cases the required expectation is computed empirically as a Monte Carlo integral. The Gibbs sampling procedure for the data-augmentation step which draws $\mathbf{z}$ from $P(\mathbf{z}|\mathbf{y}, \hat{\theta})$ can be re-used. The required expectation is computed as

$$E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y}))] = \int \log\left[P(\mathbf{y}, \mathbf{z}|\hat{\theta})\right] P(\mathbf{z}|\mathbf{y}, \hat{\theta})d\mathbf{z}. \tag{4.7}$$

Using the DA-MCMC implementation from Section 3.4.1, $\mathbf{z}^{(g)}$ is drawn from $P(\mathbf{z}|\mathbf{y}, \hat{\theta})$, where $\mathbf{z}^{(g)} = \{t_i^E, i = 1, \ldots, N\}^{(g)}$ are sample draws of infection times for all individuals in the population. Then a simulation consistent estimate of $E_{\mathbf{z}}[\log(P(\mathbf{y}, \mathbf{z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y}))]$ is given by

$$E_{\mathbf{z}}[\log(P(\mathbf{y},\mathbf{z}|\hat{\theta}(\mathbf{y}))|\mathbf{y},\hat{\theta}(\mathbf{y}))] = \frac{1}{M'}\sum_{g=1}^{M'}\log\left[P(\mathbf{y},\mathbf{z}^{(g)}|\hat{\theta})\right], \tag{4.8}$$

where M' is the number of samples used for computing the required expectation.

4.4 Computation

This section presents the computational aspects in calculating DIC$_6$ and a pseudocode is provided for calculating all the necessary components. The algorithm to compute DIC$_6$ is illustrated in Figure 4.1. Some directions are provided to compute the quantity for the corresponding full and null models and on their interpretation.

4.4.1 Numerical Implementation

The following pseudocode is provided to compute DIC$_6$ for an observed (or simulated) dataset using a computing program such as `R`. In order to implement efficiently (handle memory storage and computing time), one can write the demanding parts of code in a language such as `C++`.

Table 4.2: Pseudocode for computing DIC$_6$

Sampling from the joint posterior

```
Initialise  $\theta^{(1)} = (q^{(1)}, c^{(1)})$;
Initialise  $\{t_i^{E(1)}, i = 1, \ldots, N\}$
for m in 2 to M (the number of MCMC samples)
  for i in 1 to N
   sample $t_i^{E(m)} \sim P(T_i^{E(m)}|\theta^{(m-1)}, \mathbf{t}_{-\mathbf{i}}^{\mathbf{E(m_i)}}, m_i, t_i^S, a_i)$
   where $\mathbf{t}_{-\mathbf{i}}^{\mathbf{E(m_i)}} = \{t_1^{E(m)}, \ldots, t_{i-1}^{E(m)}, t_{i+1}^{E(m-1)}, \ldots, t_N^{E(m-1)}, \}$
 end
  sample $\theta' \sim k(\theta'|\theta^{(m-1)})$ from the proposal
  accept $\theta'$ with probability $A(\theta', \theta^{(m-1)}) = \min\left\{1, \frac{P_\theta(\theta')L(\theta';\cdot)k(\theta^{(m-1)}|\theta')}{P_\theta(\theta^{(m-1)})L(\theta^{(m-1)};\cdot)k(\theta'|\theta^{(m-1)})}\right\}$ and $\{\theta^{(m)} \leftarrow \theta'\}$
  otherwise $\{\theta^{(m)} \leftarrow \theta^{(m-1)}\}$
 end
Compute $\hat{\theta}(\mathbf{y}) = E[\theta|\mathbf{y}]$  using $\{\theta^{(m)}, m = 1, \ldots, M\}$.
```

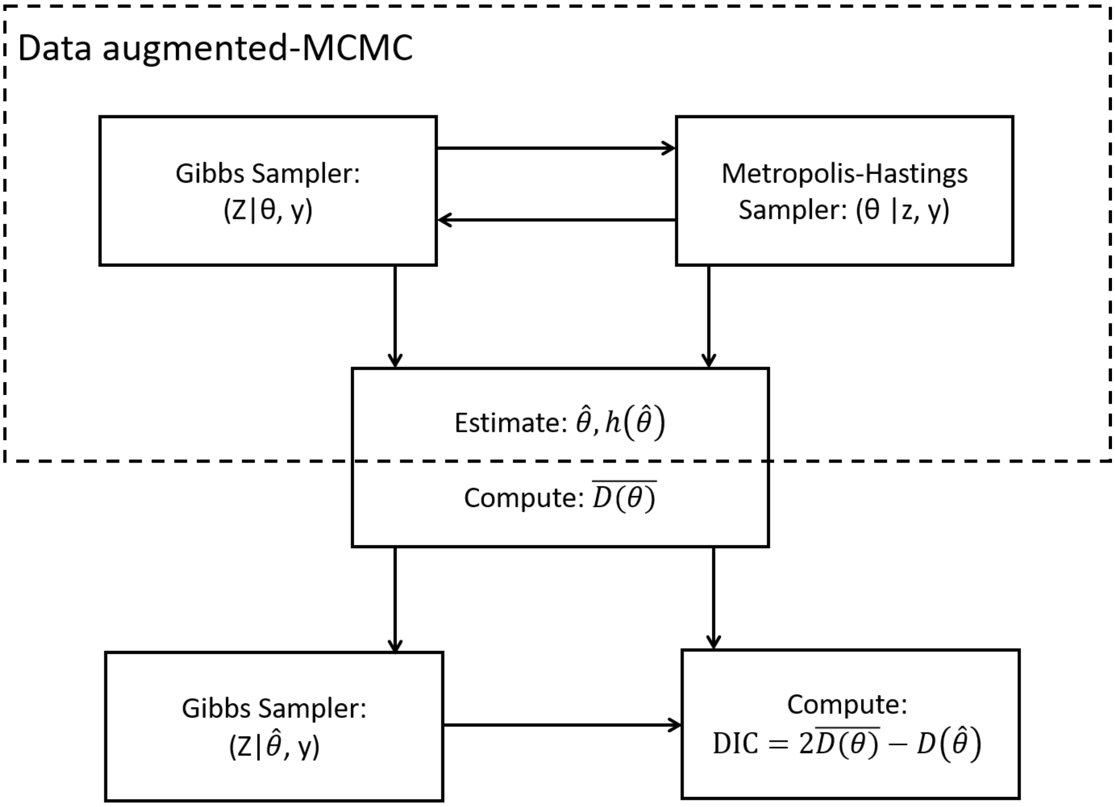
Computing DIC$_6$

```
Compute $\overline{D(\theta)} = -2E_{\theta,\mathbf{z}|\mathbf{y}}[\log P(\mathbf{y},\mathbf{z}|\theta)]$  using $\{(\theta^{(m)}, \mathbf{t}^{\mathbf{E(m)}}), m = 1, \ldots, M\}$.
 Initialise  $\{t_i^{E(1)}, i = 1, \ldots, N|\hat{\theta}\}$
 for g in 2 to M' (the number of MC samples)
   for i in 1 to N
      sample $t_i^{E(g)} \sim P(T_i^{E(g)}|\hat{\theta}, \mathbf{t}_{-\mathbf{i}}^{\mathbf{E(g_i)}}, m_i, t_i^S, a_i)$
      where $\mathbf{t}_{-\mathbf{i}}^{\mathbf{E(g_i)}} = \{t_1^{E(g)}, \ldots, t_{i-1}^{E(g)}, t_{i+1}^{E(g-1)}, \ldots, t_N^{E(g-1)}, \}$
   end
  end
Compute $E_{\mathbf{z}|\mathbf{y},\hat{\theta}}[\log(P(\mathbf{y},\mathbf{z}|\hat{\theta}))]$ as $\frac{1}{M'}\sum_{g=1}^{M'}\log\left[P(\mathbf{y},\mathbf{z}^{(g)}|\hat{\theta})\right]$ using the
   samples $\{\mathbf{z}^{(g)} = \big((t_i^{E(g)}|\mathbf{y},\hat{\theta}), i = 1, \ldots, N\big), g = 1, \ldots, M'\}$.
Calculute $D(\hat{\theta}) = -2E_{\mathbf{z}|\mathbf{y},\hat{\theta}}[\log(P(\mathbf{y},\mathbf{z}|\hat{\theta}))]$ and $\mathrm{DIC}_6 = 2\overline{D(\theta)} - D(\hat{\theta})$.
```

4.4.2 DIC$_6$ for models corresponding to the hypotheses

To compute DIC$_6$ for the null and full models (with corresponding hypotheses $H_0 : q = 1$ vs. $H_1 : q < 1$) their respective likelihoods are used. The likelihood for the null model is speficied in Section 4.3.1. The likelihood for the full model in discrete time is specified in equation (A.8) of appendix A. In the model comparison setting, a model with lower DIC is preferred (Knock & O'Neill, 2014, Alharthi et al., 2018).

Figure 4.1: Schematic representation of the MCMC algorithms to compute DIC$_6$. The figure overall presents the flow of information for computing DIC$_6$ in the present setting. The dashed box represents the scheme of data-augmented MCMC sampling.


Data augmented-MCMC
Gibbs Sampler: (Z|θ, y)
Metropolis-Hastings Sampler: (θ |z, y)
Estimate: $\hat{\theta}, h(\hat{\theta})$
Compute: $\overline{D(\theta)}$
Gibbs Sampler: (Z|$\hat{\theta}$, y)
Compute: DIC = $2\overline{D(\theta)} - D(\hat{\theta})$

Chapter 5

Simulations and Results

"It is a capital mistake to theorise before you have all the evidence. It biases the judgement."
- Sir Arthur Conan Doyle (Sherlock Holmes in A Study in Scarlet)

Simulations offer useful insights into outbreak characteristics and model properties. They are also a useful tool for evaluating inference procedures. In Section 5.1 simulated outbreak datasets are used to understand household final size distribution patterns. In Sections 5.2 simulations are used to generate data from a set of scenarios, primarily involving three settings of observed symptoms (based on the sensitivity and specificity) among others. The inference procedures presented in Chapters 3 and 4 are applied to these data in order to evaluate their performance. Sections 5.3 to 5.5 present some aspects of the model and results related to scaling, approximations and identifiability.

For the results presented using simulations in Sections 5.1 and 5.2, the outbreak datasets are simulated using a discrete-time version of the model of Section 2.5.1 (see Appendix A). In the discrete-time setting, the transmission parameters c and q are per day probabilities for escaping infection from the environment and an infectious household member, respectively.

5.1 Outbreak final size patterns in households using simulations

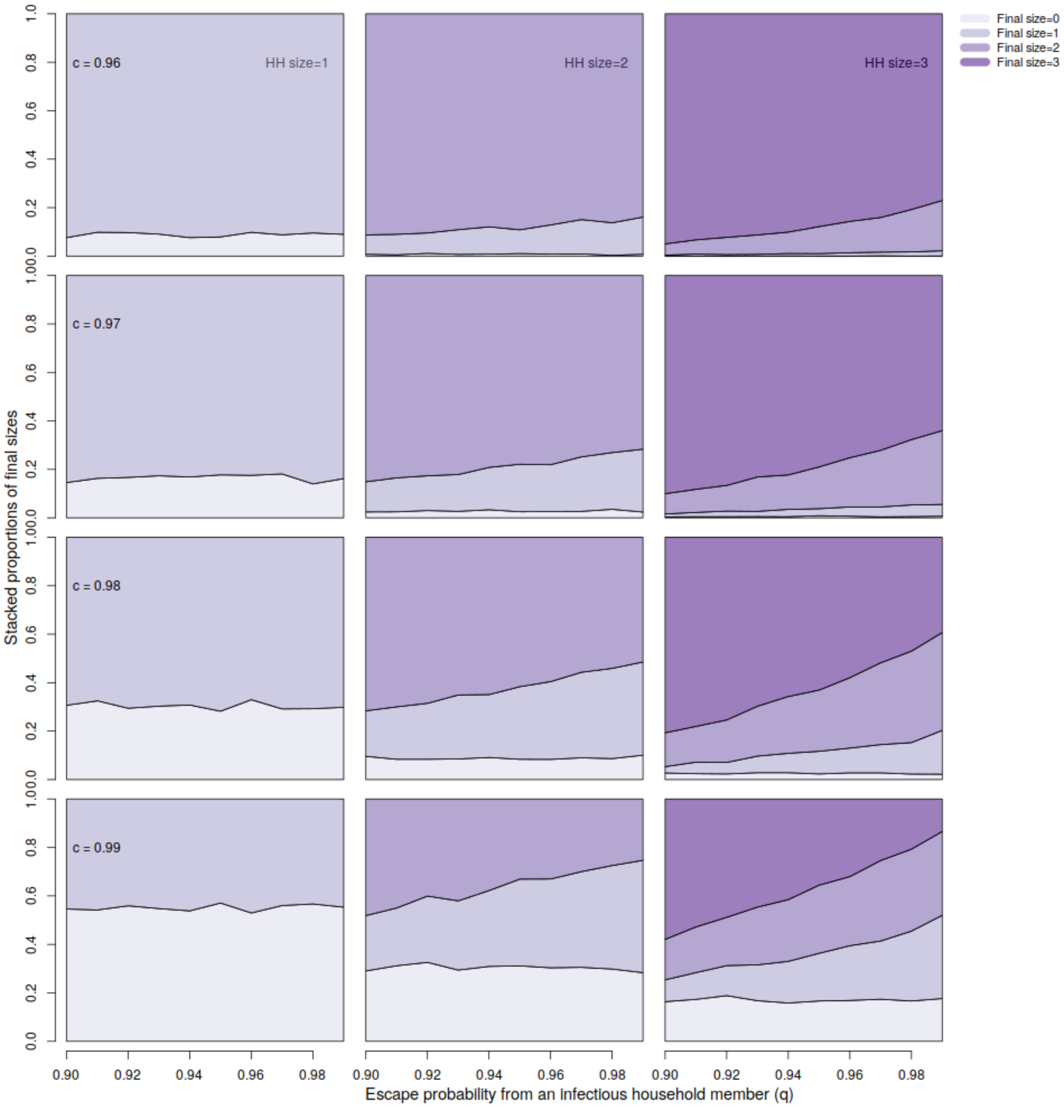
Outbreak datasets were simulated to study how the outbreak final size patterns in households are influenced by varying values of the two transmission parameters c and q (see Section 1.2.2 for a description of outbreak final size in households). A total of 500 households with sizes 1, 2 and 3 were used with proportions 0.2, 0.4 and 0.4, respectively. The environment was allowed to be infectious for 60 days ($\tau^1 = 60$). The sojourn times were bounded by $(l_{min}, l_{max}) = (4, 10)$, $(s_{min}, s_{max}) = (3, 6)$ and $(n_{min}, n_{max}) = (2, 4)$.

Infections were perfectly observed with fully sensitive and fully specific symptoms. Figure 5.1 presents the final size distributions by household size for a combination of values for c and q, with c ranging between 0.96 and 0.99 and q ranging between 0.90 and 0.99. The proportions presented are averaged from 10 simulations for each combination of values for c and q.

Interpretation of Figure 5.1
The figure presents final size distributions as twelve blocks (four rows and three columns): each row has a fixed value of c and each column has a fixed household size. Each block displays the final size distribution as stacked proportions for a fixed household size and a fixed value of c. Within each block, the value of q varies with the x-axis.

The final size distribution for households of size k gives the probabilities that some $j = 0, 1, \ldots, k$ of k household members have been infected at the end of outbreak. A final size of 0 indicates that the household remained uninfected throughout the outbreak.

Figure 5.1: Final size distributions by household size for a combination of values for c and q. The figures display the distribution (stacked proportions) of final sizes for a fixed household size and a fixed value of c. The proportions vary over increasing values of q on the x-axis. The rows represent fixed values of c ranging between 0.96 and 0.99. The columnns represent fixed household sizes of 1, 2 and 3. The time τ^1 when the environment stops to be infectious is fixed to 60 days. The proportions presented are averaged from 10 simulations for each setting.



In all the blocks, the proportion of households with final sizes of 0 is constant for all values of q. This is consistent with our model formulation: an household gets infected only from the environment and its corresponding probability is governed by c (see Section 5.3 for further discussion).

Increasing the value of c (within each column) increases the proportion of households with final size 0. For values of c less than 0.96 a negligible proportion of non-infected households (households with final size 0) were obtained in the simulations.

The effect of q pertains to households of size > 1, as it allows infections to occur based on contacts between infectious and susceptible individuals within the household. For a fixed value of c, increasing value of q affects the distribution of final sizes ≥ 1, in particular, decreasing the proportion of larger final sizes. However, the proportion of households with final size 0 (non-infected households) remains the same.

Finally, the effect of q can be seen with reference to c. For lower values of c, the probability that household members are infected from the environment is higher and thus leads to higher probabilities of larger final sizes (as can be seen from first two rows). However, for higher values of c, the probability that household members are infected from the environment is lower, allowing the effect of q to be seen more distinctly. Here, the proportion of larger final sizes decrease with increasing value of q (as can be seen from last two rows).

These combinations of values also provide insights about which values of c and q would allow the person-to-person transmission to be clearly distinguished and therefore inferred. This is also reflected in the choice of c and q used in simulation towards evaluating the inference procedures in Section 5.2.

5.2 Results: Inference and model comparison using simulated data

This section presents the results for the evaluation of performance of inference procedures developed for this application, in particular, the DA-MCMC method for estimation of the transmission parameters and the model comparison procedure using DIC$_6$ presented in Sections 3.4 and 4.3, respectively. This performance evaluation is done using simulated outbreak datasets from a set of simulation scenarios and are described as follows.

Outbreak time points and household size distributions in the population
The environment was allowed to be infectious for 60 days ($\tau^1 = 60$). T_c is the last observed symptom time after which there are no new symptomatic infections (as described in Section 2.5.1). The day of collecting serological samples (T_s) was set to a week after the outbreak was complete ($T_c + 7$). A total of 400 households containing only households of size 2 were used. Alternatively, a total of 500 households with sizes 1, 2 and 3 were drawn with proportions 0.2, 0.4 and 0.4, respectively.

Natural history of infection and observation model
The sojourn times were bounded by $(l_{min}, l_{max}) = (4, 10)$, $(s_{min}, s_{max}) = (3, 6)$ and $(n_{min}, n_{max}) = (2, 4)$. The threshold for start of antibody waning (t_w) was set larger than τ^1 and the rate of decrease per day in the detectability of antibodies (α) was set to 0.15.

Three settings were defined based on the sensitivity and specificity of observed symptoms: 1. perfectly observed symptoms ($\eta = 1, \psi = 1$), 2. non-sensitive but fully specific symptoms ($\eta < 1, \psi = 1$) and 3. non-sensitive and non-specific symptoms ($\eta < 1, \psi < 1$).

Transmission parameters for inference procedures
Two parameter sets were selected for the transmission parameter (c): 0.97 and 0.99. These values also display distinct final size patterns as can be seen in Figure 5.1. Values of $c < 0.95$ were not used as an unrealistically negligible proportion of households would remain uninfected. For the full model, which refers to presence of person-to-person transmission, the transmission parameter (q) is set to 0.95. As the null model refers to the absence of person-to-person transmission, it does not contain q and thus q was set to 1. Uniform priors were used for the transmission parameters c and q.

Computations for inference procedures
One dataset was simulated for each of the 24 combinations defined by the following: 3 settings based on sensitivity and specificity of observed symptoms, 2 models defined by the hypothesis of person-to-person transmission, 2 values for c and 2 distributions of household sizes (see Table 5.1). For each dataset the inference was performed under both the full and the null models.

The inference procedure of Section 3.4 and the model comparison procedure of section 4.3 were applied on the simulated datasets. In the analysis, the time when the environment becomes infectious (T^0) was set to the one used in simulation (τ^0). The MCMC inference and DIC$_6$ computations were performed using the following set-up: the number of MCMC samples were fixed to 10,000 for all settings with a burn-in of 1,000 samples. The number of samples for computing the deviance at posterior mean was also fixed to 10,000. The 95% credible intervals, in particular, the highest posterior density (HPD) intervals are also reported from the posterior samples of the transmission parameters.

Table 5.1: Table of variables that define twenty four combinations of simulation scenarios used towards evaluating the performance of parameter estimation and model comparison procedures

Variables defining simulation scenarios	**Notation**	**Values**
Models and their corresponding hypotheses	full model (H_1)	$q < 1$
	null model (H_0)	$q = 1$
Household size distribution	(2)	400 households of size 2 each
	(1,2,3)	500 households of sizes 1, 2 and 3 with proportions 0.2, 0.4 and 0.4, respectively
Transmission parameters	c	0.97
	c	0.99
	q	0.95
Sensitivity and specificity of symptoms	setting 1	$(\eta = 1, \psi = 1)$
	setting 2	$(\eta < 1, \psi = 1)$
	setting 3	$(\eta < 1, \psi < 1)$

All computations were performed within the `R 3.6.0` environment. The computationally intensive parts of the code such as likelihood computation and Gibbs sampler routines were written in `C++` and were called from within `R` using the `Rcpp` library routines. HPD intervals were obtained using the `coda` library routines. All computations were run on a personal computer. The running times for the three setting were approximately: 20 hours (setting 1 - table 5.2), 65 hours (setting 2 - table 5.3) and 75 hours (setting 3 - table 5.4).

5.2.1 Inference and model comparison using simulated data from setting 1: Perfectly observed symptoms

This section presents the results for the setting with perfectly observed symptoms ($\eta = 1$ and $\psi = 1$). Table 5.2 presents the marginal posterior estimates of the transmission parameters in terms of means and 95% highest posterior density (HPD) intervals. The first four datasets (1-4) were simulated under the full model ($H_1 : q < 1$) and the next four datasets (5-8) were simulated under the null model ($H_0 : q = 1$). The four datasets differ in household size distribution and transmission parameter values (q and c). Each dataset was fitted with both the full and the null models (two rows each); the posterior estimates of the model parameters, the deviance information criteria (DIC_6) and the effective number of parameters (p_{D6}) are presented.

The posterior estimates of the transmission parameters (q and c) correspond to the respective values used in simulation when inferences are made under the correct model except in dataset 2. In datasets 1-4, the value of q is consistently estimated well below 1 and the 95% HPD interval contains the value of q used in simulation (except in dataset 2). In datasets 5-8 the value of q is consistently estimated close to 1 and the upper limit of the 95% HPD interval is 1 (except in dataset 5). The values of c are estimated consistently about the respective values used in simulation under both the full model (datasets 1-4) and the null model (datasets 5-8).

For datasets simulated under the full model, the DIC_6 values for the full model are smaller than the DIC_6 values for the null model, except for dataset 2. For datasets simulated under the null model, the DIC_6 values are similar for both the full and the null models. The effective numbers of parameters (p_{D6}) are all positive and consistently higher for the full model than than the null model.

Table 5.2: Parameter estimates and DIC$_6$ in the setting 1 with perfect observation of symptoms ($\eta = 1$, $\psi = 1$). Datasets were simulated under the full and the null models (denoted by H_1 and H_0) with two household size distributions and transmission parameter settings. The posterior means and 95% highest posterior density (HPD) intervals are presented for the transmission parameters along with the deviance information criteria (DIC$_6$) and effective number of parameters (p$_{D6}$) as part of inference under each hypothesis.

Datasets	**Simulation settings**		**Inference under**	**Parameter estimate (95% HPD Interval)**		**DIC**	
	HH size	(q, c)	Hypothesis	q	c	DIC$_6$	p$_{D6}$
Simulated under $H_1 : q < 1$							
1	(2)	(0.95, 0.97)	H_1	0.949 (0.933, 0.964)	0.974 (0.971, 0.976)	9645.86	2.32
1	(2)	(0.95, 0.97)	H_0	- (-, -)	0.970 (0.968, 0.972)	9728.46	1.81
2	(1,2,3)	(0.95, 0.97)	H_1	0.966 (0.957, 0.974)	0.972 (0.970, 0.974)	13678.01	2.32
2	(1,2,3)	(0.95, 0.97)	H_0	- (-, -)	0.968 (0.966, 0.970)	13664.62	0.68
3	(2)	(0.95, 0.99)	H_1	0.939 (0.925, 0.953)	0.993 (0.992, 0.994)	6257.13	1.93
3	(2)	(0.95, 0.99)	H_0	- (-, -)	0.991 (0.990, 0.992)	6479.11	1.08
4	(1,2,3)	(0.95, 0.99)	H_1	0.953 (0.945, 0.961)	0.992 (0.991, 0.993)	9955.72	2.89
4	(1,2,3)	(0.95, 0.99)	H_0	- (-, -)	0.989 (0.988, 0.990)	10102.69	1.27
Simulated under $H_0 : q = 1$							
5	(2)	(1, 0.97)	H_1	0.992 (0.982, 0.999)	0.975 (0.973, 0.977)	9283.81	1.52
5	(2)	(1, 0.97)	H_0	- (-, -)	0.974 (0.972, 0.976)	9285.33	1.12
6	(1,2,3)	(1, 0.97)	H_1	0.996 (0.990, 1.000)	0.974 (0.972, 0.976)	12831.62	2.33
6	(1,2,3)	(1, 0.97)	H_0	- (-, -)	0.973 (0.972, 0.975)	12827.70	1.05
7	(2)	(1, 0.99)	H_1	0.998 (0.993, 1.000)	0.993 (0.992, 0.993)	5487.19	1.19
7	(2)	(1, 0.99)	H_0	- (-, -)	0.992 (0.992, 0.993)	5485.42	0.99
8	(1,2,3)	(1, 0.99)	H_1	0.998 (0.994, 1.000)	0.992 (0.991, 0.993)	7840.18	1.97
8	(1,2,3)	(1, 0.99)	H_0	- (-, -)	0.992 (0.991, 0.993)	7835.84	0.26

Figure 5.2 displays the posterior samples of transmission parameters for data generated from the full model ($H_1 : q < 1$). The density of q under the full model are well below the value 1 and are symmetric for all four datasets. The density of c under the full model have their ranges and peaks greater than their corresponding ones under the null model.

Figure 5.2: Posterior plots of the transmission parameters for datasets generated from the full model ($H_1 : q < 1$) in setting 1. The posterior samples are displayed using trace and density plots. The plots correspond to datasets 1 to 4 of Table 5.2. Each row represents a dataset: the first 4 plots in each row are posterior summaries from the full model ($H_1 : q < 1$) and the last 2 plots in each row are posterior summaries from the null model ($H_0 : q = 1$).

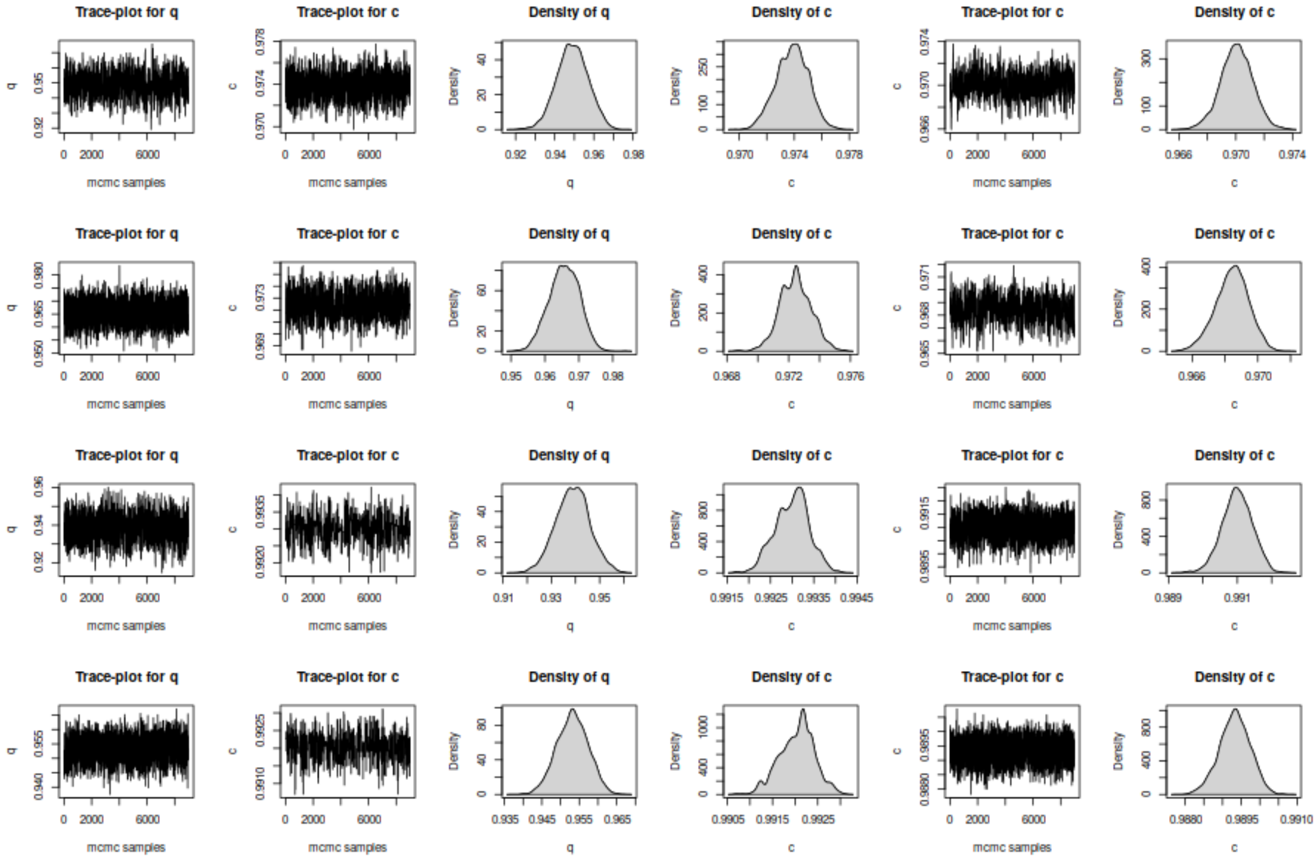
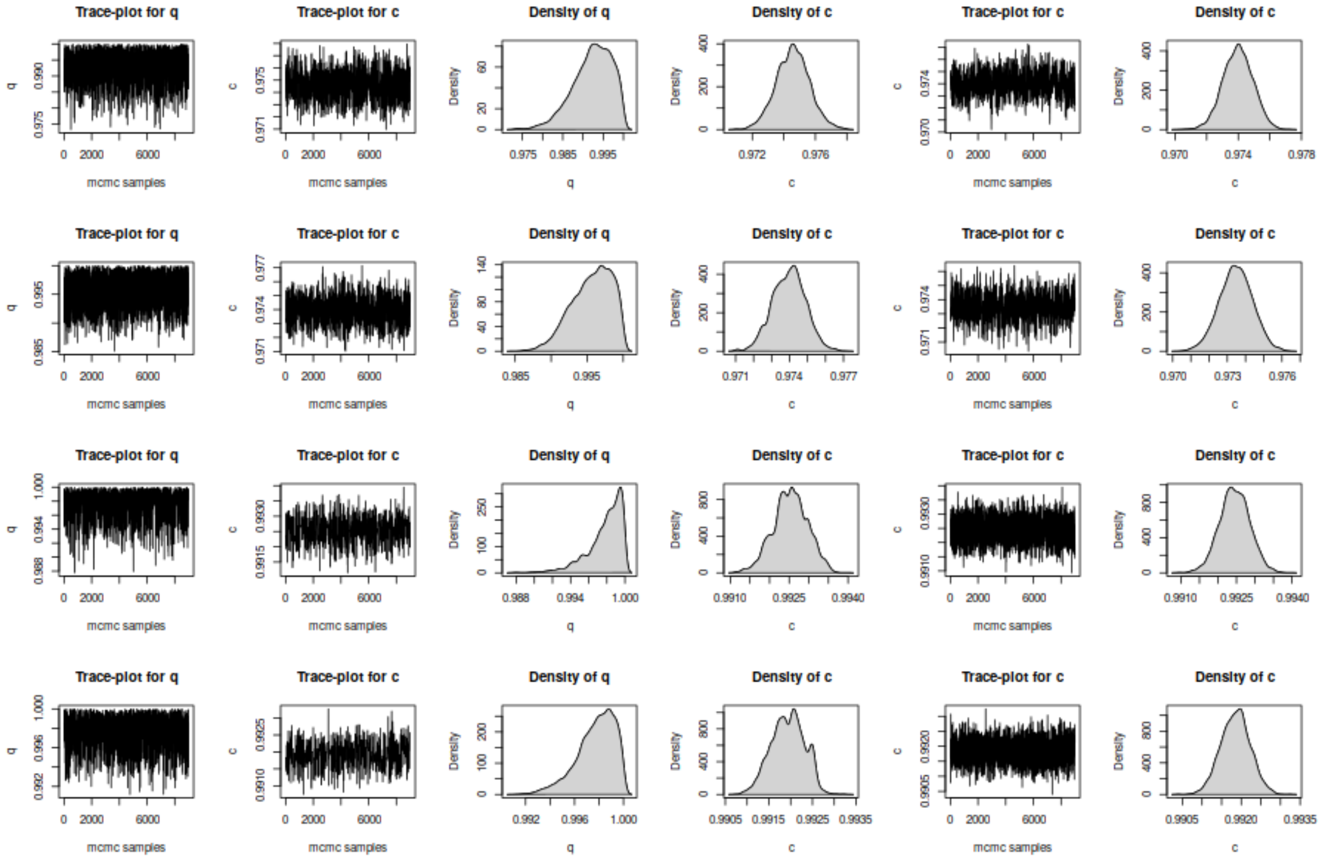


Figure 5.3 displays the posterior samples of transmission parameters for data generated from the null model ($H_0 : q = 1$). The density of q under the full model are close to the value 1 with their peaks occuring above 0.99 and are skewed for all four datasets. The density of c, on the other hand, are peaked at similar values for posterior samples under both the full and the null models.

Figure 5.3: Posterior plots of the transmission parameters for datasets generated from the null model ($H_0 : q = 1$) in setting 1. The posterior samples are displayed using trace and density plots. The plots correspond to datasets 5 to 8 of Table 5.2. Each row represents a dataset: the first 4 plots in each row are posterior summaries from the full model ($H_1 : q < 1$) and the last 2 plots in each row are posterior summaries from the null model ($H_0 : q = 1$).



5.2.2 Inference and model comparison using simulated data from setting 2: Imperfectly observed symptoms (non-sensitive)

This section presents the results for the setting 2 i.e., under a situation in which the symptoms are not fully sensitive ($\eta < 1$) but are fully specific ($\psi = 1$). Table 5.3 presents the marginal posterior estimates of the transmission parameters in terms of means and 95% highest posterior density (HPD) intervals. The first four datasets (1-4) were simulated under the full model ($H_1 : q < 1$) and the next four datasets (5-8) were simulated under the null model ($H_0 : q = 1$). The four datasets differ in household size distribution and transmission parameter values (q and c). Each dataset was fitted with both the full and the null models (two rows each); the posterior estimates of the model parameters, the deviance information criteria (DIC$_6$) and the effective number of parameters (p$_{D6}$) are presented.

In datasets 1-4, the value of q is consistently estimated well below 1 although the 95% HPD interval contains the value of q used in simulation in only one of the four datasets. In datasets 5-8 the value of q is consistently estimated close to 1 and the upper limit of the 95% HPD intervals are 1 for all four datasets. The values of c are estimated consistently about the respective values used in simulation under both the full model (datasets 1-4) and the null model (datasets 5-8).

For datasets simulated under the full model, the DIC$_6$ values for the full model are smaller than the DIC$_6$ values for the null model, except for dataset 2. For datasets simulated under the null model, the DIC$_6$ values are similar for both the full and the null models. Albeit, the effective numbers of parameters (p$_{D6}$) are negative for datasets 1 and 2, and in general, the (p$_{D6}$) values are not higher for the full models than their corresponding null models. This makes the interpetation of DIC difficult. This problem is documented for DIC$_6$ (Celeux et al., 2006).

Table 5.3: Parameter estimates and DIC$_6$ in the setting 2 with imperfect observation of symptoms (η<1, $\psi = 1$). Datasets were simulated under the full and the null models (denoted by H_1 and H_0) with two household size distributions and transmission parameter settings. The posterior means and 95% highest posterior density (HPD) intervals are presented for the transmission parameters along with the deviance information criteria (DIC$_6$) and effective number of parameters (p$_{D6}$) as part of inference under each hypothesis.

Datasets	**Simulation settings**		**Inference under**	**Parameter estimate (95% HPD Interval)**		**DIC**	
	HH size	(q, c)	Hypothesis	q	c	DIC$_6$	p$_{D6}$
Simulated under $H_1 : q < 1$							
1	(2)	(0.95, 0.97)	H_1	0.911 (0.883, 0.936)	0.967 (0.963, 0.971)	10159.05	-2.39
1	(2)	(0.95, 0.97)	H_0	- (-, -)	0.959 (0.954, 0.964)	10298.65	-2.66
2	(1,2,3)	(0.95, 0.97)	H_1	0.974 (0.964, 0.985)	0.961 (0.957, 0.964)	14210.21	-0.47
2	(1,2,3)	(0.95, 0.97)	H_0	- (-, -)	0.957 (0.953, 0.961)	14188.58	-1.44
3	(2)	(0.95, 0.99)	H_1	0.944 (0.929, 0.958)	0.993 (0.992, 0.994)	6415.23	2.59
3	(2)	(0.95, 0.99)	H_0	- (-, -)	0.991 (0.990, 0.992)	6628.15	2.15
4	(1,2,3)	(0.95, 0.99)	H_1	0.979 (0.972, 0.985)	0.991 (0.990, 0.992)	9681.51	0.64
4	(1,2,3)	(0.95, 0.99)	H_0	- (-, -)	0.990 (0.989, 0.991)	9734.09	1.31
Simulated under $H_0 : q = 1$							
5	(2)	(1, 0.97)	H_1	0.993 (0.982, 1.000)	0.973 (0.971, 0.975)	9354.52	1.02
5	(2)	(1, 0.97)	H_0	- (-, -)	0.973 (0.970, 0.975)	9360.57	2.61
6	(1,2,3)	(1, 0.97)	H_1	0.997 (0.991, 1.000)	0.972 (0.970, 0.974)	12934.95	1.19
6	(1,2,3)	(1, 0.97)	H_0	- (-, -)	0.972 (0.969, 0.974)	12934.62	0.76
7	(2)	(1, 0.99)	H_1	0.997 (0.992, 1.000)	0.993 (0.992, 0.993)	5467.93	0.95
7	(2)	(1, 0.99)	H_0	- (-, -)	0.992 (0.992, 0.993)	5469.44	1.37
8	(1,2,3)	(1, 0.99)	H_1	0.998 (0.995, 1.000)	0.992 (0.991, 0.993)	7887.62	1.55
8	(1,2,3)	(1, 0.99)	H_0	- (-, -)	0.992 (0.991, 0.993)	7885.50	0.24

Figure 5.4 displays the posterior samples of transmission parameters for data generated from the full model ($H_1 : q < 1$). The density of q under the full model are well below the value 1 and are symmetric for all four datasets. The density of c under the full model have their ranges and peaks greater than their corresponding ones under the null model.

Figure 5.4: Posterior plots of transmission parameters for datasets generated from the full model ($H_1 : q < 1$) in setting 2. The posterior samples are displayed using trace and density plots. The plots correspond to datasets 1 to 4 of Table 5.3. Each row represents a dataset: the first 4 plots in each row are posterior summaries from the full model ($H_1 : q < 1$) and the last 2 plots in each row are posterior summaries from the null model ($H_0 : q = 1$).

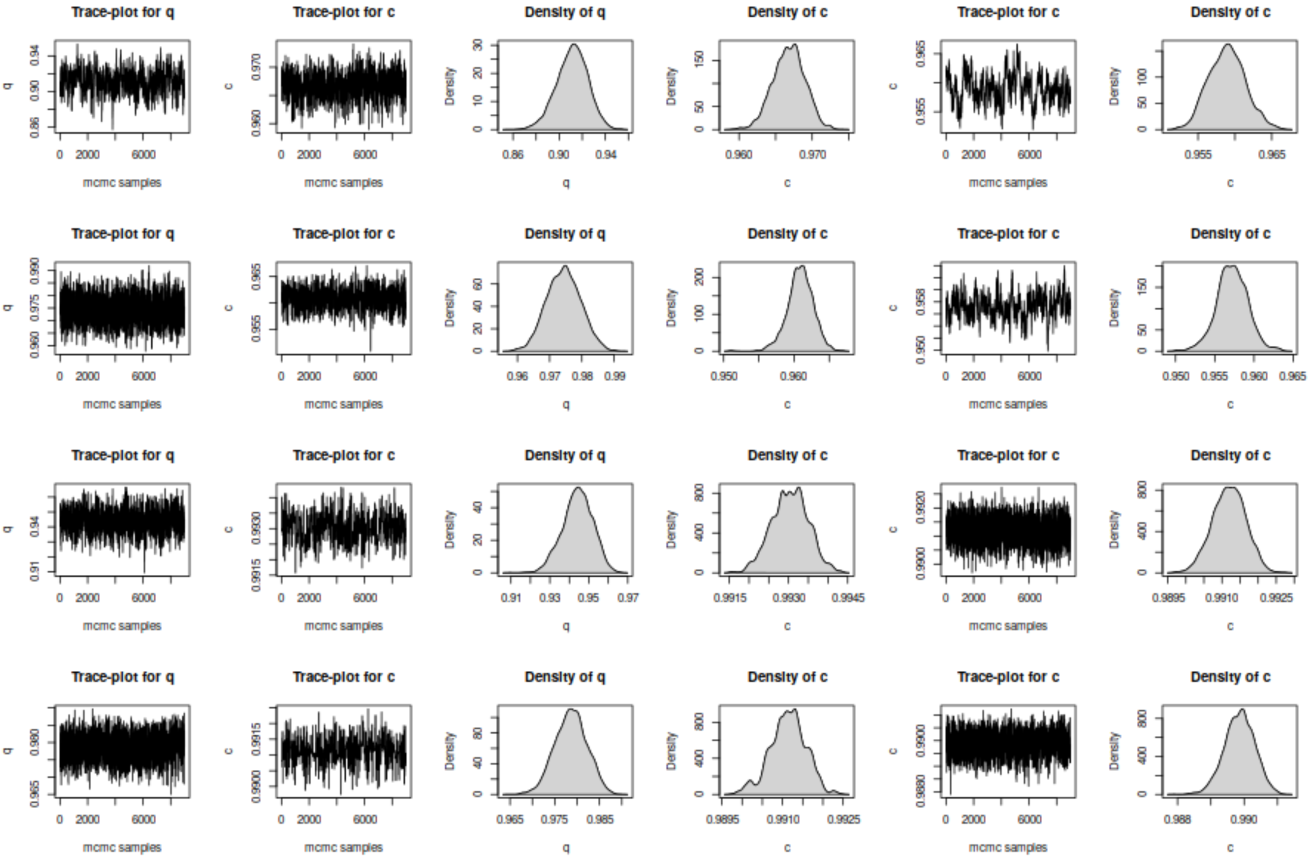
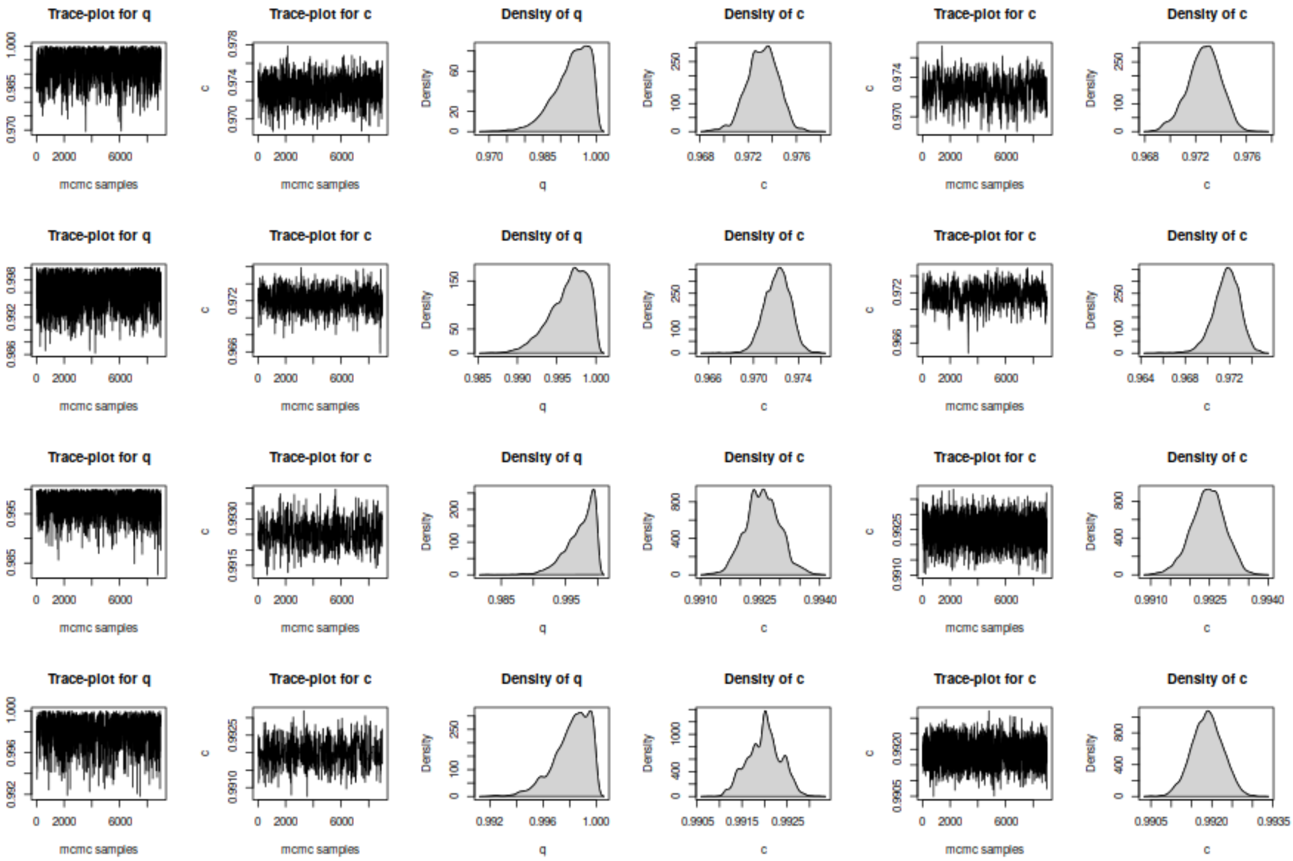


Figure 5.5 displays the posterior samples of transmission parameters for data generated from the null model ($H_0 : q = 1$). The density of q under the full model are close to the value 1 with their peaks occuring above 0.995 and are skewed for all four datasets. The density of c, on the other hand, are peaked at similar values for posterior samples under both the full and the null models.

Figure 5.5: Posterior plots of the transmission parameters for datasets generated from the null model ($H_0 : q = 1$) in setting 2. The posterior samples are displayed using trace and density plots. The plots correspond to datasets 5 to 8 of Table 5.3. Each row represents a dataset: the first 4 plots in each row are posterior summaries from the full model ($H_1 : q < 1$) and the last 2 plots in each row are posterior summaries from the null model ($H_0 : q = 1$).



5.2.3 Inference and model comparison using simulated data from setting 3: Imperfectly observed symptoms (neither fully sensitive nor specific)

This section presents the results for the setting 3 i.e., under a situation in which the symptoms are neither fully sensitive nor fully specific ($\eta < 1$ and $\psi < 1$). Table 5.4 presents the marginal posterior estimates of the transmission parameters in terms of means and 95% highest posterior density (HPD) intervals. The first four datasets (1-4) were simulated under the full model ($H_1 : q < 1$) and the next four datasets (5-8) were simulated under the null model ($H_0 : q = 1$). The four datasets differ in household size distribution and transmission parameter values (q and c). Each dataset was fitted with both the full and the null models (two rows each); the posterior estimates of the model parameters, the deviance information criteria (DIC$_6$) and the effective number of parameters (p$_{D6}$) are presented.

In datasets 1-4, the value of q is consistently estimated well below 1 although the 95% HPD interval contains the value of q used in simulation in only two of the four datasets. In datasets 5-8 the value of q is consistently estimated close to 1 and the upper limit of the 95% HPD intervals are 1 for all four datasets. The values of c are estimated consistently about the respective values used in simulation under the null model (datasets 5-8). Under the full model (datasets 1-4), the posterior estimates of c are consistent with values used in simulation when data are simulated with $c = 0.99$ but not so when simulated with $c = 0.97$.

For datasets simulated under the full model, the DIC$_6$ values for full model are less than the DIC$_6$ values for null model in all four datasets. For datasets simulated under the null model, the DIC$_6$ values are similar for both the full and the null models. The effective number of parameters (p$_{D6}$) are positive except for one null model computation in a dataset simulated under the full model. The (p$_{D6}$) values are also consistently higher for the full model than than the null model except for one dataset.

Table 5.4: Parameter estimates and DIC$_6$ in the setting 3 with imperfect observation of symptoms (η<1, ψ<1). Datasets were simulated under the full and the null models (denoted by H_1 and H_0) with two household size distributions and transmission parameter settings. The posterior means and 95% highest posterior density (HPD) intervals are presented for the transmission parameters along with the deviance information criteria (DIC$_6$) and effective number of parameters (p$_{D6}$) as part of inference under each hypothesis.

Datasets	**Simulation settings**		**Inference under**	**Parameter estimate (95% HPD Interval)**		**DIC**	
	HH size	(q, c)	Hypothesis	q	c	DIC$_6$	p$_{D6}$
Simulated under $H_1 : q < 1$							
1	(2)	(0.95, 0.97)	H_1	0.963 (0.950, 0.975)	0.994 (0.993, 0.994)	6617.35	2.80
1	(2)	(0.95, 0.97)	H_0	- (-, -)	0.992 (0.992, 0.993)	6736.63	1.34
2	(1,2,3)	(0.95, 0.97)	H_1	0.986 (0.979, 0.992)	0.996 (0.995, 0.996)	7477.07	2.00
2	(1,2,3)	(0.95, 0.97)	H_0	- (-, -)	0.995 (0.995, 0.996)	7510.00	0.84
3	(2)	(0.95, 0.99)	H_1	0.952 (0.938, 0.965)	0.995 (0.994, 0.996)	5538.45	2.40
3	(2)	(0.95, 0.99)	H_0	- (-, -)	0.994 (0.993, 0.994)	5722.25	1.25
4	(1,2,3)	(0.95, 0.99)	H_1	0.981 (0.974, 0.988)	0.993 (0.993, 0.994)	8543.95	1.62
4	(1,2,3)	(0.95, 0.99)	H_0	- (-, -)	0.992 (0.992, 0.993)	8599.01	-0.21
Simulated under $H_0 : q = 1$							
5	(2)	(1, 0.97)	H_1	0.992 (0.981, 0.999)	0.977 (0.974, 0.979)	9105.29	1.37
5	(2)	(1, 0.97)	H_0	- (-, -)	0.976 (0.974, 0.978)	9113.71	1.85
6	(1,2,3)	(1, 0.97)	H_1	0.996 (0.990, 1.000)	0.975 (0.973, 0.977)	12686.68	1.33
6	(1,2,3)	(1, 0.97)	H_0	- (-, -)	0.975 (0.973, 0.976)	12687.83	0.37
7	(2)	(1, 0.99)	H_1	0.998 (0.993, 1.000)	0.993 (0.992, 0.994)	5382.06	1.69
7	(2)	(1, 0.99)	H_0	- (-, -)	0.993 (0.992, 0.994)	5380.12	0.91
8	(1,2,3)	(1, 0.99)	H_1	0.998 (0.994, 1.000)	0.992 (0.992, 0.993)	7812.30	0.79
8	(1,2,3)	(1, 0.99)	H_0	- (-, -)	0.992 (0.992, 0.993)	7812.28	0.54

Figure 5.6 displays the posterior samples of transmission parameters for data generated from the full model ($H_1 : q < 1$). The density of q under the full model are well below the value 1 and are symmetric for all four datasets. The density of c under the full model have their ranges and peaks greater than their corresponding ones under the null model.

Figure 5.6: Posterior plots of transmission parameters for datasets generated from the full model ($H_1 : q < 1$) in setting 3. The posterior samples are displayed using trace and density plots. The plots correspond to datasets 1 to 4 of Table 5.4. Each row represents a dataset: the first 4 plots in each row are posterior summaries from the full model ($H_1 : q < 1$) and the last 2 plots in each row are posterior summaries from the null model ($H_0 : q = 1$).

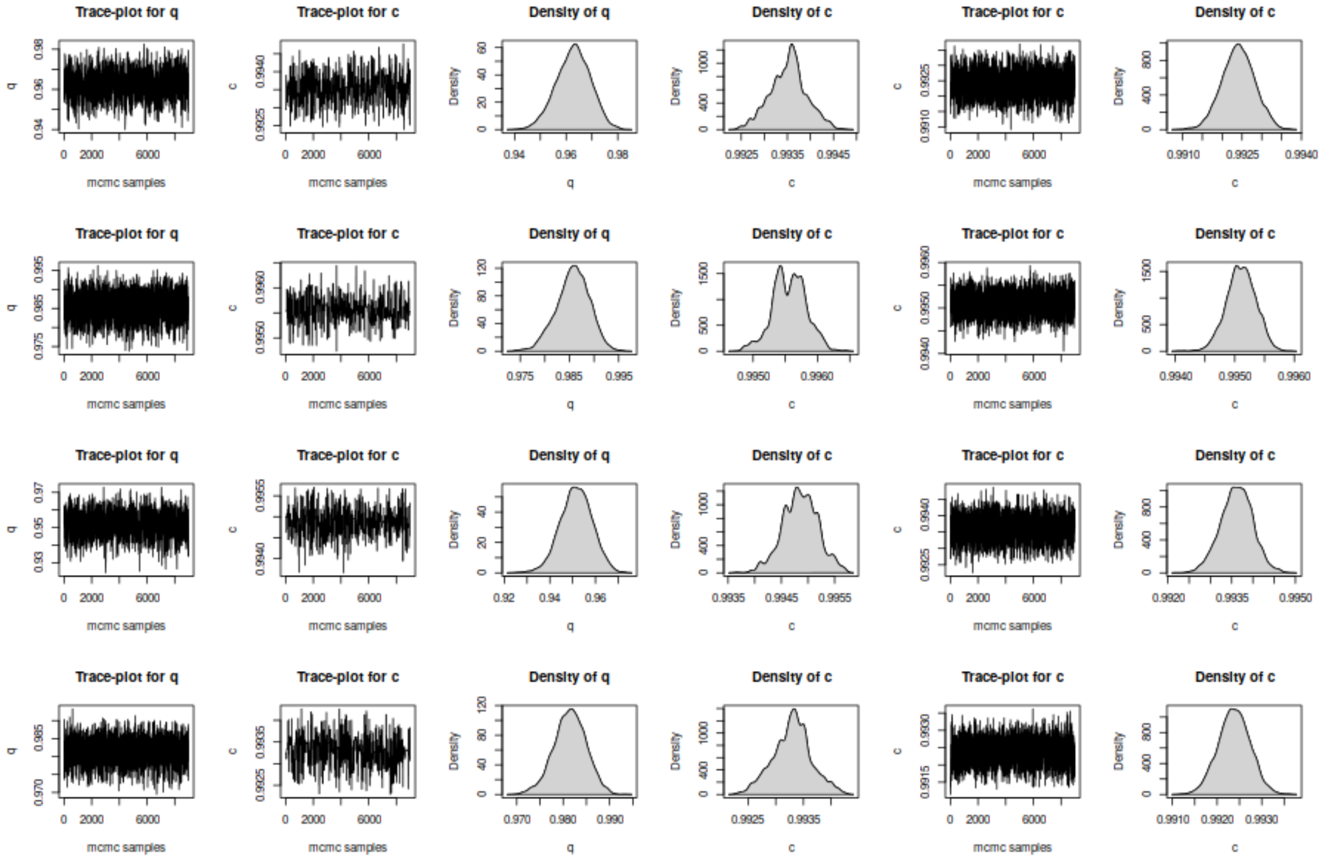
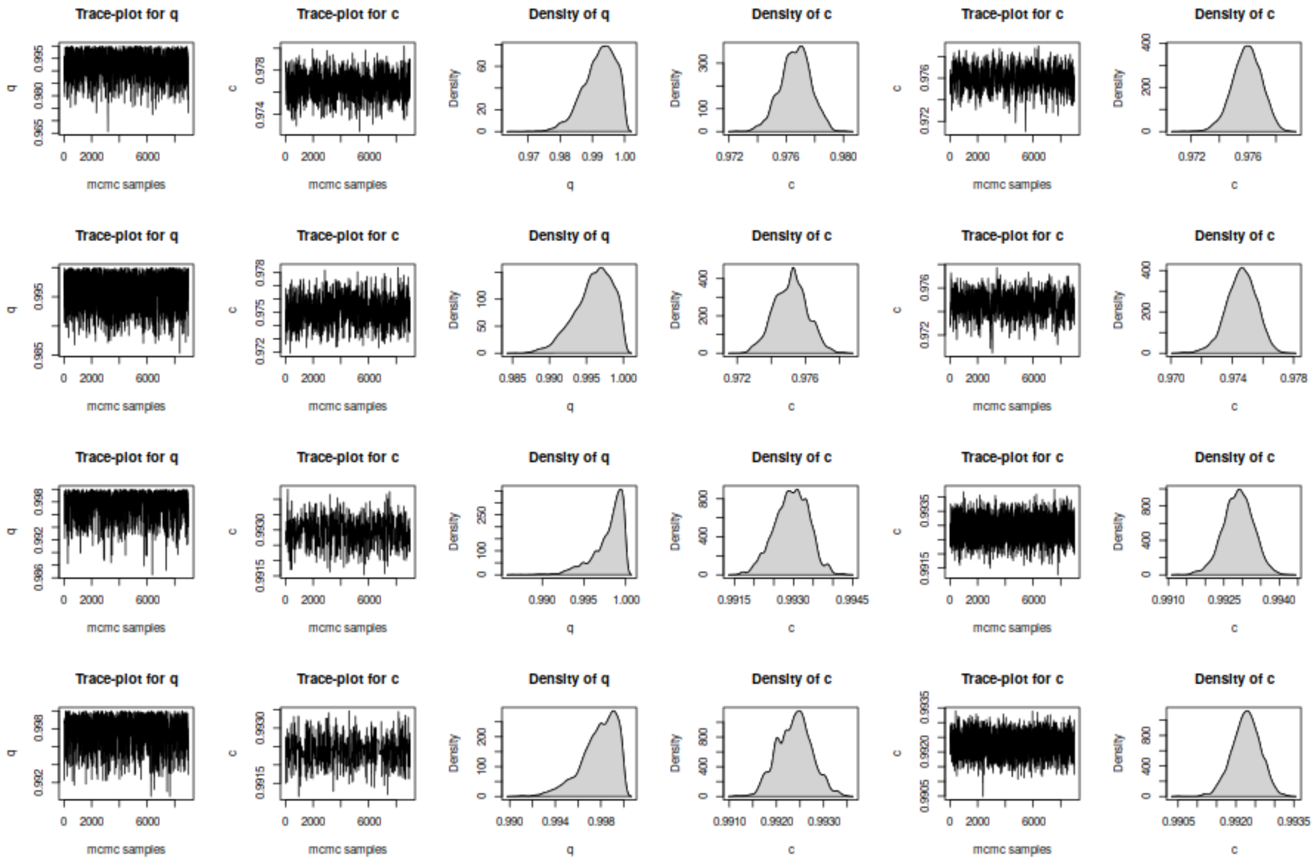


Figure 5.7 displays the posterior samples of transmission parameters for data generated from the null model ($H_0 : q = 1$). The density of q under the full model are close to the value 1 with their peaks occuring above 0.995 and are skewed for all four datasets. The density of c, on the other hand, are peaked at similar values for posterior samples under both the full and the null models.

Figure 5.7: Posterior plots of transmission parameters for datasets generated from the null model ($H_0 : q = 1$) in setting 3. The posterior samples are displayed using trace and density plots. The plots correspond to datasets 5 to 8 of Table 5.4. Each row represents a dataset: the first 4 plots in each row are posterior summaries from the full model ($H_1 : q < 1$) and the last 2 plots in each row are posterior summaries from the null model ($H_0 : q = 1$).



5.2.4 Summary of results on inference procedures using simulated data

Based on outbreak datasets simulated under a combination of scenarios (see Table 5.1), the results for the posterior estimates of the transmission parameters and DIC_6 are provided in Tables 5.2, 5.3 and 5.4. The performance of the inference procedures, namely, parameter estimation using data-augmented MCMC and the model comparison using DIC_6, are presented in this section with respect to the provided results.

Performance under different observation settings
Setting 1
In setting 1, where perfectly observed symptoms ($\eta = 1$ and $\psi = 1$) are assumed, the posterior estimates of the transmission parameters (q and c) correspond to the respective values used in simulation when inferences are made under the correct model.

For datasets simulated under the full model, the DIC_6 values for the full model are smaller than the DIC_6 values for the null model, except for one dataset. For datasets simulated under the null model, the DIC_6 values are similar for both the full and the null models.

Setting 2
In setting 2, where imperfectly observed symptoms ($\eta < 1$ and $\psi = 1$) are assumed, i.e. under the situation of non-sensitve but fully specific symptoms, the posterior estimates of the transmission parameter q lie well below the value 1, although not completely consistent with values used in the simulation for datasets simulated under the full model. However, for datasets simulated under the null model, they are completely consistent with the values used in the simulation and always include the value 1 in their 95% HPD intervals. The posterior estimates of c correspond to the respective values used in the simulation for datasets simulated under both the full and the null models.

For datasets simulated under the full model, the DIC_6 values for the full model are smaller than the DIC_6 values for the null model, except for one dataset. For datasets simulated under the null model, the DIC_6 values are similar for both the full and the null models. For datasets simulated under the full model, when the value of $c = 0.97$ in the simulation, the effective number of parameters associated with DIC_6 values are negative. Negative values for the effective number of parameters for DIC_6 has been documented in Celeux et al. (2006).

Setting 3
In setting 3, where imperfectly observed symptoms ($\eta < 1$ and $\psi = 1$) are assumed, i.e. under the situation of non-sensitve and non-specific symptoms, for datasets simulated under the full model, the posterior estimates of the transmission parameter q lie well below the value 1, although not completely consistent with the values used in the simulation. The corresponding estimates for c are consistent with values used in simulation when data are simulated with $c = 0.99$ but not so when simulated with $c = 0.97$. For datasets simulated under the null model, the posterior estimates of the transmission parameters (q and c) correspond to the respective values used in the simulation.

For datasets simulated under the full model, the DIC_6 values for the full model are smaller than the DIC_6 values for the null model, except for one dataset. For datasets simulated under the null model, the DIC_6 values are similar for both the full and the null models.

Performance under different values of c for a fixed q
For simulation under the full model, in the setting of perfectly observed symptoms ($\eta = 1$, $\psi = 1$), the posterior estimates of the transmission parameters were consistently estimated in the neighbourhood of values used in the simulation in both simulation scenarios $(q, c) = (0.95, 0.97)$ and $(q, c) = (0.95, 0.99)$. However, in the settings of imperfectly observed symptoms ($\eta < 1$, $\psi = 1$ and $\eta < 1$, $\psi < 1$), the posterior estimates of the transmission parameters were closer to the values used in the simulation for the simulation scenario of $(q, c) = (0.95, 0.99)$ than for the simulation scenario of $(q, c) = (0.95, 0.97)$.

For simulation under the null model ($q = 1$), the posterior estimates of the transmission parameters are consistently lie in the neighbourhood of values used in the simulation in both scenarios $c = 0.97$ and $c = 0.99$.

For simulations under both the full and the null models and in both scenarios $c = 0.97$ and $c = 0.99$, the DIC$_6$ values consistently indicated towards the correct model, i.e. the DIC$_6$ values are consistently smaller for the full model than the null model when the datasets are simulated under the full model in all settings and the DIC$_6$ values are consistently similar under both the full and the null models when the datasets are simulated under the null model in all the three settings.

Performance under the two models that generate the datasets
For datasets simulated under the full model, the posterior estimates of the transmission parameter q lie well below the value 1 and consistent to some extent with values used in simulation in all the three settings. The posterior estimates of c correspond to the respective values used in simulation in all the three settings. The DIC$_6$ values for the full model are almost consistently smaller than the DIC$_6$ values for the null model in all the three settings.

For datasets simulated under the null model, the parameter estimates of q is consistently estimated in the neighbourhood of 1, with the 95% HPD intervals containing 1 and the posterior estimates of c consistently lie in the neighbourhood of the value used in simulation in all the three settings. The DIC$_6$ values are consistently similar under both the full and the null models in all the three settings.

Performance under different household size distribution
The performance of parameter estimation and DIC$_6$ are not different between the two chosen household size distributions, i.e. households of size of 2 and households of sizes 1, 2 and 3.

Overall performance of estimation of the transmission parameters
Overall, for datasets simulated under the null model (datasets 5-8), the posterior estimates of the transmission parameters (q, c) are consistently estimated in the neighbourhood of the values used in simulation in all the three settings. For datasets simulated under the full model (datasets 1-4), the posterior estimates of the transmission parameters (q, c) are consistently estimated in the neighbourhood of the values used in simulation in setting 1 (perfectly observed symptoms) and consistent to some extent with the values used in simulation in settings 2 and 3 (imperfectly observed symptoms).

Overall performance of DIC$_6$
Overall, the DIC$_6$ values are consistently smaller for the full model than the null model when the datasets are simulated under the full model (datasets 1-4) under all three observation settings. The DIC$_6$ values are consistently similar for both the full and the null models when the datasets are simulated under the null model (datasets 5-8) under all three observation settings. The DIC$_6$ values consistently indicated towards the correct model in almost all scenarios and is deemed robust with respect to the simulation scenarios used.

5.3 Analysis of final size of outbreak and its implications for presenting results based on simulated data

This section presents some approximations of the model of Section 2.5.1 for the following reasons: (i) to aid understanding of the epidemic behaviour in relation to certain model parameters and (ii) to identify any parameter redundance in the presentation of results in order to make them generalisable without being tied to any particular set of input values for the parameters. Note that the model here correponds to continuous time and the following parameterisation only holds as an approximation to illustrate some scaling issues and is not referred elsewhere in this thesis.

As described in Section 2.3, μ and λ are the hazards of a still susceptible individual being infected from the environment and an infectious household member, respectively. Let $\pi_c = e^{-\mu\tau^1}$ be the probability that a still susceptible individual escapes infection from the environment during the outbreak when none of the household members are infectious. Let $\pi_h = e^{-\lambda T_I}$ be the probability that a still susceptible individual escapes infection from an infected household member whose (fixed) infectious duration is T_I.

Let $k = 1, 2, \ldots, K$, be the possible household sizes. Let $m_k, k = 1, 2, \ldots, K$, be the number of households of size k in the population. The household-size distribution in the population is then given as

$$p_k = \frac{m_k}{\sum_{k=1}^{K} m_k}, k = 1, 2, \ldots, K. \tag{5.1}$$

(i) Proportion of households infected by household size

Let a be the number of household members infected from the environment in a household of size k. In fact, in our model, the only way an household becomes infected is from the environment. Then $a|k, \pi_c \sim Bin(k, 1-\pi_c)$ and the probability that an household of size k is ever infected during the outbreak, (i.e., $a \geq 1$) is exactly given by

$$P_k(a \geq 1) = 1 - P_k(a = 0) = 1 - \pi_c^k. \tag{5.2}$$

This is the probability that a household of size k is infected during an oubreak. Figure 5.1 displays the proportion $P_k(a = 0) = \pi_c^k$ being almost constant for a fixed value of c irrespective of the value of q (although the graph is generated from a discrete-time model wherein this probability would be $P_k(a = 0) = c^{k\tau^1}$).

(ii) Final size of outbreak by household size

Let j be the final size in a household with k susceptible members at the start of the outbreak (i.e., j of k members are infected over the outbreak). When the household gets infected during the outbreak, both the environment and the infectious household members contribute to the final size within the household. Based on Longini & Koopman (1982) and O'Neill et al. (2000) the probability that a household of size k has final size j is given by

$$P_k(j|\pi_c, \pi_h) = \binom{k}{j} P_j(j)(\pi_c \pi_h^j)^{k-j}. \tag{5.3}$$

Here $(\pi_c \pi_h^j)^{k-j}$ refers to the probability that some $k - j$ uninfected individuals escape infection from the environment and each of the j infected household members. The term $P_j(j)$ refers to the probability that some j individuals are infected, implicitly denoting all pathways of acquiring the infection. Although the expression is provided for SIR models, it also holds for SEIR models.

Computing $P_j(j)$ is a non-trivial task. For infections that only spread by direct contact, all infections from the community can be considered as initial infections in the household (i.e., they occur only to introduce the infection to the household). In such cases $P_j(j)$ can be obtained in more than one ways. Longini & Koopman (1982) provide a recursive formula $P_j(j) = 1 - \sum_{i=0}^{j-1} P_j(i)$ in such setting. O'Neill et al. (2000) provide closed form approximations using Gontcharoff polynomials under different assumptions for heterogeneity in infectivity.

Ball & Lyne (2002, 2006) and Shaw (2016) also provide a more general recursive formulation for $P_{k,a}(j|\lambda^{(k)})$, where a is the number of initial infections from the community and $\lambda^{(k)}$ is a household-size-dependent within-household transmission rate.

The above formulations for $P_j(j)$, however, consider a single epidemic chain occurring within the household and account for all infections from the environment at the beginning of the epidemic chain. These formulations do not account for multiple epidemic chains within the household resulting from re-initiation of infections from the community (which is the case of the application in this thesis). A more tedious method is to explicitly model $P_j(j)$ using all possible epidemic chains within the household as random variables. However, with final size data and the order of occurrence of infection events being unknown, integrating over all possible chains may not be analytically or numerically feasible. Data augmentation procedures can be used for inference in these settings (O'Neill, 2009, Knock & O'Neill, 2014).

(iii) Final size of outbreak in a population of households

The expected total number of individuals infected during the outbreak in the population of households is given by

$$w' = \sum_{k=1}^{K} m_k \sum_{j=1}^{k} j P_k(j|\pi_c, \pi_h). \tag{5.4}$$

For a population of size $N = \sum_{k=1}^{K} k m_k$ we then have $w = w'/N$ as the expected proportion of individuals infected during the outbreak in the population of households.

(iv) Implications for outbreak simulation

When there is no direct transmission, i.e. $\pi_h = 1$, all infections are due to the environment. Hence a lower bound (w'_l) for the final size (w') can be obtained as

$$w'_l = \sum_{k=1}^{K} m_k \sum_{j=1}^{k} j \binom{k}{j} (1-\pi_c)^j (\pi_c)^{k-j}. \tag{5.5}$$

Arguing similarly, when individuals are highly infectious, the probability of avoiding infection from an infectious household member decreases ($\pi_h \to 0$). For a household of size k, on becoming infected with probability $(1-\pi_c^k)$ (see equation (5.2)), all k members of the household will be infected. Hence an upper bound (w'_u) for the final size (w') can be obtained as

$$w'_u = \sum_{k=1}^{K} m_k k (1-\pi_c^k). \tag{5.6}$$

On using the limits of direct transmission parameter ($0 < \pi_h \leq 1$) the final size is seen to be bounded by $w'_l \leq w' \leq w'_u$ where the upper and lower bounds are governed by $\pi_c = e^{-\mu\tau^1}$. Thus μ and τ^1 jointly contribute to the final size of the outbreak through the cumulative risk $1-\pi_c$.

When multiple parameters jointly define simulation settings, the analysis such as above enable to identify smaller parameter groups. Fixing some parameters within the groups and scaling others with respect to the fixed parameter leads to simulation settings defined by smaller number of parameters. This allows efficient generalisation of the simulation settings.

From a simulation viewpoint, with respect to the role of infection from the environmnent, the parameterisation is one-dimensional in that μ and π_c have a one-to-one correspondence for a fixed τ^1. Thus, as long as π_c is constant, i.e., $\mu\tau^1$ is constant, the probability model for infection from the environment is the same, irrespective of the individual values of μ and τ^1. Therefore results from simulation settings can be generalised by varying μ for fixed τ^1.

In practise, however, τ^1 is fixed (or approximated) in the observed data and is conditioned upon in the estimation of μ. As shown in the case of μ and π_c, one can similarly argue that λ and π_h have a one-to-one correspondence for a fixed duration of infectiousness T_I.

5.4 Outbreak time points and their approximations

This section discusses the outbreak time points defined for the model used in this application (see Section 2.1 and Figure 2.1) and how they are approximated in the inference procedures presented in Section 5.2. Other ways of handling them are also discussed with reference to relevant literature.

Unobserved outbreak time points and their approximation
The time point τ^0 when the environment becomes infectious was assumed to be known when analysing the simulated data in Section 5.2. In practise this time point is not observed. For an observed outbreak, some time point $T^0 \leq \min(t_i^S) - (l_{min} + s_{min})$, defined based on $\min(t_i^S)$, the earliest symptom time in the population, can be used as an approximation. In this application, T^0 was set to the one used in simulation (τ^0) (see, for example, Yang et al. (2007a) for a method to analyse outbreak data when τ^0 is known and Yang et al. (2007b) for application of the method to avian influenza A (H5N1) when τ^0 is known).

The estimability of c (or equivalently μ) depends on T^0 as transmission events are intrinsically conditioned up on this time point. The case-ascertained follow-up design in section 2.5.3 is an exception as the time period from τ^0 to t' (the earliest possible infection time of the index case in the household) does not contribute to the conditional likelihood.

In the context of Bayesian inference the unknown τ^0 can also be jointly sampled along with the parameters of interest. This requires specification of a model $P(\tau^0|\mathbf{z}, \mathbf{y}, \theta)$ and defining a prior distribution for τ^0. O'Neill & Roberts, 1999 jointly estimate the infection time of the index case in the population (a time-point with similar implications as τ^0) in the general stochastic epidemic model setting.

For the models of Sections 2.5.1 and 2.5.3 (completed outbreak), the time of the end of outbreak τ_N is approximated by T_c. This seems to be a reasonable approximation for the end of outbreak as no new symptomatic infections are observed beyond T_c, especially when $\eta \approx 1$.

Similarly, τ^1, the time when the environment stops being infectious, is not observed. An approximate value T^1 can be provided given T_c is observed (for example, $T^1 = T_c$ is used in this application). On a related note, in the simulation scenarios, τ^1 was chosen to be smaller than t_w, the threshold for start of antibody waning since time of infection. This was based on preliminary examinations wherein the posterior estimates of the transmission parameters were poor when $\tau^1 > t_w$ for imperfectly observed symptoms.

5.5 Parameter redundance, identifiability and misclassification

The inference procedures in chapters 3 and 4 assume that the sojourn time distributions $g(\cdot)$, $u(\cdot)$ and $v(\cdot)$ and the observation model (parameters characterising the symptom data (η, ψ) and the model for serology (α, t_w)) are fully known. In addition, the outbreak timelines are approximated by other observed timelines. Only the transmission model parameters (μ, λ) are estimated.

Identifiability
Estimation of parameters from such complex models is a non-trivial task and some questions are in order when inferring the model parameters from data. Are there redundant parameterisations in the model which would lead to non-unique solutions, i.e. is the model *identifiable*? Analysis undertaken in the previous section describes the relationship among μ, τ^1 and the outbreak final size. This explains why the unknown τ^1 should be fixed by design and approximated using T^1 instead of being estimated.

Similar arguments could also be laid out for λ and the sojourn time distributions. Yang et al. (2009) report in an SIR model with a similar observation setting that distributions of incubation and infectious periods cannot be jointly estimated in addition to model parameters associated with transmission.

When $\eta = 1$ and $\psi = 1$ the observational model is completely deterministic. The likelihood then corresponds to equation (2.11) which is identifiable. On the other hand, when $\eta < 1$ but unknown, it is already complicated due to joint identifiability of (μ, λ, η). Estimating these parameters depend not only on the proportion of households with final size ≥ 2, but also on the serological data.

In general, if the study is designed to estimate the effects of covariates, vaccine or antiviral treatment effects, additional care and design considerations are required to ensure the identifiability of model parameters from the observed data.

Misclassification

In this application, the complete data model was utilised by introducing latent infection times (t_i^E). Inference methods using data augmentation procedures to account for latent variables can have *misclassification errors* in augmented data when there is dependence among model parameters. For example, when symptoms are imperfectly observed, the infection times are sampled from a mixture. Dependence among c, q and t_i^E may affect the sampled infection times. Implementing blocking strategies for updating of latent data and transmission parameters may reduce misclassification, albeit, with additional effort due to developing the required sampling algorithms and the associated computational complexity.

Chapter 6

Discussion and Conclusion

"How often have I said to you that when you have excluded the impossible, whatever remains, however improbable, must be the truth?"
- Sir Arthur Conan Doyle (Sherlock Holmes in Sign of Four)

6.1 Discussion of Results

6.1.1 What was the problem?

The main aim of this thesis was to address the following question: "Can person-to-person transmission be inferred using partially observed data from an outbreak in a population of households?". Hepatitis A virus infection was used to motivate the state-space of the model due to its complex natural history of infection. As is common in such outbreaks, the partial observation of infection process is due to unobserved infection times, presence of asymptomatic infections and not fully sensitive serological test results. We assumed that the observations include symptom onset times and serological test results at the end of outbreak. It was further assumed that the observed data corresponds to a completed outbreak in the entire population of households (full cohort).

6.1.2 What has been done?

Model specification
The state-space of the model consisted of five states: susceptible, exposed, infectious, symptomatic and recovered. The specification of the transmission model accounted for infection from outside the household due to the environment and from within the household due to infectious household members. Two model parameters governed the two sources of infection. The specification of the observation model accounted for observing serological test results at the end of outbreak, proportion of symptomatic cases and their symptom onset times.

Furthermore, two models corresponding to two hypotheses for person-to-person transmission were considered: the full model (with both sources of infection - environment and infectious household members) and the null model (with environment as the only source of infection).

Simulation of outbreak datasets
Outbreak datasets were simulated based on the specified models for 24 combinations (2 models corresponding to the 2 hypotheses, 3 settings based on sensitivity and specificity of symptoms, 2 sets of transmission parameter values and 2 household size distributions in the population). After simulation of the outbreak datasets, only information pertaining to that of partially observed outbreaks were further in statistical inference. Bayesian inference procedures were performed on each of the 24 datasets under both the full and the null models.

Parameter estimation and model comparison
The unknown infection times were considered as latent variables and were included in the complete-data likelihood representation. Under the Bayesian inference framework, posterior estimates of the two model parameters that govern the transmission process were obtained using Markov chain Monte Carlo (MCMC) sampling for all the simulated datasets. The unknown infection times were also jointly sampled as part of the data-augmentation procedure.

In addition to estimating the two transmission parameters, the two fitted models (the full and the null models) were compared using a version of the deviance information criteria (DIC) for each dataset from the 24 combinations. This version of DIC, namely DIC_6, is based on missing data model that uses the complete-data likelihood respresentation as well as having the required focus of inference (focus on the transmission parameters and not on missing data).

6.1.3 What were the findings and their implications?

Findings
The marginal posterior estimates of the two transmission parameters were presented as mean and 95% highest posterior density (HPD) intervals (Tables 5.2 to 5.4). The marginal posterior estimates were quite consistent with the underlying model that generated the datasets, in that, for the datasets simulated under the null model, the model parameter for person-to-person transmission included the null value ($q = 1$) in the 95% HPD intervals in each of the 4 datasets in all 3 settings. And for the datasets simulated under the full model, the upper limits of the 95% HPD intervals for the model parameter for person-to-person transmission were consistently below 1.

The DIC_6 values for datasets simulated under the null model were similar for both the null and the full models for all 4 datasets in each setting. This was due to the marginal posterior of q estimated correctly in the neighbourhood of one under the null model. With similar marginal posterior estimates under both models, the DIC_6 values were similar as expected.

For datasets simulated under the full model, more often, the DIC_6 values for the full model were lower than the DIC_6 values for the null model. This occurred in 3 out of 4 datasets in settings 1 and 2 (i.e. perfectly observed symptoms and non-sensitive but fully specific symptoms) and in all 4 datasets in setting 3 (non-sensitive and non-specific symptoms).

Implications from the findings
Results from both marginal posterior estimates of the two transmission parameters and the corresponding DIC_6 values indicate towards the correct model that was used for simulating a dataset. Especially, for the datasets simulated under the null model, the results are highly consistent. The performance of parameter estimation and DIC_6 are not different between the two chosen household size distributions (households of size of 2 and households of sizes 1, 2 and 3). It seems plausible that these procedures would perform similarly with other household size distributions, including those with larger household sizes than considered here.

For the datasets simulated under the full model, the results are consistent for the tansmission parameter values $(q, c) = (0.95, 0.99)$ in all three settings. However, for the tansmission parameter values $(q, c) = (0.95, 0.97)$, the results are less consistent in all three settings. The reason for this behaviour can be seen from Figure 5.1 and its interpretation in Section 5.1, i.e. the effect of q is more pronounced and easily distinguished for higher values of c such as 0.99 than a lower value 0.97. This is because, for higher values of c, the environment merely introduces the infection to the household and the within-household transmission is propogated through the effect of q. However, for lower values of c, the environment continues to contribute to the household final size significantly even after the initial infections within the household. Further analytical and simulation-based studies would be required to look at the effect of q *per se* and the ratio (q/c). It should be noted, however, that model comparison results using DIC_6 were not very different between the scenarios of $c = 0.97$ and $c = 0.99$.

The observation setting governed by sensitivity and specificity of symptoms (η, ψ) plays a key role in contributing to the complexity of partial observation of outbreaks. This can be seen in the performance of estimation of the two transmission parameters: the results were more consistent with values used in simulation in the perfectly observed setting ($\eta = 1, \psi = 1$) than in the imperfectly observed settings ($\eta < 1, \psi = 1$ and $\eta < 1, \psi < 1$). However, the performance of model comparison using DIC_6 was consistent across all the three observation settings.

For the purpose of model comparison, which is the primary aim of the application in this thesis, DIC_6 consistently indicates towards the correct model in almost all scenarios and is deemed robust with respect to the simulation scenarios used. Additionally, the estimation of the transmission parameters through data augmented-MCMC procedures, which is reasonably consistent across all simulation scenarios, provides methods to sample from the joint posterior distribution. Together, they offer methods to analyse partially observed outbreaks in households for infections that closely follow the natural history used in this application and observation settings that fall within the context of the presented simulation scenarios.

6.2 Design aspects

6.2.1 Study designs for epidemics in households

As households form important observation units in the study of infectious diseases, understanding the associated study designs is crucial to inform methods to collect data. The observation design primarily considered in this thesis is the completed outbreak in a full cohort which is a more general set-up. The model and inference procedures in this set-up can be further extended to any modifications in the study design.

Chapter 2 on modelling outbreak discusses some adaptations of the full cohort design. The case-ascertained follow-up design are ubiquitous in the epidemic modelling literature as data from such designs are plentiful. Research concerning the design aspects of such epidemic data from households are few (as outlined in Section 1.2) and need further exposition to fully understand the implications of this design.

Other design adaptations such as data from an ongoing outbreak are also important, especially when the pathogen poses a potential for pandemic (like the case of influenza from zoonotic origins) or large scale epizootics (such as foot-and-mouth diease outbreaks in the UK). Here, the need for inference on the transsmission parameters in real-time is key for infection management and control.

It should be noted, however, that outbreak in progress and case-ascertained follow-up designs are only possible under the assumptions that symptoms are sensitive and specific. The completed outbreak in whole population is a more general case with respect to observation of symptoms.

6.2.2 Data observed from the outbreak

The current application assumes that observed data would include symptom times and serological test result (based on IgM antibodies that wane over time) at a specific outbreak time point T_s. Chapter 2 provides a model for serology also based on IgG antibodies, although not used in this application. The model can be extended to include laboratory test data and other time-varying covariates collected over the duration of outbreak. Simulations can be performed in studying the optimal observation time points for these variables that maximise information towards more accurate inference.

The current application assumes that the observation model and the natural history of infection are known. While this might be reasonable to assume for endemic or well studied pathogens such as hepatitis A virus, for a novel pathogen, this can be a challenging assumption. When these quantities are unknown, justifiable assumptions on those quantities along with appropriate sensitivity analyses for the assumptions can be used.

6.3 Potential extensions to data and modelling

The data described in this application can be extended to include other features. Here we consider two such extensions: inclusion of prior immunity or vaccination and availability pathogen sequence data from individuals infected during the outbreak.

6.3.1 Immunity and Vaccination

For some applications, it is possible that a certain proportion of the population is already immune before the outbreak onset, especially in populations where the disease is endemic. If the immune status is not observed for the individuals before the outbreak onset, the proportion of immunes is unknown and should be estimated.

Parameters for time-dependent vaccination can be included as covariates as briefly discussed in Section 2.3 (following equation (2.4)). In either case, whether it is to account for prior immunity or vaccination, additional precaution must be taken in terms of observation design and model formulation to ensure that the associated parameters are identfiable along with the transmission parameters.

6.3.2 Pathogen sequence data

Many recent outbreak datasets include pathogen genetic sequence data from infected individuals along with event times, symptoms and laboratory test results data. When such data are available, the model can include components that account for pathogen evolution between infected individuals (for example, a metric to compare between pathogen genetic sequences for any two individuals). This would increase the accuracy in inferring transmission parameters. Additionally, the underlying transmission tree describing the sequence of transmission events can be inferred (Morelli et al., 2012, Klinkenberg et al., 2017).

6.3.3 Structures within and between households

The models presented in Chapter 2 assumed homogeneity of susceptibility and infectivity. When more detailed covariates are available to stratify individuals (for example, based on age, comorbidity, contact patterns or immunity), the present model can be extended to a multitype model.

One of the crucial assumptions in models of Chapter 2 is that of households being independent, i.e., no transmission between households. Endo et al. (2019) describes this as a pseudo-likelihood assumption wherein interaction among households is neglected. It should be noted that more information about contacts between households and the neighbourhood structure are required to account for between household transmission. Additional information is also required to discern infections due to between household and community transmission.

Yang et al. (2007a) and Yang et al. (2007b) use models with three parameters with respect to transmission from the community, from those within the household or close contacts and from those outside the household or casual contacts. This requires more specific data such as residential location (neighbourhood and household) to estimate the parameters (Yang et al., 2007b). However, it should be noted that the natural history of infection and observation setting in the cited studies are simpler than the ones used in the application presented in this thesis.

6.4 Some aspects of Bayesian inference

6.4.1 Prior and posterior predictive distributions

The performance of methods developed in this thesis were tested using simulated outbreak data. The uniform distribution was used as the priors for the two transmission parameters c and q. As pointed out in Section

3.4.4, the transmission parameters in the continuous time model can use gamma priors in view of conjugacy towards sampling from the posterior density.

6.4.2 Model comparison

Bayesian model comparison for partially observed stochastic epidemic models has been a growing focus of research in recent years (Gibson et al., 2018). In this thesis, we used deviance information criteria (DIC) based on a complete-data model representation for choosing between the full ($q < 1$) and the null ($q = 1$) models. A key drawback that is still unaddressed in DIC in general is the possibility of obtaining negative effective numbers of parameters.

Other methods such as Bayes factors (BF) and procedures based on posterior predictive distributions have also been used for model comparison. When compared to such methods, DIC has been considered less Bayesian (Gibson et al., 2018) as its interpretation is not based on posterior probability directly. DIC is considered as a measure of fit and complexity.

As pointed out in Section 4.1.2, computing BF for partially observed stochastic epidemic models is not straightforward as difficulties arise in implementing the required RJ-MCMC procudures. Computing BF using other approaches such as the one in Alharthi et al. (2018) should be developed and evaluated against DIC for its performance on model choice.

6.5 Conclusions

The presented stochastic modelling framework in Chapter 2 is useful for modelling outbreak data from first principles based on an epidemiologically meaningful parameterisation. The outbreak data model can be adapted to other study designs and observed variables besides the settings presented in Chapter 2. Such models yield well to probabilistic inference methods towards estimation of model parameters and model comparison.

The data augmented Markov chain Monte Carlo procedures for sampling from the posterior distribution presented in Chapter 3 are useful in the presence of latent variables in the model. The procedures presented in Chapter 4 are useful in computing DIC (i.e., DIC_6) under complete-data representation with focus only on the unknown model parameters, but can be extended to other focus of inference (i.e., DIC_4 or DIC_5). Both the parameter estimation and model comparison procedures can be modified based on complete, observed or conditional data model representations as described in Chapters 3 and 4.

From the results using simulated outbreak datasets, the Bayesian estimation procedures from Chapter 3 provide estimates of transmission parameters that are quite consistent with values used in the simulation. The DIC_6 presented in Chapter 4 consistently indicates towards the correct model in almost all simulation scenarios and is robust across all the presented simulation scenarios.

The model and inference procedures are useful in inferring the presence of person-to-person transmission from outbreak data in households. The model and inference procedures are flexible and can be modified as necessary to the appropriate study design and focus of inference.

Appendix A

Discrete-time transmission model and complete-data likelihood

In discrete-time, the model is specified in the time unit of days (see Section 2.6 for details). The model formulation follows that of Rampey et al. (1992).

Discrete-time transmission model (for Section 2.3)

The probability that an individual i is not infectious on a given day is denoted as $W_i(t|\mathcal{F}_{it-1})$. Let $W_i^{m=1}(t|\mathcal{F}_{it-1})$ be the probability that an individual i with symptomatic infection is not infectious on day t, given the history $\mathcal{F}_{it}$:

$$W_i^{m=1}(t|\mathcal{F}_{it-1}) = \begin{cases} 1 & \text{if} \quad X_i(t-1) = U, \\ 1 - \sum_{\tau=t_i^E+l_{min}}^{t-1} g(\tau|t_i^E) & \text{if} \quad X_i(t-1) = E, \\ \sum_{\tau=t_i^S}^{t-1} v(\tau|t_i^S) & \text{if} \quad X_i(t-1) = S. \end{cases} \tag{A.1}$$

Let $W_i^{m=0}(t|\mathcal{F}_{it-1})$ be the probability that an individual i with asymptomatic infection is not infectious on day t, given the history $\mathcal{F}_{it}$:

$$W_i^{m=0}(t|\mathcal{F}_{it-1}) = \begin{cases} 1 & \text{if} \quad X_i(t-1) = U, \\ \{1 - \sum\limits_{\tau=t_i^E+l_{min}}^{t-1} g(\tau|t_i^E)\} + \sum\limits_{\tau=t_i^E+l_{min}}^{t-1} g(\tau|t_i^E)F(t|\tau) & \text{if} \quad X_i(t-1) \geq E, \end{cases} \tag{A.2}$$

where $F(t|\tau) = \sum\limits_{u=\tau}^{t} f(u|\tau)$. Obviously, the probability of an individual i not being infectious before being infected is given by $W_i(t|\mathcal{F}_{it-1}) = 1$.

Let q be the probability that a still susceptible individual escapes infection from a single infectious person within his/her household on a single day. The probability that a still susceptible individual i escapes infection from individual j on day t is given as

$$q_{ij}(t|\mathcal{F}_{it-1}, \mathcal{F}_{jt-1}) = W_j^m(t|\mathcal{F}_{jt-1}) + [1 - W_j^m(t|\mathcal{F}_{jt-1})]q. \tag{A.3}$$

Let c be the probability that a still susceptible individual escapes infection from the environment on a single day. This probability c is constant as long as the environment is infectious. If the environment is infectious up to τ^1, we define the infectiousness of the environment at any time t as

$$c(t|\tau^1) = 1_{\{t>\tau^1\}} + [1_{\{t\leq\tau^1\}}]c. \tag{A.4}$$

It is assumed that a susceptible individual escapes infection from all sources independently. The probability that a still susceptible individual i escapes infection from all infective sources on day t is $e_i(t|\mathcal{H}_{it-1})$ and is expressed as

$$e_i(t|\mathcal{H}_{it-1}) = c(t|\tau^1) \prod_{\substack{j\in H_i, \\ j\neq i}} q_{ij}(t|\mathcal{F}_{it-1}, \mathcal{F}_{jt-1}). \tag{A.5}$$

The probability that an individual is infected on day t is given by

$$Z_i(t|\mathcal{H}_{it-1}) = \prod_{u=\tau^0}^{t-1} e_i(u|\mathcal{H}_{iu-1})\{1 - e_i(t|\mathcal{H}_{it-1})\}. \tag{A.6}$$

The probability that an individual is not infected during the outbreak up to some time t is given by

$$Q_i(t|\mathcal{H}_{it-1}) = \prod_{u=\tau^0}^{t} e_i(u|\mathcal{H}_{iu-1}). \tag{A.7}$$

Discrete-time complete-data likelihood (for Section 2.5.1)
The parameters of interest are $\theta = (c, q)$. Given the history $\mathcal{H}_{iT_c}$ of all household members of individual i up to day T_c for the complete data C_i, the likelihood contribution from an individual i, $L_i(\theta; \mathcal{H}_{iT_c})$ is,

$$= \begin{cases} \eta R(T_c - t_i^E) \sum\limits_{t_i^I = t_i^E + l_{min}}^{min(t_i^S, t_i^E + l_{max})} u(t_i^S|t_i^I) g(t_i^I|t_i^E) Z_i(t_i^E|\mathcal{H}_{it_i^E-1}), & \text{if } C_i{=}(t_i^E{<}T_c, m_i{=}1, t_i^S{<}T_c, a_i{=}1), \\ (1-\eta) R(T_c - t_i^E) Z_i(t_i^E|\mathcal{H}_{it_i^E-1}), & \text{if } C_i{=}(t_i^E{<}T_c, m_i{=}0, t_i^S{=}T_c{+}1, a_i{=}1), \\ \eta(1 - R(T_c - t_i^E)) \sum\limits_{t_i^I = t_i^E + l_{min}}^{min(t_i^S, t_i^E + l_{max})} u(t_i^S|t_i^I) g(t_i^I|t_i^E) Z_i(t_i^E|\mathcal{H}_{it_i^E-1}), & \text{if } C_i{=}(t_i^E{<}T_c, m_i{=}1, t_i^S{<}T_c, a_i{=}0), \\ (1-\eta)(1 - R(T_c - t_i^E)) Z_i(t_i^E|\mathcal{H}_{it_i^E-1}), & \text{if } C_i{=}(t_i^E{<}T_c, m_i{=}0, t_i^S{=}T_c{+}1, a_i{=}0), \\ (1-\psi) Q_i(T_c|\mathcal{H}_{iT_c-1}), & \text{if } C_i{=}(t_i^E{=}T_c{+}1, m_i{=}1, t_i^S{<}T_c, a_i{=}0), \\ \psi Q_i(T_c|\mathcal{H}_{iT_c-1}), & \text{if } C_i{=}(t_i^E{=}T_c{+}1, m_i{=}0, t_i^S{=}T_c{+}1, a_i{=}0). \end{cases} \tag{A.8}$$

References

1. Alharthi M, Kypraios T, O'Neill PD (2018). Bayes factors for partially observed stochastic epidemic models. Bayesian Analysis.

2. Andersen PK, Borgan O, Gill RD, Keiding N (1993). Statistical Models based on Counting Processes. Springer, New York.

3. Andrieu C, Roberts GO (2009). The psuedo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics 37 (2): 697-725.

4. Ball F, Lyne O. (2002) Epidemics among a population of households. In Mathematical Approaches for Emerging and Reemerging Infectious Diseases. Part II: Models, Methods, and Theory Volume 126: 115-42.

5. Ball F, Lyne O. (2006) Optimal vaccination schemes for epidemics among a population of households, with application to variola minor in Brazil. Statistical Methods in Medical Research 15: 481-497.

6. Ball FG, Mollison D, Scalia-Tomba G (1997). Epidemics with two levels of mixing. Annals of Applied Probability 7: 46-89.

7. Beaumont MA (2003). Estimation of population growth and decline in genetically monitored populations. Genetics 164: 1139-1160.

8. Becker NG (1989). Analysis of infectious disease data. London: Chapman and Hall

9. Becker NG, Britton T (1999). Statistical studies of infectious disease incidence. Journal of the Royal Statistical Society Series B 61 (2): 287-307.

10. Becker NG, Hasofer AM (1997). Estimation in epidemics with incomplete observation. Journal of the Royal Statistical Society series B 59 (4): 415-429.

11. Britton T (1997). Test to detect clustering of infected individuals within families. Biometrics 53: 98-109.

12. Britton T (1998). Estimation in multitype epidemics. Journal of the Royal Statistical Society series B 60 (4): 663-679.

13. Britton T (2010). Stochastic epidemic models: a survey. Mathematical Biosciences 225 (1): 24-35.

14. Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boelle PY (2004). A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statistics in Medicine 23 (22): 3469-87.

15. Cauchemez S, Ferguson NM (2011). Methods to infer transmission risk factors in complex outbreak data. Journal of the Royal Society Interface.

16. Celeux G, Forbes F, Robert CP, Titterington DM (2006) Deviance information criteria for missing data models. Bayesian Analysis 1 (4):651-674.

17. Davison AC. (2003). Statistical Models. Cambridge University Press, Cambridge.

18. Demiris N, O'Neill PD (2005). Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. Journal of the Royal Statistical Society, Series B 67: 731-746.

19. Endo A, Uchida M, Kucharski AJ, Funk S (2019). Fine-scale family structure shapes influenza transmission risk in households: Insights from primary schools in Matsumoto city, 2014/15. PLoS Computational Biology 15(12):e1007589.

20. Fintzi J, Cui X, Wakefield J, Minin VN (2017). Efficient data augmentation for fitting stochastic epidemic models to prevalence data. Journal of Computational and Graphical Statistics 26 (4): 918-929.

21. Forrester M, Pettitt A, Gibson G (2007). Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data. Biostatistics 8: 383-401.

22. Fraser C (2007) Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. PLoS ONE 2(8): e758.

23. Gibson GJ, Streftaris G, Thong D (2018). Comparison and assessment of epidemic models. Statistical Science. 33(1), 19-33. DOI: 10.1214/17-STS615

24. Gordon A, Tsang TK, Cowling BJ, Kuan G, Ojeda S et al (2018). Influenza transmission dynamics in urban households, Managua, Nicaragua, 2012-2014. Emerging Infectious Diseases 24 (10): 1882-1888 (along with technical appendix).

25. Gough KJ (1977). The estimation of latent and infectious periods. Biometrika 64 (3): 559-565.

26. Grassly NC, Fraser C (2008). Mathematical models of infectious disease transmission. Nature Reviews Microbiology 6 (6): 477-487.

27. House T, Keeling MJ (2008). Determinisitc epidemic models with explicit household structure. Mathematical Biosciences 213: 29-39.

28. House T, Inglis N, Ross JV, Wilson F, Suleman S, Edeghere O, et al. (2012). Estimation of outbreak severity and transmissibility: Influenza A(H1N1)pdm09 in households. BMC Medicine 10: 117.

29. Kinjanjui TM, Pellis L, House T. (2016) Information content of household-stratified epidemics. Epidemics 16: 17-26.

30. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput Biol 13(5): e1005495.

31. Klinkenberg D, Nishiura H (2011). The correlation between infectivity and incubation period of measles, estimated from households with two cases. Journal of Theoretical Biology 284 (1): 52-60.

32. Knock ES, Kypraios T (2018). Bayesian non-parametric inference for infectious disease data. arXiv:1411.2624 [stat.ME]. https://arxiv.org/pdf/1411.2624.pdf

33. Knock ES, O'Neill PD (2014). Bayesian model choice for epidemic models with two levels of mixing. Biostatistics 15 (1): 46-59.

34. Kumbang J, Ejide S, Tedder S, Ngui SL (2012). Outbreak of hepatitis A in an extended family after importation by non-immune travellers. Epidemiology and Infection 140: 1813-1820.

35. Kypraios T, Minin VN (2018). Introduction to the special section on inference for infectious disease dynamics. Statistical Science 33 (1): 1-3.

36. Kypraios T, Neal P, Prangle D (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. Mathematical Biosciences 287: 42-53.

37. Lau MSY, Cowling BJ, Cook AR, Riley S (2015). Inferring influenza dynamics and control in households. Proceedings of the National Academy of Sciences 112 (29) 9094-9099.

38. Lima LR, Almeida AJD, Tourinho RdS, Hasselmann B, Lewis Ximenez LL, De Paula VS (2014). Evidence of Hepatitis A Virus person-to-person transmission in household outbreaks. PLoS ONE 9 (7): e102925.

39. Longini IM, Koopman JS (1982) Household and community transmission parameters from final distributions of infections in households. Biometrics 38: 115-126.

40. McKinley TJ, Ross JV, Deardon R, Cook AR (2014) Simulation-based Bayesian inference for epidemic models. Computational Statistics and Data Analysis 71: 434-447.

41. Morelli MJ, Thébaud G, Chadouf J, King DP, Haydon DT, Soubeyrand S (2012) A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. PLoS Comput Biol 8(11): e1002768.

42. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. (2008) Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. PLoS Med 5(3): e74.

43. O'Neill PD, Roberts GO (1999). Bayesian inference for partially observed stochastic epidemics. JRSS-A 162 (1): 121-129.

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

45. O'Neill PD (2009). Bayesian inference for stochastic multitype epidemics in structured populations using sample data. Biostatistics 10 (4): 779-791.

46. O'Neill PD, Balding DJ, Becker NG, Eerola M, Mollison D. (2000) Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. Applied Statistics 49 (4): 517-542.

47. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

48. Rampey AH, Longini IM, Haber M, Monto AS (1992) A discrete-time model for the statistical analysis of infectious disease incidence data. Biometrics 48 (1): 117-128.

49. Rhodes PH, Halloran ME, Longini IM (1996) Counting process models for infectious disease data: Distinguishing exposure to infection from susceptibility. JRSS-B 58 (4): 751-762.

50. Robert CP (2014). Bayesian Computational Tools. Annual Review of Statistics and Its Application 1: 153-177.

51. Robert CP, Casella G (2004). Monte Carlo Statistical Methods. Springer.

52. Robert CP, Casella G (2010). Introducing Monte Carlo Methods with R. Springer

53. Sato T (1988). Sequentially-occurring transmission of hepatitis A in a family. Tohoku Journal of Experimental Medicine 155 (4): 387-388.

54. Shaw LM. (2016) SIR epidemics in a population of households. PhD thesis, University of Nottingham.

55. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, series B 64: 583-616.

56. Streftaris G, Gibson GJ (2004) Bayesian inference for stochastic epidemics in closed population. Statistical Modelling 4: 63-75.

57. Tanner MA, Wong WH (2010). From EM to data augmentation: The emergence of MCMC Bayesian computation in the 1980s. Statistical Science 25 (4): 506-516.

58. Victor JC, Surdina TY, Suleimenova SZ, Favorov MO, Bell BP, Monto AS (2006). Person-to-person transmission of hepatitis A virus in an urban area of intermediate endemicity: Implications for vaccination strategies. American Journal of Epidemiology 163 (3): 204-210.

59. Yang Y, Longini IM, Halloran ME. (2006) Design and evaluation of prophylactic interventions using infectious disease incidence data from close contact groups. Applied Statistics 55 (3): 317-330.

60. Yang Y, Longini IM, Halloran ME (2007a). A resampling-based test to detect person-to-person transmission of infectious disease. The Annals of Applied Statistics 1 (1): 211-228.

61. Yang Y, Halloran ME, Sugimoto JD, Longini IM (2007b). Detecting human-to-human transmission of avian influenza A (H5N1). Emerging Infectious Diseases 13 (3): 1348-53.

62. Yang Y, Halloran ME, Longini IM (2009a). A Bayesian model for evaluating influenza antiviral efficacy in household studies with asymptomatic infections. Biostatistics 10 (2): 390-403.

63. Yang Y, Sugimoto JD, Halloran ME, Basta NE, Chao DL, Matrajt L, et al. (2009b). The transmissibility and control of pandemic influenza A (H1N1) virus. Science 326 (5953): 729-733.

64. Zhang X-S, Iacano GL (2018). Estimating human-to-human transmissibility of hepatisis A virus in an outbreak at elementary school in China, 2011. PLoS ONE 13 (9): e0204201.