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

SIR epidemics in a population of households

Thesis submitted to The University of Nottingham for the degree of Doctor of Philosophy

Laurence Matthew Shaw, MSc.

December 2016

Abstract

The severity of the outbreak of an infectious disease is highly dependent upon the structure of the population through which it spreads. This thesis considers the stochastic SIR (susceptible → infective → removed) household epidemic model, in which individuals mix with other individuals in their household at a far higher rate than with any other member of the population. This model gives a more realistic view of dynamics for the transmission of many diseases than the traditional model, in which all individuals in a population mix homogeneously, but retains mathematical tractability, allowing us to draw inferences from disease data.

This thesis considers inference from epidemics using data which has been acquired after an outbreak has finished and whilst it is still in its early, 'emerging' phase. An asymptotically unbiased method for estimating within household infectious contact rate(s) from emerging epidemic data is developed as well as hypothesis testing based on final size epidemic data. Finally, we investigate the use of both emerging and final size epidemic data to estimate the vaccination coverage required to prevent a large scale epidemic from occurring. Throughout the thesis we also consider the exact form of the households epidemic model which should be used. Specifically, we consider models in which the level of infectious contact between two individuals in the same household varies according to the size of their household.

Acknowledgements

I would first like to thank Professor Frank Ball who has provided me with invaluable advice and direction throughout my studies. He has also introduced me to a fascinating area of mathematics in which I expect I shall always retain a keen interest. I would also like to acknowledge the contributions of my other supervisor, Dr Theo Kypraios, as well as Professor Philip O'Neill and Dr David Hodge. They have all have reviewed my work at various stages over the past four years and have made important suggestions to improve it. The help of Theo and Frank in finding me short-term statistical work whilst writing up my thesis is also greatly appreciated.

I am grateful for the continued support of my family, especially my Mum, Dad, Mary, Phil and Paige. Special thanks to Cerian for putting up with me during the writing up stage of the thesis. I am also thankful to my many friends in Nottingham, both in the mathematics department and on the various pool tables in the local area, who have helped me to feel comfortable and enjoy life in my adopted city. I have discovered that having a strong set of family and friends is of great importance at times when research isn't going your way.

Finally, my gratitude is extended to the EPSRC, whose funding made this research possible.

Contents

CHAPTER 1

Introduction

This introduction seeks to give the reader a detailed but non-technical insight into the purpose of this thesis by providing a historical context. Section 1.1 offers a motivation for the involvement of mathematicians in epidemiology. The initial breakthroughs of mathematical epidemiologists are explored in Section 1.2. An insight into the early development of the specific model used in this thesis is given in Section 1.3 before Section 1.4 considers early ideas in mathematics surrounding the prevention of epidemics, most notably through vaccination schemes. Section 1.5 places the work of this thesis in its current context, discussing recent literature that is closely related to this thesis and literature detailing other mathematical ideas currently being implemented in epidemiology. Finally, an outline of the thesis and a very brief summary of the key results is given in Section 1.6.

1.1 Motivation

From the Athenian epidemic of approximately 430 B.C. to the outbreaks of malaria, dengue fever, AIDS and Ebola that still affect the world today, communicable disease has been a one of the greatest scourges to affect the history of mankind. Perhaps the most startling example of this was the Spanish influenza pandemic of 1918, which is estimated to have killed around 75 million people worldwide, dwarfing even the 37 million casualties of the Great War from the four previous years. Whilst it seems clear than one cannot prevent disease from occurring altogether, it is worth asking whether it is possible to eradicate certain diseases, such as smallpox, and to at least mitigate the effects of those that cannot be eradicated (e.g. influenza). Short of placing an entire population in quarantine at the first sign of an epidemic, this may still seem to be infeasible. However, one must wonder whether the Black Death of the 14th century would have been quite so deadly (some estimates suggest that 70% of Europeans were killed) had people understood the airborne nature of the disease and been careful to avoid coughing or sneezing on others. By contrast, the work of John Snow on the Broad Street epidemic of 1854 has virtually eradicated the possibility of a cholera epidemic in countries where the water supply is sanitised properly and it is worth considering just how many lives this has saved.

Despite being a physician, Snow used mathematical methods to trace the cause of the epidemic and, in the case of a disease such as cholera where something as simple as clean water can stop an outbreak altogether, there proved to be no need for further mathematical input. In 1760 however, the only recognised method available to prevent the spreading of smallpox was variolation, an early and less effective form of vaccination in which subjects were infected with a mild form of smallpox in the hope of inducing immunisation to more fatal variants. In this year Daniel Bernoulli submitted a paper to the Academy of Sciences in Paris investigating the effectiveness of variolation. Dietz and Heesterbeek [2002] show how formulae within this paper can be used to calculate the increase in life expectancy as the result of a proportion of a population being successfully immunized. Although this was the only known mathematical work of note prior to the late 19^{th} century on the spread of infectious disease, this paper does give an indication as to the place of mathematics within epidemiology and illustrates the general idea behind the mathematical approach used in this thesis in which is to parameterise an epidemic in order to assess the extent to which it spreads among a population. One can then assess the potential impact of a given intervention strategy, with the hope of eventually choosing a strategy that prevents a major outbreak from occurring.

It should also be noted at this point that curtailing the severity of an epidemic has huge economic as well as humanitarian benefits. Sickness prevents people from attending their place of work, and also incurs treatment and rehabilitation costs. An epidemic which causes a spike in the number of sick individuals in a population can place a huge strain on the healthcare infrastructure of a community, potentially forcing authorities to implement costly emergency measures such as a mass quarantine. A mathematical model can also give an indication of the extent to which vaccination of a population, quarantining or other intervention is required to prevent an epidemic from becoming established. Even if a conservative estimate of the required vaccination coverage is used compared to the estimate given by a mathematical model, this could still greatly reduce the expenditure needed to combat a given disease.

Presenting a generalised mathematical model for the spread of infectious diseases has clear problems. The intricacies surrounding transmission are numerous, with a person's age, gender, occupation, living arrangements and social activities being among the variety of factors that could feasibly impact upon the probability of them becoming infected by a disease and the number of people that they would then pass that disease on to. It is impractical, both in terms of collecting the necessary data and mathematically, to implement a model based on all possible variables affecting the spread of disease. However, by considering a simpler model, one may be able to estimate the severity of an epidemic and the effectiveness of intervention strategies to the extent that one can determine the most efficient strategy for preventing the outbreak from affecting a significant proportion of the population. By using additional knowledge from outside of the mathematical model when executing this strategy, it should be possible to prevent major outbreaks from occurring if the necessary resources are available. An example of this is vaccinating people whom one may consider more vulnerable to infection or more likely to spread the disease themselves based on key determinants of health (e.g. children or people who work with the general public). We can also look to build upon simpler models to include more variables as further mathematical techniques become available and it is this aspiration which motivates much of the work presented in this thesis.

For the work presented here we use the stochastic SIR (susceptible → infective → recovered) model for a closed population of households. We define a closed population to be a population in which there is no migration. The purpose of a closed households model is to mimic the population structure of urban settlements which a large proportion of the world's human population lives in. The assumption of a closed population is reasonable since the rate of migration in and out of urban populations is generally far smaller than the rate at which epidemics spread within them. Splitting a population into small groups may also be useful in modelling the spread of disease between animals or plants on farms, since animals my be kept in small groups at night (e.g. separate barns or sties) and plants may be grouped according to their plot. Under the SIR model, all individuals start off as being susceptible to a given disease which is introduced to the population. A susceptible individual who makes contact with an infective individual contracts the disease and becomes infected themselves for a certain amount of time after which they recover and are no longer able to contract or transmit the disease. Under the households model considered here, all individuals are considered equally susceptible and make contact with all other individuals in the population with equal frequency. The only exceptions to this are individuals in the same household, with whom contacts are made with additional frequency. That is to say that an infective is just as likely to transmit the disease to one given susceptible in the population as any other, except for those susceptibles within the same household for whom infectious contact becomes more likely.

This model is given in a more detailed, mathematical manner in Chapter 2. However, for now it should be noted that the phrases "infectious contact" and "contact" are used as generally interchangeable to refer to a contact made between an infective and a susceptible which results in the susceptible becoming infected unless specifically stated otherwise. (This does not occur until Chapter 5.) For the remainder of the introduction, we consider the history of mathematical epidemiology, with a particular emphasis on the history of the stochastic SIR households model.

1.2 Early history of mathematical epidemiology

Attempts to model epidemics in the 19^{th} century were largely based around fitting curves to incidences of a disease over time and extending them to predict the future course of an outbreak. While this is a reasonable starting point for mathematical epidemiology, such work is widely considered to be redundant now since the predictions made from curve-fitting proved to be highly inaccurate when compared to observations of outbreaks (see p.10 of Bailey [1975]).

Heesterbeek [2002], however, does cite the work of En'ko [1889, 1989] as being an important development as this is possibly the first work in epidemiology in which a mathematical model implies the existence of a threshold for the infectiousness of a disease beyond which epidemics can occur. Specifically, En'ko notes that conditions for an epidemic to spread are much more favourable in large populations with a strong communication network between individuals. This could be considered to allude to *population density*, an idea which formed the reference points for the earliest explicit threshold parameters in mathematical epidemiology. En'ko is probably also the first to consider a stochastic epidemic model and even considers the idea of immunity after infection, laying the foundations for an SIR model. Unfortunately this work was originally printed in Russian and this may well explain why it appears to have gone largely unnoticed by early 20^{th} century mathematicians, such as Kermack, McKendrick and Bailey, who are mentioned below.

The work of Ross [1911] provides the first application of the threshold concept. His 'mosquito theorem' suggests that a certain density of mosquitoes is needed for a malaria outbreak to occur and that therefore it is not necessary to remove all mosquitoes from a given area to cut short a malaria outbreak. One simply has to reduce their number to below the critical density required for an epidemic to occur. His subsequent papers (Ross [1916], Ross and Hudson [1917a,b]) on '*a priori* pathometry' developed the first epidemic model using prior assumptions regarding the manner of disease transmission in a population. This idea would underpin future models, in the sense that we approach analysis of an epidemic with a model in mind in advance (such as the stochastic SIR households model). Data from an outbreak are then used to estimate the parameters of that model, rather than using the data alone to suggest an adequate model as well as its parameters.

For a long time, McKendrick [1925] was widely credited with introducing the first stochastic epidemic model (see Bailey [1975]), since Ross' model was deterministic and En'ko's work was unknown. Despite this, his most important contribution to mathematical epidemiology was in Kermack and McKendrick [1927], which introduces the SIR model for a deterministic epidemic (the first of five concluding comments notes that under their model "complete immunity is conferred by a single attack") and generalises Ross' 'mosquito theorem' into the celebrated threshold theorem for infectious diseases. The second concluding comment explains this theorem:

> “In general a threshold density of population is found to exist, which depends upon the infectivity, recovery and death rates peculiar to the epidemic. No epidemic can occur if the population density is below this threshold value.”

The further comments emphasise the importance of the threshold theorem, noting that small increases in infectivity could be the cause of a major outbreak and that the termination of an epidemic is related to the time at which enough recoveries have occurred such that the density of susceptibles falls below that required by the threshold theorem to allow for the possibility of a large scale epidemic. As such, Kermack and McKendrick note that “an epidemic, in general, comes to an end before the susceptible population is exhausted”. This threshold density would eventually become the reproduction number, R_0, with an epidemic being able to take place only if $R_0 > 1$. The concept of R_0 was actually introduced before Kermack and McKendrick’s seminal paper by Dublin and Lotka [1925]. However, this parameter was as a ratio of births in a demographic context rather than as a parameter in epidemic modelling. It would be another 50 years before R_0 became synonymous with modelling the spread of infectious disease.

The next great leap would come from Bailey [1953]. In this work, Bailey defines a stochastic SIR epidemic with infection and recovery rates. He then goes on to give a method for estimating their ratio, which he defines as the “relative removal rate”, using statistical techniques (specifically maximum likelihood estimation) and uses this estimated parameter to give a distribution for the final size of an epidemic. The subsequent paper by Whittle [1955] simplifies the calculation for the final size distribution and, perhaps more importantly, uses Bailey’s relative removal rate as a threshold parameter by comparing it to the total population size. Whittle shows that when the relative removal rate exceeds the population size, a large scale epidemic cannot occur, creating a threshold theorem for stochastic SIR epidemics. He also gives the probability of a major outbreak occurring when the relative removal rate falls below the population size. By using a statistical approach to estimate parameters in an epidemic model and formulating a stochastic equivalent of Kermack and McKendrick's threshold theorem, Bailey and Whittle set the standard for analysing epidemics using the stochastic SIR model. Important contributions at this time were also made by Kendall [1956] and Bartlett [1956]. In particular both make an effort to approximate how epidemics reach their final size and thus provide some of the first analyses of epidemics whilst they are in progress.

Although the relative removal rate introduced by Bailey and analysed by Whittle provides a threshold parameter for stochastic SIR epidemics, its form is rather untidy since its value needs to be compared to the population size in order for it to take on any meaning. The basic reproduction number, R_0, is a more satisfying threshold parameter since any communicable disease can only take off if $R_0 > 1$. Consequently, it is R_0 rather than Bailey's relative removal rate which has become the standard threshold parameter in mathematical epidemiology. In an interesting parallel with Ross' mosquito theorem preceding Kermack and McKendrick's threshold theorem, the inspiration for a threshold parameter with a critical value of 1 in mathematical epidemiology would also come from the study of malaria. Macdonald [1955] uses the term 'basic reproduction rate' for his value z_0 and notes that

> "The critical level is 1.0, rates below which determine the progressive elimination of the disease."

Macdonald had actually discussed the basic reproduction rate of malaria in an earlier paper (Macdonald [1952]) but it is the introduction of the parameter z_0 that gives his work a striking resemblance to epidemiology's R_0. In his paper dedicated to the history and calculation of R_0, Heesterbeek [2002] credits Dietz [1975] with introducing the first clearly defined threshold parameter for mathematical epidemic models which has a critical value of 1. This value, R, would quickly become the R_0 which is now so familiar in epidemiology. He also cites Hethcote [1975] and Becker [1975] as making valuable contributions in this development. Credit should also be given here to Bartoszyński [1967] who equated epidemic models to branching processes. Branching process theory provides a standard framework to understand the concept of R_0 by equating births and non-extinction in a standard branching process to infectious contacts in a epidemic and an epidemic becoming established in a population, respectively. Branching processes have also been used to develop and understand more complicated epidemic models, such as the households model.

Whilst Bailey and Whittle were formalising their continuous-time SIR model, an analogous discrete-time model was being introduced to the mathematical community. Reed and Frost developed their model in the 1920s but it was Abbey [1952] who first fully explained the model in published writing. Although there are some clear differences between the Reed-Frost model and the continuous-time, stochastic SIR model (such as the inclusion of a latent period in the Reed-Frost model) they do share many qualities and both can be equated to branching processes in their initial stages. As such, enhancing one's understand of one model can often be beneficial in learning about the other.

1.3 Development of the households model

The techniques for analysing epidemic models discussed above all assume that the population in which a given disease spreads is homogeneously mixing. That is to say that any infective individual in the population has an equal chance of infecting any given susceptible. This is clearly an unrealistic assumption since an individual is far more likely to infect somebody that they live or work with than someone chosen at random from the population. Therefore, we have motivation to look at an epidemic model in which the population is partitioned into small groups, such as households. The first attempt to introduce a model without homogeneous mixing was made by Rushton and Mautner [1955]. They developed a deterministic epidemic model (with no recoveries) in which several communities interact with homogeneous mixing taking place between communities and additional homogeneous mixing taking place within a given community. Watson [1972] introduced a stochastic SIR version of this model and included a notion of epidemic severity based on the number of different communities affected by a disease as well as a threshold theorem which gives a minimum requirement for a generalised epidemic, in which most communities are affected, to be possible. This work considers a population split into homogeneously mixing communities, but does rely on each of those communities containing a large number of individuals (unlike households in a realistic population which are generally small). However, the threshold theorem produced by Watson does use a key idea in that it considers the spread of an epidemic based on the proliferation of infected communities, rather than infected individuals.

Bartoszyński [1972] was the first to look at epidemics among a population split into smaller communities, such as households, with no minimum requirement for their size. The distinction from the work of Watson [1972] being that Bartoszyński [1972] considers a population consisting of a large number of groups of fixed size rather than a population consisting of a fixed number of large groups. This work uses a deterministic, discrete-time model rather than the continuous time, stochastic SIR model which is predominantly used in this thesis. However, his notion that, in the early stages of an epidemic, all infectious contacts between households can be considered to infect individuals in fully susceptible households only, forms the basis for the threshold theorem for all household epidemic models, including the stochastic SIR model. It would be some time before a threshold theorem for households would be explicitly determined. In the interim, Longini and Koopman [1982] fitted a stochastic, discrete-time households model to real data, with a focus on estimating the infectious rates within households and in the population as a whole. These are the values which determine the threshold parameter for households and therefore whether a large scale epidemic is possible. Meanwhile, Ball [1986] and Addy et al. [1991] worked on introducing arbitrary but specified infectious periods into epidemic models, considering the final size distribution of epidemics in a homogeneously mixing population and a multipopulation, similar to that used by Watson. Addy et al. apply this information to a households model and attempt to estimate the parameters of an epidemic from final outcome data.

The breakthrough of a threshold parameter, R_*, for the stochastic SIR households model came with the parallel works of Becker and Dietz [1995] and Ball et al. [1997]. This is analogous to R_0 from the homogeneous mixing case in the sense that both parameters have to take a value greater than 1 in order for there to be any possibility of an epidemic taking off. For the households model this can be seen as an event similar to Watson's generalised epidemic, but in this case referring to a large number of households in the population being affected by the epidemic. The parameter is calculated by considering the number of new households that one infective household is expected to infect rather than considering the proliferation of infected individuals. The rationale behind this is that an individual in a fully susceptible household has far more "targets" through local infectivity than an individual who is the last in their household to be infected. Such differences in the expected number of infectious contacts an individual makes means that calculating a threshold parameter based on infected individuals is a highly cumbersome task, although Pellis et al. [2012] do give a reproduction number using this method. Ball et al. [1997] go on to give a detailed analysis of the model and discuss the final size distribution of an SIR households epidemic in their paper. They also consider estimation of R_* from final outcome data of an epidemic. Ball and Lyne [1999] extend this argument to show how final size data can be used from an SIR households epidemic can be used to estimate the global and local infectious rate parameters by using maximum likelihood estimation. This improves upon the estimators of Addy et al. [1991] who had assumed that within-household epidemics occurred independently of each other.

A key area of research in more recent years has been to consider the behaviour of emerging epidemics. We have outlined the work describing how the parameters of an epidemic can be estimated using data from its final outcome in a population. However, it is often desirable to try to understand the dynamics of the spreading of a disease before these data are available so that the epidemic can be combated before it is established in other populations (such as nearby towns) and to curb the effects in the population where the epidemic is establishing itself currently. It has already been noted that branching processes have been used to approximate the early stages of an epidemic, suggesting that there is a mathematical structure attached to emerging epidemics.

Pellis et al. [2011] note that the *real-time growth rate* is one of the first pieces of information available from an emerging epidemic and go on to show how this rate relates to the other parameters of an epidemic in a households setting. Wallinga and Lipsitch [2007] explain the relationship between the real-time growth rate and reproduction numbers of an epidemic, whilst Fraser [2007] uses the real-time growth rate for estimation purposes, showing how it can be used to give estimates of the threshold parameters R_0 and R_*. Little work has been done however on making use of observed data in the emerging phase of an epidemic in order to estimate its parameters.

Another interesting idea is that of a local infectious rate that is determined by household size. Cauchemez et al. [2004] moot this idea and suggest a potential households model to incorporate it. It seems sensible to suggest that an infective individual is more likely to infect a given susceptible in the same household if they are the only other individual in the household rather than if they are one of many other individuals in the household. In the latter case one would assume that the level of contact between the specific susceptible and infective would be less frequent. As such, this concept is included in the model given in Chapter 2 that is used throughout this thesis.

1.4 Vaccination in epidemic models

Upon finding that the threshold parameter of an epidemic is greater than 1 (be that R_0 in the homogeneous case or R_* for epidemics among a population of households) the priority of any authority dealing with the outbreak should be to introduce preventative measures in order to reduce the threshold parameter to 1 and thereby eliminate the possibility of a large scale outbreak. An obvious example of such a measure which has received considerable attention from mathematicians is vaccination.

The first mention of vaccination among mathematicians looking at epidemic models (excluding Bernoulli's pioneering work on inoculation mentioned in Section 1.1) is credited to Neyman and Scott [1964], who looked to use a Galton-Watson process to show how immunisation could reduce the expected size of a stochastic discrete-time epidemic. Becker [1972] builds on this work with the aim of determining the minimum number of individuals that need to be vaccinated in order to curtail the spread of a disease. In Becker's case he was attempting vaccinate as few people as possible in cases where vaccines were known to have potentially harmful side effects and was therefore looking for a balance between preventing an epidemic from spreading and avoiding having too many people fall ill from taking a vaccine which they did not need. Becker was effectively trying to reduce the threshold parameter to below 1 at the minimum cost possible in terms of vaccines used (although this was not explicitly mentioned in his paper since R_0 was still three years away from being formally introduced to the mathematical community). By reducing the threshold parameter of an epidemic to below 1, one can be certain that a major outbreak cannot occur under a deterministic model. The same is true under a stochastic model as the population size $N \to \infty$. Hethcote and Waltman [1973] show how to use vaccination to reduce the spread of an epidemic below a fixed value, effectively providing a mathematical framework to achieve this goal for a deterministic model.

The survey paper of Wickwire [1977] cites the Taylor [1968] paper on the spread of bovine viral diarrhoea (BVD) among cattle as the first to consider the effect of vaccination in a stochastic epidemic model but also comments that it is Becker [1975] who considers vaccination in terms of the early stages of an epidemic which can be approximated by a branching process. Taylor should also be credited with noticing potential problems with vaccine models, such as the potential for vaccines to fail and the possibility of new strains of a disease to develop which would nullify the effect of a successful vaccination, however he does not elaborate on these ideas mathematically. Wickwire notes from Becker's work that vaccinating a population such that the epidemic has a "birth rate" of 1 prevents a major outbreak and that there is little value in vaccinating a population any further. This is equivalent to reducing the basic reproduction number, R_0, to 1 and provides a stochastic version of Hethcote and Waltman's work in the deterministic setting.

Hethcote [1978] extends this work to vaccinating a heterogeneously mixing population with a disease spreading under a deterministic model. This population contained large groups, such as a town (c.f. the Rushton and Mautner [1955] model), but paved the way for work on vaccinating a population split into households of small size. The subsequent publication by May and Anderson [1984] demonstrates the value of extending the work to a heterogeneously mixing population, since they show that if one falsely assumes a homogeneously mixing population, then under-vaccination will occur under the deterministic model if one allocates vaccines randomly. If, however, one knows the manner in which a heterogeneously mixing population is split then, by allocating vaccines properly, it is possible to prevent an epidemic by vaccinating fewer people than the false homogeneous model suggests is necessary.

The aforementioned papers Becker and Dietz [1995] and Ball et al. [1997] were key in moving the theory of vaccination in mathematical epidemiology to a households model where the locally mixing groups are small in size. Becker and Dietz introduce a post-vaccination reproduction number R_v for stochastic SIR epidemics among a community of households which, as with other reproduction numbers, needs to be reduced to 1 in order to prevent a major outbreak. They also consider different vaccination strategies and discussed whether it is preferable to vaccinate whole households or random individuals in order to contain an outbreak at the minimum possible cost. Ball et al., meanwhile, make a conjecture as to the actual optimal vaccination strategy for this epidemic model which they call *equalisation*. The idea of this strategy is to choose to vaccinate an individual in a household with the maximum available number of unvaccinated susceptibles, effectively removing them from the population and, so far as is possible, equalising household sizes throughout the population in terms of the number of susceptible individuals within them. We discuss this strategy in greater detail in Chapter 5.

All of the work on vaccination mentioned above assumes that vaccines are certain to render the individual that they are given to fully immune from a given disease. Becker and Starczak [1997] allowed for a vaccine to have a random response and responses which could make an individual partially immune to a disease rather than making them fully immune or having no effect whatsoever. Ball and Lyne [2002a,b, 2006] discuss these vaccine models in more detail and show how an imperfect vaccine can affect the optimal vaccination strategy for an epidemic. With an imperfect vaccine, individuals cannot ensure their effective removal from the population post-vaccination and so the equalization strategy outlined above is generalised. Chapter 12 of Andersson and Britton [2000] perhaps offers the best mathematical introduction to the effects of vaccination on epidemics and also includes a section describing a potential method for estimating the effectiveness of an imperfect vaccine.

1.5 Recent literature

We focus on the model of Ball et al. [1997]. Over the course of this thesis we develop the theory outlined in this introduction by introducing hypothesis testing to data observed at the end of an epidemic and a maximum pseudolikelihood method for parameter estimation from emerging epidemic data to this model.

We also include a study on vaccination, with a focus on how using the households model of Cauchemez et al. [2004] affects results seen previously in the literature. A perfect epidemic model would be mathematically tractable whilst also reflecting the complex and random nature of a given outbreak. In the earlier sections of this introduction, we have outlined the approach of mathematicians who have started with a simple, stochastic epidemic model and have increased its complexity over the course of the last 125 years to give the stochastic SIR households model. It is hoped that this thesis serves to develop both the complexity and the potential for inference from this model.

In this section we attempt to offer an overview of the wealth of other research in the epidemiological field that is currently being carried out by those who have taken alternative approaches, be that in terms of model selection or in developing methods for statistical inference from epidemic data. Since the field is so vast, we largely restrict ourselves to discussing literature related to stochastic SIR models (and even then we barely scratch the surface of the available literature). Note, however, that other compartmental models, such as the SI, SIS and SIRS models, and deterministic epidemic models are still widely studied. See Keeling and Rohani [2011] for an overview of these alternative epidemic models.

An obvious starting point is to consider current literature which looks at the SIR households epidemic model, as used in this thesis. Bayesian approaches to inference have been used for some time on this model. Cauchemez et al. [2004] suggest a model for local infectivity that is used throughout this thesis and adopt a Bayesian MCMC (Markov Chain Monte Carlo) approach to parameter estimation for their model while Clancy and O'Neill [2007] use a rejection sampling methodology when considering exact model selection for outbreaks of influenza from real life final outcome data from influenza epidmeics. Parameter estimation for these same influenza data sets under an SIR households epidemic model is also considered by Demiris and O'Neill [2005], who use a Bayesian MCMC approach, and Neal and Kypraios [2015] by using data augmentation. Chapter 3 of this thesis considers parameter estimation and model selection from a frequentist perspective using the same influenza outbreak data. This focus on real household studies forms an important part of the current literature. For example, a wealth of studies based on the influenza A(H1N1) pandemic in 2009 have been taken using various statistical approaches and these have been reviewed by Lau et al. [2012] and House et al. [2012].

Several other approaches have been used in recent times to analyse SIR households epidemic data, particularly using Bayesian techniques. Neal [2012] uses approximate Bayesian computation (ABC) methodology and develops an ABC based algorithm which is shown to be computationally efficient when analysing households data from an epidemic. The sequential Monte Carlo (SMC) and non-parametric methods of Toni et al. [2009] and Knock and Kypraios [2014] respectively for inference from SIR epidemic data can also be applied to the households model used here. Britton and Giardina [2014] provide a more in depth review in inferential methods for infectious disease.

This thesis only considers a frequentist approach to inference from households epidemic data. This generally has the advantage of being computationally less expensive then many of the Bayesian approaches listed above and has the usual frequentist advantage of not needing to use prior distributions. Our methods do have their limitations however and in many circumstances Bayesian methods may prove more fruitful than the frequentist approach. For example, we discuss in Section 4.7 that a Bayesian approach may well provide the best way of approximating the standard error of the key estimator that is derived in Chapter 4. As such, the frequentist approach taken here should only be seen as a preference of the author rather than a dismissal of the important role that Bayesian inference has to play in this field.

Increasing the complexity of the stochastic SIR households model to reflect real epidemic dynamics is also a key current area of research. Neal [2016] extends the households epidmeic model by introducing a notion that an individual is only ever mixing within their household or in their wider community and thus cannot have infectious pressure on both at the same time. There is also the households-workplace model of Ball and Neal [2002] which considers individuals belonging to two separate local groups in which there are increased levels of mixing. This model generally assumes that individuals in the same household do not work in the same workplace. Ouboter et al. [2015] propose and analyse a hierarchical model, in which members of the same household also have the same workplace. This is a particularly useful model for analysing diseases which are particularly prevalent amongst children, since siblings are often likely to attend the same school.

There are also many potential parameters of interest in epidemic models that are not covered in this thesis but have received considerable attention elsewhere. Goldstein et al. [2009] define several reproduction numbers for an epidemic among households and give inequalities related to them. These values include a households reproduction number and a real-time growth rate, which are used in this thesis, but also include an individual reproduction number, which is the true equivalent of the value R_0 discussed earlier in this introduction. In addition, Goldstein *et al.* relate these reproduction numbers to the vaccination coverage needed to prevent the possibility of a major outbreak. Pellis et al. [2012] and Ball et al. [2016] extend this work by introducing further reproduction numbers, refining the calculation technique for R_0 and extending the methodology to a households workplace model. Scalia-Tomba et al. [2010] consider the use of generation times (time needed to pass on the disease to another individual after becoming infective) under the households model. The authors note that creating an unbiased depiction of the dynamics of an epidemic is not simple using generation times but they offer potential solutions to this problem and note that generation time is a concept that is often easy to observe in practice.

Further inferential methods for epidemics have been studied outside of the confines of a households model. For example, the distribution of the final size of an stochastic epidemic is of great interest. Methods for computing this probability mass function in a homogeneously mixing outbreak are reviewed and compared by House et al. [2013]. Others have considered modelling the transmission of specific diseases for which the households epidemic model may not be appropriate. For example, Ainseba and Iannelli [2012] discuss the use of screening methods in curbing the transmission of infections such as HIV while reproduction numbers for the varicella-zoster virus (VZV) in different countries are calculated using various socio-demographic factors by Santermans et al. [2015]. Ideas for increasing model complexity have also stemmed from those starting out with models which do not include local mixing in small groups such as households. Examples include O'Neill and Wen [2012], who suggest using a model in which the rate at which a given susceptible in a population is exposed to infectious contact does not increase linearly with the number of infectives in the population and Arino and Portet [2015], who consider dividing a large population (such as a city) into its central area and a series of smaller satellite communities or suburbs.

Alternative data collection techniques are also being used to gain a deeper insight into the dynamics of SIR epidemics. Tsang et al. [2016] discuss the advantages of collecting data during an epidemic by targeting households that have become infected for collection, rather than taking a random sample of households in the population (as is assumed within this thesis). This method of data collection has limitations, particularly with respect to the accuracy of parameter estimation, but could prove to be a useful tool in gaining a very early insight into the dynamics of a households epidemic.

A modelling method that has received considerable attention in recent times is the network epidemic model. The theory behind this model is to produce a directed graph in which nodes represent members of the population and an edge from person i to person j denotes that person i will make an infectious contact with person j should person i becomes infected at some point during the epidemic. Any nodes which can be reached on this graph, starting from those nodes representing the initial infectives, represent individuals who are ultimately infected by the outbreak (see Newman [2002]). For the remainder of this section we offer brief overview of the reasons behind the emergence of the network model and some of the results obtained from this model in the literature.

The epidemic models discussed earlier in this introduction and throughout this thesis may be incorporated into a network epidemic model. For example, Ball and Neal [2002] show how the stochastic SIR households epidemic model may also be expressed as a network model. These models have the additional advantage of being able to model further complexities within the population structure. For example, it is generally unrealistic to assume that a given individual interacts with an entire population. By using a network model, one can easily restrict the number of other individuals with whom a given individual in a population may make contacts with. In their review paper on network epidemic models, Keeling and Eames [2005] note that such restrictions change the dynamics of an epidemic considerably and that networks have the advantage of being able to model measures, such as contact tracing, that may be used to control an outbreak.

Over the past 10 years, the inferential tools available for network epidemic models have been developed. Ball et al. [2010b] derive a reproduction number, the probability of a global outbreak occurring and the expected final size of a global outbreak for a network model in which the number of potential contacts for a given individual is restricted. Real-time growth rate for an epidemic on an unclustered network is computed by Pellis et al. [2015]. Thus the key parameters used for inference in the more general households epidemic model used in this thesis are increasingly being made available for the more complex network model.

Other tools have also been developed for improving inference from epidemics on networks. A deterministic ordinary differential equation (ODE) model is used to estimate the spread of a disease on such a network by Ma et al. [2012], who also consider the effect of clustering on the final outcome of an epidemic. Hsu et al. [2015] use a network structure to develop a Bayesian hierarchical model for the spreading of influenza which can be used to determine several levels of heterogeneity in an outbreak from final size data. A network model is also used by López-Garcia [2016] to study a stochastic SIR epidemic among a small population in which all individuals display heterogeneity.

Of course, the level of information required to set up an accurate network epidemic model is often unlikely to be available. For example, our only information on a given population may come from census data which are unlikely to offer details beyond total population size, ages, genders and knowledge of how the population is divided into households. Thus, whilst the use of networks with a random degree specification (see Trapman [2007]) can go some way to making up for these deficiencies, the simplicity afforded by the model used in this thesis still has a huge role to play in epidemiology.

1.6 Thesis outline and key results

The outline of this thesis is as follows. Mathematical preliminaries are given in Chapter 2. We introduce the SIR households epidemic model from a mathematical point of view, with a particular emphasis on allowing the rate at which infectives pass on a disease to others in their own household to vary according to household size, and give a formula for the calculation of the threshold parameter R_* for this model. This chapter also contains some discussion regarding when an epidemic can be considered to have “taken off” and offers some results adapted from the literature as to the final outcome of such an epidemic.

Ball and Lyne [2001] provide a central limit theorem for the final outcome of an SIR households epidemic. In Chapter 3 we use this theorem to develop hypothesis tests. The theory behind these hypothesis tests is presented in a general setting however, as the chapter progresses, the focus turns towards tests relating to model selection for the local dynamics of a households epidemic. We show that if the hypotheses being tested only relate to local infectious parameters of our model, we need not know the proportion of the entire population within the sample we have available. The tests are then illustrated using real influenza data for which we consider three possible nested models for the local dynamics of a households epidemic.

Chapter 4 considers the problem of parameter estimation whilst an epidemic is still in progress. We introduce the notion of an emerging epidemic, as defined in the literature, and show that intuitive estimators of local contact rates using emerging epidemic data using pseudolikelihood methods turn out to be biased. A new, asymptotically unbiased estimator for this model is developed in this chapter and is adapted to define a similar estimator for the discrete-time Reed-Frost epidemic. A series of illustrations using simulated data are then used to ascertain that the new estimator can perform well in practice. This chapter is based on the papers of Ball and Shaw [2015, 2016].

We consider the effects of vaccination on an epidemic in Chapter 5. Models for vaccine action, vaccination strategies and a post vaccination threshold parameter that have been considered previously in the literature are introduced at the beginning of this chapter. We then discuss the notion of an optimal vaccination strategy to prevent an epidemic from taking off in more detail and outline how using a model in which local contact rates can depend on household size can cause this strategy to change. This is followed by a series of illustrations which are used to examine the impact of error in parameter estimation (be that as a result of the variance of the estimators or an incorrect model selection) in terms of whether a population may be under/over-vaccinated. If a population has been under-vaccinated, we examine the affect this has on the expected final outcome of an epidemic. Finally, some concluding comments and suggestions for future research are given in Chapter 6.

CHAPTER 2

Mathematical preliminaries

In the introduction, the history of the households SIR (susceptible → infective → recovered) epidemic model was presented, along with an overview of the most recent literature in similar areas of mathematical epidemiology, in order to provide a context and motivation for the work of this thesis. This chapter presents a general mathematical overview of the SIR households epidemic model, as well as some results from the literature regarding the threshold parameter and the final outcome of an epidemic that are relevant to the thesis as a whole.

2.1 Model

The epidemic model used throughout this thesis is based on that analysed by Ball et al. [1997]. We consider a closed, finite population of individuals, each of whom resides within exactly one household. Given that the population is finite, there is a maximum household size which is denoted by n_{max}. For $n = 1, 2, ..., n_{max}$, let m_n be the number of households of size n, let $m = \sum_{n=1}^{n_{max}} m_n$ be the total number of households and let $N = \sum_{n=1}^{n_{max}} nm_n$ be the total number individuals in the population. Further, let $\alpha_n = m_n/m$ be the proportion of households of size n and $\tilde{\alpha}_n = nm_n/N$ be the proportion of individuals residing in households of size n in the population. The row vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_{n_{max}})$ shall be referred to as the *population structure*. The vector $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \tilde{\alpha}_2, ..., \tilde{\alpha}_{n_{max}})$ may be used equivalently to define the population structure since there is a one-to-one correspondence between $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$.

It is assumed that a small number of individuals in the population become infected by some external force at time $t = 0$ and that all other individuals in the population are susceptible at this time. The amount of time that a given individual spends as an infective is determined by a random variable, T_I, which is independently and identically distributed for each individual in the population. The distribution of T_I is arbitrary but, unless stated otherwise, is assumed to be known. During its infectious period, an infected individual makes *global contacts* with a given susceptible in the population at the points of a homogeneous Poisson process with rate λ_G/N. For $n = 2, 3, ..., n_{max}$, an infective in a household of size n makes additional *local contacts* with a given susceptible in the same household at points of a homogeneous Poisson process with rate $\lambda_L^{(n)}$. (Observe that infectives in households of size 1 cannot make local contacts and thus it is unnecessary to include $\lambda_L^{(1)}$ as a parameter of the model. Alternatively $\lambda_L^{(1)}$ may assume an arbitrary value.) It is assumed that all of the Poisson processes describing infectious contacts and all infectious periods are mutually independent. Note that this is an extension to the model of Ball et al. [1997], who assume a single local contact parameter, independent of household size.

Once a susceptible has been contacted by an infective, be it globally or locally, the susceptible immediately becomes infected themselves. Once an infective ends its infectious period it recovers and no longer makes infectious contacts, nor is it affected by infectious contact. The epidemic ends as soon as no infectives remain in the population. It is worth noting that the "R" in the SIR epidemic model is traditionally referred to as removed rather than recovered in the literature, dating back to Kermack and McKendrick [1927]. The change to recovered is only made to clarify that the population size N does not change and thus nor does rate at which global contacts are made between two individuals. Recovery simply implies that the given individual is no longer affected by infectious contact.

A more significant change to much of the previous literature is the use of household size dependent local contact rates. This was suggested by Cauchemez et al. [2004], whose particular model is explored in later chapters. Note that a single local contact rate independent of household size is a special case of our current model which may be achieved by setting $\lambda_L^{(n)} = \lambda_L$ $(n = 2, 3, ..., n_{max})$ for some, $\lambda_L \geq 0$. Also note that if we take $\lambda_L = 0$, we recover the traditional, homogeneously mixing model of Bailey [1953], in which any partitioning of the population is assumed to have no effect.

For the sake of convenience, we generally assume that there is no latent period between a susceptible being contacted by an infective and the onset of their own infectious period. Results relating to the final outcome of an epidemic are invariant to general assumptions concerning a latent period, as justified in Section 3 of Ball et al. [1997]. Thus our assumption of no latent period is reasonable. A similar argument shows that one may assume, without loss of generality, that $\mathbb{E}[T_I] = 1$ when dealing with final size data (with other parameters being rescaled accordingly). However, details relating to latent periods and $\mathbb{E}[T_I] = 1$ cannot be ignored when making inferences about epidemics that are still in progress. This issue is addressed in Chapter 4 when we discuss emerging epidemics.

2.2 Threshold parameter

At any given time during the course of an epidemic, we define a *fully susceptible household* to be a household containing susceptible individuals only and an *infected household* to be a household which contains, or has contained, at least one infected individual. We also define a *single-household epidemic* to be the progress of the epidemic due to local contacts alone in an infected household with one or more initial infectives which have been contacted externally, be that from global infectious contact within the population or as one of the initial infectives in the epidemic. The size of a single-household epidemic is defined to be the eventual number of individuals that become infected (or recovered).

Let the initially infected households belong to the 0^{th} generation of the epidemic and let any household that becomes infected as a result of global contact from an i^{th} generation household belong to the $(i+1)^{th}$ generation $(i = 0, 1, 2, ...)$. Note that generations can and often will overlap in real time. For a community in which the total number of households, m, is large, it is highly probable that all global contacts made by individuals in infected households in the early stages of an epidemic are with individuals in fully susceptible households. (This probability tends to 1 as $m \to \infty$ provided "early stages" has been defined appropriately. See Section 4.6 for further details.) These contacts also occur independently of each other as described in Section 2.1. The proliferation of infected households on a generation-by-generation basis in the initial stages of an epidemic can therefore be approximated by a branching process with the following offspring distribution, the mean of which is the threshold parameter, R_*. The idea of using a generational approach to epidemic analysis is justified by Ludwig [1975], who observes that the dependence of the dynamics of an epidemic on the time elapsed since its inception can be ignored if attention is restricted to the final outcome of the epidemic.

Consider an individual that has been contacted globally by an infective in the initial stages of an epidemic. From the statement above we can assume that this individual is in a fully susceptible household and therefore initiates a single-household epidemic with one initial infective in that household. For $n = 1, 2, ...,$ n_{max}, $\tilde{\alpha}_n$ is the probability that the given individual resides in a household of size n and, for $a = 1, 2, ..., n$, let $\mu_{n,a}(\lambda_L^{(n)})$ denote the mean size of a typical single-household epidemic in a household of size n with a initial infectives. Each infective has an expected infectious period of $\mathbb{E}[T_I]$ and, given the way the model is defined in Section 2.1, infectives make global contacts with individuals chosen uniformly at random from the population at a rate of λ_G in the initial stages of an epidemic, when almost all of the population is still susceptible. Thus, by considering the number of global contacts made by infectives in a single-household epidemic of size n with 1 initial infective (and hence the number of newly infected households in the epidemic), we obtain the threshold parameter for the proliferation of infected households of Ball et al. [1997], given by

$$R_* = \lambda_G \mathbb{E}[T_I] \sum_{n=1}^{n_{max}} \tilde{\alpha}_n \mu_{n,1}(\lambda_L^{(n)}). \tag{2.2.1}$$

By standard branching process theory (for example, Athreya and Ney [1972] p.7) the approximating branching process described above becomes extinct with probability 1 if $R_* \leq 1$ and with probability strictly less than 1 if $R_* > 1$. However, epidemics under our model always become 'extinct' (in the sense that they terminate) since, unlike the approximating branching process, they take place in a finite population. It is therefore relevant to ask whether the threshold parameter R_* serves any useful purpose in this context.

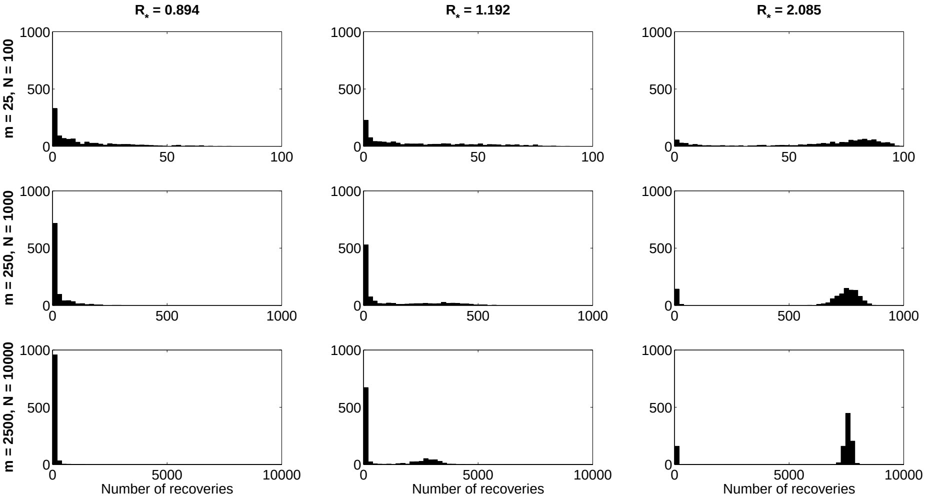

Figure 2.1: Histograms depicting the number of people infected in each of nine sets of 1000 simulations of epidemics. Three different threshold parameters (increasing from left to right) and three different population sizes (increasing from top to bottom) were considered. In each epidemic the population was partitioned into households of size 4, T_I took a negative exponential distribution with mean of 1 and $\lambda_L^{(4)} = 1$. Vertical axis denotes frequency

Figure 2.1 offers an illustrative example of the importance of the threshold parameter R_* and, in particular, the critical value $R_* = 1$. We consider populations of 100, 1000 and 10000 individuals partitioned into 25, 250 and 2500 households respectively, each of size 4, and epidemics for which T_I takes a negative exponential distribution with mean 1 and $\lambda_L^{(4)} = 1$. The histograms in Figure 2.1 show the final outcomes of three sets of 1000 epidemic simulations of each population size. In the left hand plots $\lambda_G = 0.3$, in the central plots $\lambda_G = 0.4$ and in the right hand plot $\lambda_G = 0.7$. Since $\mu_4(\lambda_L^{(4)}) = 2.979$, calculation of which is explained in Section 2.3, (2.2.1) implies that $R_* = 0.894$ for the first set of epidemic simulations, $R_* = 1.192$ for the second set and $R_* = 2.085$ for the third set. Note from all of the plots that the epidemic can die out early. This may be equated to extinction in the approximating branching process.

The difference between the histograms lies in the fact that when $R_* \leq 1$ (in the left hand column of Figure 2.1) all of the epidemics appear to die out in the initial stages whereas, when $R_* > 1$, a greater proportion of individuals can become infected. This is particularly clear in the bottom right plot of Figure

2.1 which contains the largest population and a value of R_* somewhat greater than the critical value of 1. Here the histogram is clearly split into two parts, between epidemics that have died out quickly and those which go on to become infect a sizeable proportion of the population. It is the set of epidemics in the second part, which ultimately appear to infect around $6000 - 8000$ individuals under these parameters, which can be equated to the approximating branching process not becoming extinct, since their eventual termination is caused only by the population being finite.

We refer to these as epidemics which have *taken off* or as *global epidemics* and it is these epidemics which are studied throughout this thesis. Since epidemics can only take off under this model if $R_* > 1$, we only consider epidemics meeting this criterion. The eventual aim is to reduce the value of the threshold parameter to $R_* \leq 1$, by measures such as vaccination, to ensure that a given epidemic cannot take off. However, Figure 2.1 does present some issues in recognising such epidemics. When R_* is close to one, the expected number of individuals infected by a global epidemic decreases and the variance appears to increase. Thus we observe some overlap in the central column of Figure 2.1 between epidemics which have theoretically taken off and those which have not. Similarly, as the population size decreases, the variance of the number of individuals ultimately infected appears to increase proportionally to N, causing further difficulty in determining whether an epidemic has become global in the theoretical sense. Thus, the distinction between major and minor outbreaks is not always clear, with small population size and closeness to criticality being the key factors in reducing this clarity.

2.3 Final outcome

If an epidemic does take off in the manner described in Section 2.2, the right hand histograms of Figure 2.1 suggest that its final outcome, whilst not being completely determined given the stochastic nature of the epidemic, is predictable to some extent. For $n = 1, 2, ..., n_{max}$ and $a = 1, 2, ..., n$, consider a single-household epidemic in a household of size n initiated by a infected individuals within the household. For $j = a, a+1, ..., n$, let $P_{n,a}(j|\lambda_L^{(n)})$ denote the probability that j individuals (including the initial infectives) are ultimately infected by the single-household epidemic. By Equation (2.5) of Ball [1986], $P_{n,a}(j|\lambda_L^{(n)})$ may be determined by the following triangular system of linear equations

$$\sum_{j=a}^{k} \binom{n-j}{k-j} \frac{P_{n,a}(j|\lambda_L^{(n)})}{\phi((n-k)\lambda_L^{(n)})^j} = \binom{n-a}{k-a}, \quad k = a, a+1, ..., n, \tag{2.3.1}$$

where $\phi(t) = \mathbb{E}[e^{-tT_I}]$ is the moment generating function of the infectious period T_I. The probabilities calculated from (2.3.1) may be used to evaluate threshold parameter R_* using (2.2.1), since

$$\mu_{n,a}(\lambda_L^{(n)}) = \sum_{j=1}^{n} jP_{n,a}(j|\lambda_L^{(n)}). \tag{2.3.2}$$

If $R_* > 1$ and an epidemic takes off, (2.3.1) and (2.3.2) also form the basis for predicting its final outcome. The ensuing argument describing the final outcome of an epidemic is approximate but does give an exact result as $m \to \infty$. The exact result is usually proved by an embedding argument, such as that given in Section 4 of Ball et al. [1997]. Although Ball et al. consider the simpler model in which local contact rates are independent of household size, their methods are easily adapted to our model since households of different sizes are considered on a term by term basis (see (2.3.4) below).

Following Section 3.4 of Ball et al. [1997], let π be the probability that a given individual within the population avoids global infectious contact throughout the course of an epidemic and let z be the proportion of individuals in the population that are ultimately infected by the outbreak (so that Nz is the expected number of individuals infected by the epidemic). Since global infectious contacts occur at points of a Poisson process, the probability that a given individual avoids infection from a single given infective is $\exp(-\mathbb{E}[T_I]\lambda_G/N)$. Recall that Poisson processes governing infectious contact occur independently of each other, thus

$$\pi = [\exp(-\lambda_G \mathbb{E}[T_I]/N)]^{Nz} = \exp(-\lambda_G z \mathbb{E}[T_I]). \tag{2.3.3}$$

The expected proportion of individuals in the population as a whole that become infected can be calculated as follows. For $n = 1, 2, ..., n_{max}$ the number of individuals in a given household of size n that are contacted globally by an infective throughout the course of the epidemic is given by a $Binomial(n, 1-\pi)$

distribution. By conditioning on the number of initial infectives (globally contacted individuals) in a household, considering the expected size of the ensuing single-household epidemic and weighting based on the proportion of individuals in each household size, it follows that

$$z = \sum_{n=1}^{n_{max}} \frac{\tilde{\alpha}_n}{n} \sum_{a=1}^{n} \binom{n}{a} (1-\pi)^a \pi^{n-a} \mu_{n,a}(\lambda_L^{(n)}). \tag{2.3.4}$$

The calculation for z given in (2.3.4) is that given in Ball and Lyne [2002a]. Although the methodology here differs from the calculation of z given in (3.24) of Ball et al. [1997], it should be noted that the two are equivalent except for the use of household size dependent local contact parameters under our model. By substituting the value of π given in (2.3.3) into (2.3.4), we obtain an implicit equation for z. Clearly $z = 0$ ($\pi = 1$) is always a solution to this equation. Ball et al. [1997] show that in the case where $R_* \leq 1$ it is the only solution but that there is a second solution, with $z \in (0, 1)$, when $R_* > 1$. It is this value which gives expected proportion of individuals in the population that become infected when the epidemic takes off.

To illustrate this, consider again the epidemics simulated to create Figure 2.1 in Section 2.2. For the first set of simulations with $R_* = 0.894$, $z = 0$ is the only root of the implicit equation for z given by (2.3.3) into (2.3.4). For the third set of epidemics, in which $R_* = 2.085$, $z = 0.7577$ emerges as a second root of the equation and thus as the expected proportion of the population infected by a major outbreak. A cursory glance at the third column of histograms of Figure 2.1 suggests that this is a reasonable assertion, particularly when looking at $N = 10000$ plot. This is strengthened by the knowledge that the mean size of epidemics with more than 10% of the population infected from these sets of simulations (i.e. those epidemics which took off) was 71.87, 754.1 and 7575 for the $N = 100$, 1000 and 10000 epidemics respectively. For the $R_* = 1.192$ epidemics, the second root of (2.3.4) is $z = 0.2731$ and the mean size of epidemics infecting more than 10% of the population in the $N = 100$, 1000 and 10000 epidemic simulations was 38.72, 310.2 and 2714 respectively. Again, the number of infected individuals seems to correspond to expectation in the larger population, however it is clear that in the smaller populations, we have excluded some epidemics which had theoretically taken off. This illustrates the difficulties discussed at the end of Section 2.2 in determining whether an epidemic has taken off if the population is small and/or R_* is close to unity.

For the remainder of this thesis z and π shall refer to the non-trivial solutions of (2.3.3) and (2.3.4) since these are the values which are relevant to global epidemics. From (3.25) of Ball et al. [1997], it is clear that z is strictly increasing in λ_G. Consequently, (2.3.3) shows that π is strictly decreasing in z and λ_G. Thus, any one of the parameters π, z and λ_G determines the other two if the local contact rates $\lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})}$, population structure $\boldsymbol{\alpha}$ and distribution of the infectious period T_I are fixed. Since (2.2.1) also shows a clear one-to-one correspondence between R_* and λ_G, we observe that any one of π, z and R_* may be used to replace λ_G as the parameter explaining the global infectious dynamics of an epidemic, without loss of information.

Let $\boldsymbol{\theta} = (\pi, \lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n)})$ be a row vector denoting the individual to individual contact rates of an epidemic among a population of households. For $n = 1, 2, ...n_{max}$ and $j = 0, 1, 2, ..., n_{max}$ let $P_n(j, \boldsymbol{\theta})$ be the probability that j individuals are ultimately infected in a given household of size n at the end of a global epidemic under parameters given by $\boldsymbol{\theta}$. Noting that $P_{n,0}(k, \lambda_L^{(n)}) = 1$ if $k = 0$ and zero for any other value of k and using a similar logic to the derivation of (2.3.4) yields

$$P_n(j|\boldsymbol{\theta}) = \sum_{a=0}^{j} \binom{n}{a} (1-\pi)^a \pi^{n-a} P_{n,a}(j-a|\lambda_L^{(n)}). \tag{2.3.5}$$

Addy et al. [1991] show that these probabilities may also be determined by their own triangular system of linear equations, given by

$$\sum_{j=0}^{k} \binom{n-j}{k-j} \frac{P_n(j|\pi, \lambda_L^{(n)})}{\phi((n-k)\lambda_L^{(n)})^j \pi^{n-k}} = \binom{n}{k}, \quad j = 0, 1, ..., n. \tag{2.3.6}$$

These probabilities form the basis of all parameter estimation from final size data in this thesis and (2.3.6) offers a computationally less intensive method for their calculation than using a combination of (2.3.1) and (2.3.5). It should be noted however that both methods become unstable for large n and so a sensible cutoff for n_{max} should be imposed.

CHAPTER 3

Hypothesis testing using final size data for an SIR households epidemic

In this chapter we use hypothesis testing to determine the most appropriate households epidemic model given some final size data. Throughout this chapter, it is assumed that the distribution of the infectious period T_I and household distribution $\boldsymbol{\alpha}$ are known but that parameters determining the rates of global and local infectious contacts are unknown. Following the discussion in Section 2.3, we consider π rather than λ_G as our unknown global contact parameter. Let $\boldsymbol{\theta} = (\pi, \lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n)})$ be a row vector of length n_{max} denoting the parameters of the epidemic model to be estimated.

The final outcomes of an epidemic in different households within the same population are not independent, although the dependence is weak if the number of households, m, is large. (Ball and Lyne [2016] note that the dependence is of the order $1/m$.) Therefore, standard maximum likelihood estimation (MLE) procedures for parameter estimation and the subsequent central limit theorem that would be used for independent observations are potentially inadequate. Section 3.1 provides a central limit theorem for the households epidemic model which takes account of this dependence between households. This theorem was originally presented in Ball and Lyne [2001] but has been adapted to our setting in which the local contact rate is dependent on household size. We also include the notion of there being unobserved households in the population, as suggested by Ball and Lyne [2016].

Parameter estimation and applications of the theory discussed in Section 3.1

are then explored in Section 3.2. This is then used to provide the theory for hypothesis testing on final size data which is explained in Section 3.3. These sections follow the work of Ball and Lyne [2016] although this work has been extended in Section 3.3 to establish specific hypotheses for the developed tests which relate directly to the setting of this thesis.

The remainder of this chapter focuses on application of the preceding theory. Section 3.4 provides a detailed explanation of the calculation of the covariance matrices which are needed to carry out the hypothesis tests established in Section 3.3. Results relating to the proportion of households that have been observed in an epidemic are given in Section 3.5. Specifically, we establish that knowing the proportion of the population which has been observed is unnecessary when applying the specific hypothesis tests established in Section 3.3. The hypothesis tests are then applied to real life data in Section 3.6 and concluding comments on the chapter are made in Section 3.7.

3.1 Key convergence theorem - Ball and Lyne [2001]

The purpose of this section is to provide a central limit theorem for the final outcome of a global epidemic under the model established in Chapter 2. The original version of this theorem appeared in Ball and Lyne [2001] and was extended by Ball and Lyne [2016] to include unobserved households. On a practical level, this may either refer to households for which data could not be obtained, or data only being obtained from a sample of households in a larger population. We extend the theorem to account for the possibility of a local contact rate that is dependent on household size. As in Ball and Lyne [2001], the embedding methods of Scalia-Tomba [1985, 1990] and Ball et al. [1997] to construct epidemics and provide an asymptotic final size distribution before a central limit theorem for the final outcome is derived.

3.1.1 Construction and outcome of a completed epidemic

We follow the Selke-type construction of Ball and Lyne [2001] to construct realisations of an epidemic. Britton [2010] notes that the advantage of this method is that it shows that the final outcome of an outbreak is a random process which can be constructed using a contributions from a random number of independent and identically distributed processes and as such it does obey a central limit theorem. At the end of this subsection we derive an equation using the Selke construction which is used as the basis for the law of large numbers proved in Section 3.1.2 and, eventually, the central limit theorem which is proven in Section 3.1.3.

In order to give a realisation of a typical epidemic, E, we first consider the spread of the epidemic in a household of size n. Label individuals in the household $1, 2, ..., n$ and, for $i = 1, 2, ..., n$, let Q_i^L and Q_i^G be random variables that are negative exponentially distributed with mean 1 and let Q_i^I be a random variable distributed according to T_I which determines the length of the infectious period if individual i becomes infected. All of the random variables Q_i^L, Q_i^G and Q_i^I are independent of each other $(i = 1, 2, ..., n)$.

Suppose each individual in the household is exposed to $t \in [0, \infty]$ units of global infection (this is effectively a rescaling of time). For $i = 1, 2, ..., n$, individual i is infected globally if $Q_i^G < t$. If any individual in the household is infected globally then a local epidemic follows. If y is the number of infectives in the household at a given time, susceptible individuals at that time accumulate local exposure to infection at rate $y\lambda_L^{(n)}$ and a susceptible individual, i say, becomes infected locally when its local exposure reaches Q_i^L. This provides a realisation of the single-household epidemic $E_n(\lambda_L^{(n)}, e^{-t})$, where e^{-t} is the probability that a given individual avoids global infection, and the epidemic terminates when infectives no longer remain in the household (i.e. $\pi = e^{-t}$ for a completed epidemic).

For a single-household epidemic $E_n(\lambda_L^{(n)}, e^{-t})$, let $Y^{(n)}(t)$ be the number of individuals ultimately infected, $A_n(t)$ be the sum of their infectious periods (henceforth known as the *severity* of the epidemic) and $\boldsymbol{R}_n(t) = f_n(Y_n(t))$ be some finite, deterministic, vector-valued function of $Y_n(t)$. For example, $\boldsymbol{R}_n(t)$ could take the form of an indicator function denoting whether the household has been infected, a score statistic vector for unknown parameters of the epidemic or, to take the most trivial case, $Y_n(t)$ itself. For $n = 1, 2, ..., n_{max}$, $k = 1, 2, ...$ let $\{(\boldsymbol{R}_{n.k}(t), A_{n,k}(t))\}$ be independent, identically distributed copies of $\{(\boldsymbol{R}_n(t), A_n(t)) : t \geq 0\}$ which can be realised by generating the random variables $(Q_i^L,\ Q_i^G, Q_i^I)$ $(i = 1, 2, ...n)$ using the construction method above.

A realisation of the final outcome E can then be constructed by first supposing that the population is exposed to T_0 units of global infectious time, meaning that a typical susceptible in the population is exposed to $T_0\lambda_G/N$ initial units of global exposure. For $n = 1, 2, ..., n_{max}$, label the households of size n $1, 2, ..., m_n$ and let

$$\{(\boldsymbol{R}_\bullet(t), A_\bullet(t))\} = \sum_{n=1}^{n_{max}} \sum_{k=1}^{m_n} \{(\delta_{n,k}\boldsymbol{R}_{n,k}(t), A_{n,k}(t))\}, \tag{3.1.1}$$

where $\delta_{n,k}$ is the indicator function of the event that the final outcome in household k of size n has been observed. By considering the single-household epidemics triggered by the initial T_0 units global infectious time, it is clear that a further $A_\bullet(T_0\lambda_G/N)$ units of further global infectious time are introduced to the epidemic which may trigger further local infection. This process will continue and so, for $l = 1, 2, ...$, it is useful to define T_l as total units of global infectious time within the epidemic after l phases of this construction have been completed. Specifically,

$$T_{l+1} = T_0 + A_\bullet(T_l\lambda_G/N)$$

and $l^* = \min\{l : T_{l+1} = T_l\}$ is well-defined since the population is finite ($l^* \leq N$). Let $T_\infty = T_{l^*}$ so that

$$T_\infty = T_0 + A_\bullet(\lambda_G T_\infty/N). \tag{3.1.2}$$

Note that T_∞ represents the final severity of E and $\boldsymbol{R}_\bullet(\lambda_G T_\infty/N)$ the desired vector-valued function of the final outcome of E.

3.1.2 Asymptotics of severity

We now consider the asymptotic distribution of the severity of an epidemic $A_\bullet(t)$. This is achieved by considering a sequence of epidemics indexed by $\nu = 1, 2, ...$ and will be used in Section 3.1.3 to find the asymptotic distribution of $\boldsymbol{R}_\bullet(t)$, which in turn will be used to give a central limit theorem for our function of interest $\boldsymbol{R}_\bullet(T_\infty)$.

Consider a sequence of epidemics $E^{(\nu)}$, all governed by the same unknown infectious parameters $\boldsymbol{\theta}$ and known infectious period parameter T_I but with different population structures, initiated by $T_0^{(\nu)}$ units of global infectious time ($\nu = 1, 2, ...$). The set of independent processes $\{(\boldsymbol{R}_{n,k}(t), A_{n,k}(t))\}$ can be used to construct realisations of the processes $\{(\boldsymbol{R}_{\bullet}^{(\nu)}(t), A_{\bullet}^{(\nu)}(t))\}$ which in turn provide a realisation of $E^{(\nu)}$. Note that n_{max} is retained as the maximum household size across all epidemics in the sequence.

For $\nu = 1, 2, ...$ let $m^{(\nu)}$ and $N^{(\nu)}$ be the total number of households and individuals respectively in the population of $E^{(\nu)}$. Suppose $m^{(\nu)} \to \infty$ as $\nu \to \infty$. For $n = 1, 2, ..., n_{max}$, let $\alpha_n^{(\nu)} = m_n^{(\nu)}/m^{(\nu)}$, $\alpha_n = \lim_{\nu\to\infty} \alpha_n^{(\nu)}$, $a_n(t) = \mathbb{E}[A_n(t)]$, $a^{(\nu)}(t) = \sum_{n=1}^{n_{max}} \alpha_n^{(\nu)} a_n(t)$ and $a(t) = \sum_{n=1}^{n_{max}} \alpha_n a_n(t)$ $(t \geq 0)$.

Lemma 3.1.1. *Suppose that,*

(i) $a^{(\nu)}(t) \to a(t)$ *as* $\nu \to \infty$ *and*

(ii) $\mathbb{E}[T_I^2] < \infty$.

Then

$$\sup_{t\in[0,\infty]} |(m^{(\nu)})^{-1} A_{\bullet}^{(\nu)}(t) - a(t)| \xrightarrow{a.s.} 0 \text{ as } \nu \to \infty. \tag{3.1.3}$$

Proof. We follow the proof of Lemma 1 of Ball and Lyne [2001] which is easily adapted to this setting. For $n = 1, 2, ..., n_{max}$ and fixed $t \geq 0$,

$$(m^{(\nu)})^{-1} A_{\bullet}^{(\nu)}(t) = (m^{(\nu)})^{-1} \sum_{n=1}^{n_{max}} \sum_{k=1}^{m_n^{(\nu)}} A^{(n,k)}(t) \xrightarrow{a.s.} a(t) \text{ as } \nu \to \infty$$

by the strong law of large numbers and condition (i). (Note that the strong law of large numbers can be applied due to condition (ii).) Hence,

$$(m^{(\nu)})^{-1} A_{\bullet}^{(\nu)}(t) \xrightarrow{a.s.} a(t) \text{ as } \nu \to \infty \tag{3.1.4}$$

for $t \in (\mathbb{Q} \cap [0, \infty]) \cup \{\infty\}$. Fix $\epsilon > 0$. Now, $a(0) = 0$, $a(\infty) < \infty$ and $a(t)$ is non-decreasing in t and hence there exists $r \in \mathbb{N}$ and $t_1, t_2, ..., t_r \in \mathbb{Q}$ such that $0 = t_0 < t_1 < t_2 < ... < t_r < t_{r+1} = \infty$ and

$$a(t_{i+1}) - a(t_i) < \epsilon/2 \quad i = 0, 1, ..., r.$$

From (3.1.4) there also exists $K \in \mathbb{N}$ such that

$$|(m^{(\nu)})^{-1} A_{\bullet}^{(\nu)}(t_i) - a(t_i)| < \epsilon/2 \quad i = 0, 1, ..., r+1, \ \nu > K$$

and since $A_{\bullet}^{(\nu)}(t)$ is also non-decreasing in t it follows that

$$|(m^{(\nu)})^{-1} A_{\bullet}^{(\nu)}(t) - a(t)| < \epsilon \quad \nu > K.$$

Since $\epsilon > 0$ is an arbitrary choice,

$$|(m^{(\nu)})^{-1}A_\bullet^{(\nu)}(t) - a(t)| \xrightarrow{a.s.} 0 \text{ as } \nu \to \infty$$

and the lemma follows. □

For $\nu = 1, 2, ...$, let $m_H^{(\nu)} = N^{(\nu)}/m^{(\nu)}$ and $m_H = \lim_{\nu\to\infty} m_H^{(\nu)} = \sum_{n=1}^{n_{max}} n\alpha_n$, observing that $m_H^{(\nu)}, m_H \leq n_{max}$ (and hence are finite). Also, considering (3.1.2) and letting $T_0^{(\nu)}$ and $T_\infty^{(\nu)}$ be defined in the obvious manner for $E^{(\nu)}$ ($\nu = 1, 2, ...$), note that

$$m_H^{(\nu)} T_\infty^{(\nu)}/N^{(\nu)} = m_H^{(\nu)} T_0^{(\nu)}/N^{(\nu)} + A_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})/m^{(\nu)}. \tag{3.1.5}$$

Suppose that conditions (i) and (ii) above are satisfied and $T_0^{(\nu)}/N^{(\nu)} \to 0$ as $\nu \to \infty$. Then, letting $\nu \to \infty$ in (3.1.5) and using Lemma 3.1.1 implies that $T_\infty^{(\nu)}/N^{(\nu)}$ converges to a random variable satisfying

$$\begin{aligned} m_H t &= a(\lambda_G t) \\ &= \sum_{n=1}^{n_{max}} \alpha_n \mu_n(\lambda_L^{(n)}, e^{-\lambda_G t})\mathbb{E}[T_I], \end{aligned}$$

where $\mu_n(\lambda_L^{(n)}, e^{-\lambda_G t})$ is the expected size of a single household epidemic in which each individual has probability $e^{-\lambda_G t}$ chance of avoiding global infection. By noting that $m \sum_{n=1}^{n_{max}} \alpha_n \mu_n(\lambda_L^{(n)}, e^{-\lambda_G t})$ gives the expected number of individuals that become infected by an epidemic in which individuals have a $e^{-\lambda_G t}$ chance of avoiding global infection, it is easy to see the equivalence of (3.1.5) to (2.3.4). Thus (3.1.5) has a root at $t = 0$ and also at $t = \tau$ if, where $\tau > 0$ if and only if the threshold parameter $R_* > 1$. Specifically, $\tau = z\mathbb{E}[T_I]$ and thus $\pi = e^{-\lambda_G \tau}$.

It should be noted that condition (ii) of Lemma 3.1.1, as well as the eventual list of conditions required for Theorem 3.1.3, have been simplified compared to Ball and Lyne [2001]. This is largely as a result of our imposing a maximum household size, n_{max}, on our epidemic, following the general setting of this thesis. Specifically, for Lemma 3.1.1 to hold if no maximum household size is imposed on the epidemic, condition (ii) should changed to state that there exists $\kappa > 2$ such that $\mathbb{E}[T_I^\kappa] < \infty$. The only restrictions on the population under this setting are that it is closed and that the population size, N, is finite. Also, the conditions for Lemma 3.1.1 below and for Theorem 3.1.2 in Section

3.1.3 only require minor alterations to allow our eventual central limit theorem to be adapted to the multitype setting of Ball and Lyne [2001]. Appendix B of Ball and Lyne [2016] should be consulted for further details on both scenarios outlined here.

3.1.3 Central limit theorem

This subsection provides a functional central limit theorem for $(\boldsymbol{R}_{\bullet}(t), A_{\bullet}(t))$ in Theorem 3.1.2 which is used in Theorem 3.1.3 to derive a central limit theorem for $(\boldsymbol{R}_{\bullet}(T_{\infty}), A_{\bullet}(T_{\infty}))$. In Section 3.2, we consider an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ obtained by maximum pseudolikelihood estimation. By having a central limit theorem for $\boldsymbol{R}_{\bullet}(T_{\infty})$, we have a central limit theorem for the score statistic and Fisher information of a pseudolikelihood function of $\boldsymbol{\theta}$ following an observation of an epidemic that has reached its conclusion. Using these we are eventually able to derive a central limit theorem for $\hat{\boldsymbol{\theta}}$ in Section 3.2.

We begin by introducing some new notation which will eventually be used to define the covariance matrix for our central limit theorem for $\boldsymbol{R}_{\bullet}(T_{\infty})$. For $n = 1, 2, ..., n_{max}$, let $\beta_n = m^{-1} \sum_{k=1}^{m_n} \delta_{n,k}$ be the proportion of households in the population that are both observed and of size n and let $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_{n_{max}})$. Let $\boldsymbol{R}_n(t) = (R_{n1}(t), R_{n2}(t), ..., R_{np}(t))$ where p is the dimension of $\boldsymbol{R}_n(t)$ and let $\boldsymbol{r}(t) = \sum_{n=1}^{n_{max}} \beta_n \boldsymbol{r}_n(t)$ where $\boldsymbol{r}_n(t) = (r_{n1}(t), r_{n2}(t), ..., r_{np}(t)) = \mathbb{E}[\boldsymbol{R}_{n,1}(t)]$. Let $\boldsymbol{C}_{RR}(t) = \sum_{n=1}^{n_{max}} \beta_n \boldsymbol{C}_{RR}^n(t)$ where $\boldsymbol{C}_{RR}^n(t) = \text{var}(\boldsymbol{R}_n(t))$ is a $p \times p$ matrix whose i, j^{th} element, $c_{ij}^n(t,t)$, is given by $\text{cov}(R_{ni,1}(t), R_{nj,1}(t))$. Let $C_{AA}(t) = \sum_{n=1}^{n_{max}} \alpha_n C_{AA}^n(t)$ where $C_{AA}^n(t) = \text{var}(A_{n,1}(t))$ and $\boldsymbol{C}_{RA}(t) = \sum_{n=1}^{n_{max}} \beta_n \boldsymbol{C}_{RA}^n(t)$ where the column vector $\boldsymbol{C}_{RA}^n(t) = \text{cov}(\boldsymbol{R}_{n,1}(t), A_{n,1}(t))$. Finally, let $\boldsymbol{B} = D_{\boldsymbol{R}}(m_H - D_A)^{-1}$ where $D_{\boldsymbol{R}}$ is a column vector whose i^{th} element is given by $\partial r_i(\lambda_G \tau)/\partial \tau$ and $D_A = \partial a(\lambda_G \tau)/\partial \tau$. For $\nu = 1, 2, ...$, let $\beta_n^{(\nu)}$ be defined as expected and let $\beta_n = \lim_{\nu \to \infty} \beta_n^{(\nu)}$.

Theorem 3.1.2. *For $c > 0$ and $t \in [0, c]$*

$$(m^{(\nu)})^{-1/2}(\boldsymbol{R}_{\bullet}^{(\nu)}(t) - \mathbb{E}[\boldsymbol{R}_{\bullet\nu}(t)]) \xrightarrow{W} \boldsymbol{X}(t) \text{ as } \nu \to \infty,$$

where $\boldsymbol{X}(t) = (X_1(t), X_2(t), ..., X_p(t))$ is a zero-mean Gaussian process with covariance function given by $\text{cov}(X_i(t), X_j(s)) = \sum_{n=1}^{n_{max}} \beta_n c_{ij}^n(t,s)$ $(t, s \in [0, c];\ i, j =$

$1,2,...,p$) and $\xrightarrow{W}$ denotes weak convergence in the product space of bounded functions on $[0,c]$.

Provided the conditions of Theorem 5.2 of Ball and Lyne [2001] are satisfied, the argument of the proof of that theorem may be used to prove Theorem 3.1.2. The conditions are as follows.

(i′) $\lim_{\nu\to\infty}\sum_{n=1}^{n_{max}}\alpha_n^{(\nu)}r_{ni}(t)=\sum_{n=1}^{n_{max}}\alpha_n r_{ni}(t)<\infty \quad (t\in[0,c];\ i=1,2,..,p);$

(ii′) $\lim_{\nu\to\infty}\sum_{n=1}^{n_{max}}\alpha_n^{(\nu)}c_{ij}^{n}(s,t)=\sum_{n=1}^{n_{max}}\alpha_n c_{ij}^{n}(s,t)<\infty \quad (s,t\in[0,c];\ i,j=1,2,..,p);$

(iii′) for some $\kappa>2$,

$$\lim_{\nu\to\infty}\frac{1}{(M^{(\nu)})^{\kappa/2-1}}\sum_{n=1}^{n_{max}}\alpha_n^{(\nu)}\mathbb{E}[(\boldsymbol{R}_{n,1}(t))^{\kappa}]=0 \quad (t\in[0,c];\ i=1,2,..,p);$$

(iv′) if

$$F_i^{(\nu)}(t)=\sum_{n=1}^{n_{max}}\alpha_n^{(\nu)}\mathbb{E}[(\boldsymbol{R}_{n,1}(t))\boldsymbol{R}_{n,1}(c))] \quad (t\in[0,c];\ i=1,2,..,p)$$

and

$$D^{(\nu)}=\max_{i=1,2,...,p}\max_{t\in[0,c]}\frac{\mathrm{d}}{\mathrm{d}t}F_i^{(\nu)}(t),$$

then there exist $A,B>0$ such that $A\le D^{(\nu)}\le B$ for all sufficiently large ν.

Conditions $(i')-(iii')$ are satisfied since the sums are over the finite index set $\{1,2,...,n_{max}\}$. A similar theorem holds for $A_{\bullet}^{(\nu)}(t)$ provided that condition (ii) of Lemma 3.1.1 is satisfied since (iii') follows immediately from (ii). Note also that (iii') follows immediately from the adapted version of (ii) given at the end of Section 3.1.2 that is to be used when no maximum household size is imposed on the epidemic.

It is assumed in Ball and Lyne [2001] that, for $i=1,2,...,p$, $R_{ni}(0)=0$ and that $R_{ni}(t)$ is non-decreasing in t for $t\ge 0$. This assumption and condition (iv') are addressed at the end of this subsection. We now give a central limit theorem for $\{(\boldsymbol{R}_{\bullet}^{(\nu)}(t),A_{\bullet}^{(\nu)}(t))\}$ in a similar manner to Theorem 5.3 of Ball and Lyne [2001]. In the statement of this theorem $\mathbf{0}$ refers to a row vector of zeros of size p but in general $\mathbf{0}$ will be used to refer to a vector or matrix of zeros of an appropriate size throughout this chapter.

Theorem 3.1.3. *Suppose that the conditions required for Lemma 3.1.1 are satisfied and as,* $\nu \to \infty$,

(a) $(m^{(\nu)})^{1/2} T_0^{(\nu)} / N^{(\nu)} \xrightarrow{p} 0$;

(b) for $n = 1, 2, ..., n_{max}$, $(m^{(\nu)})^{1/2}(\alpha_n^{(\nu)} - \alpha_n) \to 0$ *and* $(m^{(\nu)})^{1/2}(\beta_n^{(\nu)} - \beta_n) \to 0$,

where $\xrightarrow{p}$ *denotes convergence in probability. Then, in the event of a global epidemic,*

$$(m^{(\nu)})^{-1/2}[\boldsymbol{R}_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)} \boldsymbol{r}(\lambda_G \tau)] \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Sigma}) \text{ as } \nu \to \infty,$$

where $\boldsymbol{\Sigma} = \boldsymbol{C}_{RR}(\tau) + \boldsymbol{B}\boldsymbol{C}_{AR}(\tau) + \boldsymbol{C}_{RA}(\tau)\boldsymbol{B}^\top + \boldsymbol{B}C_{AA}(\tau)\boldsymbol{B}^\top$ *and* $\boldsymbol{C}_{AR}(\tau) = (\boldsymbol{C}_{RA}(\tau))^\top$ *(and thus is a row vector).*

Proof. Since we have specified that we are dealing with a global outbreak, we have $T_\infty^{(\nu)}/N^{(\nu)} \xrightarrow{p} \tau$ (see Section 3.1.2) and hence $\lambda_G T_\infty^{(\nu)}/N^{(\nu)} \xrightarrow{p} \lambda_G \tau$ as $\nu \to \infty$ by the continuous mapping theorem. Thus, by the continuous mapping theorem (see Theorem 1.3.6 of van der Vaart and Wellner [1996]),

$$(m^{(\nu)})^{-1/2}(\boldsymbol{R}_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)} \boldsymbol{r}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})) \xrightarrow{W} \boldsymbol{X}(\lambda_G \tau) \quad (3.1.6)$$

as $\nu \to \infty$, where $\boldsymbol{r}^{(\nu)}(t) = \sum_{n=1}^{n_{max}} \beta_n^{(\nu)} \boldsymbol{r}_n(t)$.

By van der Vaart and Wellner [1996], Addendum 1.5.8, $\boldsymbol{X}$ is separable since almost all of its sample paths are continuous, hence

$$\begin{aligned}
&(m^{(\nu)})^{-1/2}[\boldsymbol{R}_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)} \boldsymbol{r}(\lambda_G \tau)] \\
&\qquad = (m^{(\nu)})^{-1/2}[\boldsymbol{R}_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)} \boldsymbol{r}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})] \\
&\qquad + (m^{(\nu)})^{1/2}[\boldsymbol{r}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - \boldsymbol{r}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})] \qquad (3.1.7a) \\
&\qquad + (m^{(\nu)})^{1/2}[\boldsymbol{r}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - \boldsymbol{r}(\lambda_G \tau)]. \qquad (3.1.7b)
\end{aligned}$$

First note that (3.1.7a)$\xrightarrow{p}$ $\boldsymbol{0}$ as $\nu \to \infty$ since

$$\begin{aligned}
&(m^{(\nu)})^{1/2}[\boldsymbol{r}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - \boldsymbol{r}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})] \\
&= (m^{(\nu)})^{1/2} \left[\sum_{n=1}^{n_{max}} \left\{ \beta_n^{(\nu)} \boldsymbol{r}_n(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - \beta_n \boldsymbol{r}_n(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) \right\} \right] \\
&= (m^{(\nu)})^{1/2} \left[\sum_{n=1}^{n_{max}} (\beta_n^{(\nu)} - \beta_n) \boldsymbol{r}_n(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) \right] \\
&\to \ \boldsymbol{0} \text{ as } \nu \to \infty
\end{aligned}$$

by condition (b).

We now turn our attention to (3.1.7b). For $i = 1, 2, ..., p$ Let $f_i(t) = r_i(t)$ and note that f_i is continuous since r_i is continuous. Let D_{R_i} be the i^{th} component of $D_{\boldsymbol{R}}$. Then, by the mean value theorem,

$$\begin{aligned}
&(m^{(\nu)})^{1/2}[r_i(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - r_i(\lambda_G \tau)] = (m^{(\nu)})^{1/2}(T_\infty^{(\nu)}/N^{(\nu)} - \tau) f_i'(k^{(\nu)}) \\
&= (m^{(\nu)})^{1/2}(T_\infty^{(\nu)}/N^{(\nu)} - \tau) f_i'(\tau) + (m^{(\nu)})^{1/2}(T_\infty^{(\nu)}/N^{(\nu)} - \tau)[f_i'(k^{(\nu)}) - f_i'(\tau)] \\
&= (m^{(\nu)})^{1/2}(T_\infty^{(\nu)}/N^{(\nu)} - \tau) D_{R_i}^\top + K_i^{(\nu)},
\end{aligned}$$

for some $k^{(\nu)} \in (T_\infty^{(\nu)}/N^{(\nu)}, \tau)$. Clearly $k^{(\nu)} \xrightarrow{p} \tau$ as $\nu \to \infty$, therefore, $[f_i'(k^{(\nu)}) - f_i'(\tau)] \xrightarrow{p} \mathbf{0}$ as $\nu \to \infty$ by the continuous mapping theorem. Hence it is sufficient to show that $(m^{(\nu)})^{1/2}(T_\infty^{(\nu)}/N^{(\nu)} - \tau)$ is bounded to show that $K_i^{(\nu)} \to \mathbf{0}$ as $\nu \to \infty$. Recalling (3.1.5) and that $m_H \tau = a(\lambda_G \tau)$ from Section 3.1.2,

$$\begin{aligned}
(m^{(\nu)})^{1/2}(T_\infty^{(\nu)}/N^{(\nu)} - \tau) = {} & (m^{(\nu)})^{1/2} T_0^{(\nu)}/N^{(\nu)} \\
& + (m^{(\nu)})^{-1/2}[A_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)} a(\lambda_G \tau)] m_H^{-1} \\
& + (m^{(\nu)})^{-1/2} A_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})((m_H^{(\nu)})^{-1} - m_H^{-1}).
\end{aligned} \tag{3.1.8}$$

Note that the first term of (3.1.8) converges in probability to 0 as $\nu \to \infty$ by condition (a). For the final term, observe that

$$\begin{aligned}
\frac{(m^{(\nu)})^{-1/2} A_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})}{((m_H^{(\nu)})^{-1} - m_H^{-1})} &= \frac{A_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})}{(m^{(\nu)})^{1/2}(N^{(\nu)})^{1/2}} \frac{(m^{(\nu)})^{1/2}(m_H - m_H^{(\nu)})}{m_H} \\
&= \frac{A_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})}{m^{(\nu)}} \frac{(m^{(\nu)})^{1/2}(m_H - m_H^{(\nu)})}{(m_H^{(\nu)})^{1/2} m_H}.
\end{aligned}$$

This converges in probability to 0 as $\nu \to \infty$ by (b), since $m^{-1} A_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})$ is bounded as $\nu \to \infty$ (see Section 3.1.2) and $(m_H^{(\nu)})^{-1/2} m_H^{-1} \to m_H^{-3/2}$ as $\nu \to \infty$ which is also bounded since $m_H \geq 1$. Hence, $K^{(\nu)} \to \mathbf{0}$ as $\nu \to \infty$.

Thus, using (3.1.8) again,

$$\begin{aligned}
&(m^{(\nu)})^{-1/2}[\boldsymbol{R}_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)} \boldsymbol{r}(\lambda_G \tau)] \\
&\qquad = (m^{(\nu)})^{-1/2}[\boldsymbol{R}_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)} \boldsymbol{r}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})] \\
&\qquad + (m^{(\nu)})^{-1/2}[A_\bullet^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)} a(\lambda_G \tau)] m_H^{-1} D_{\boldsymbol{R}}^\top + F^{(\nu)}
\end{aligned}$$

where $F^{(\nu)} \xrightarrow{p} \mathbf{0}$ as $\nu \to \infty$. The above argument also holds for determining the value of the term $(m^{(\nu)})^{-1/2}[A_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)}a(\lambda_G \tau)]$, so that

$$\begin{aligned}
&(m^{(\nu)})^{-1/2}[A_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)}a(\lambda_G \tau)] \\
&\quad = (m^{(\nu)})^{-1/2}[A_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)}a^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})] \\
&\quad + (m^{(\nu)})^{-1/2}[A_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)}a(\lambda_G \tau)]m_H^{-1}D_A + G^{(\nu)} \\
&\quad = (m^{(\nu)})^{-1/2}[A_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)}a^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})](1 - m_H^{-1}D_A)^{-1} \\
&\quad + G^{(\nu)},
\end{aligned}$$

where $G^{(\nu)} \xrightarrow{p} \mathbf{0}$ as $\nu \to \infty$. Hence

$$\begin{aligned}
&(m^{(\nu)})^{-1/2}[\boldsymbol{R}_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)}\boldsymbol{r}(\lambda_G \tau)] \\
&\qquad = (m^{(\nu)})^{-1/2}[\boldsymbol{R}_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)}\boldsymbol{r}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})] \qquad (3.1.9a) \\
&\qquad + (m^{(\nu)})^{-1/2}[A_{\bullet}^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)}) - m^{(\nu)}a^{(\nu)}(\lambda_G T_\infty^{(\nu)}/N^{(\nu)})]\boldsymbol{B}^\top + H^{(\nu)} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.1.9b)
\end{aligned}$$

where $H^{(\nu)} \xrightarrow{p} \mathbf{0}$ as $\nu \to \infty$. The theorem follows by application of Theorem 3.1.2 on (3.1.9a) and (3.1.9b). It is shown in Section 3.4 that τ is a proper crossing point of $m_H t = a(\lambda_G t)$, hence $D_A \neq m_H$ so $\boldsymbol{B}$ is well defined. □

We now return to showing that condition (iv') holds and that the assumption, for $i = 1, 2, ..., p$, $R_{ni}(t)$ is non-decreasing in t for $t \geq 0$, may be relaxed. Let $\mathbb{1}_{\{A\}}$ denote the indicator function on event A and, for $n = 1, 2, ...n_{max}$; $i = 0, 1, ..., n$, $k = 1, 2, ..., m_n$ and $t > 0$, let $\chi_{n,k,i}(t) = \mathbb{1}_{\{Y_{n,k}(t) = i\}}$ where $Y_{n,k}(t)$ denotes the number of individuals ultimately infected in household k of size n. Hence,

$$\boldsymbol{R}_{n,k}(t) = \sum_{i=0}^{n} f_n(i)\chi_{n,k,i}(t)\delta_{n,k},$$

and so, since n_{max} is finite, a central limit theorem for $\boldsymbol{R}_{\bullet}(t)$ follows immediately from one for $\boldsymbol{R}_{\bullet}^*(t) = \{R_{n,i}^*(t) : n = 1, 2, ..., n_{max}, k = 0, 1, ..., n\}$ where

$$R_{n,i}^* = \sum_{k=1}^{m_n} \chi_{n,k,i}(t)\delta_{n,k}.$$

For $n = 1, 2, ...n_{max}$; $i = 0, 1, ..., n$, $k = 1, 2, ..., m_n$ and $t > 0$, let $\tilde{\chi}_{n,k,i}(t) = \mathbb{1}_{\{}Y_{n,k}(t) \geq i\}$ and note that $\tilde{\chi}_{n,k,i}(t)$ can be expressed as a linear combination of the $(\chi_{n,k,l}(t);\ i \leq l \leq n)$ using the "Möbius inversion" method of Martin-Löf [1986]. Let $\tilde{\boldsymbol{R}}^*_\bullet(t) = \{\tilde{R}^*_{n,i}(t) : n = 1, 2, ..., n_{max},\ k = 0, 1, ..., n\}$, where

$$\tilde{R}^*_{n,i} = \sum_{k=1}^{m_n} \tilde{\chi}_{n,k,i}(t)\delta_{n,k}.$$

A central limit theorem for $\boldsymbol{R}^*_\bullet(t)$ (and hence for $\boldsymbol{R}_\bullet(t)$) follows immediately from one for $\tilde{\boldsymbol{R}}^*_\bullet(t)$ and note that $\tilde{\boldsymbol{R}}^*_\bullet(t)$ is non-decreasing in t since $\tilde{\chi}_{n,k,i}(t)$ is non-decreasing in t. Thus a central limit theorem for $\tilde{\boldsymbol{R}}^*_\bullet(t)$ exists if condition (iv') is satisfied.

Following the construction for E described at the beginning of this section, note that, for $t \in [0, c]$ $(c > o)$

$$\mathbb{E}[\tilde{\chi}_{n,1,i}(t)\tilde{\chi}_{n,1,i}(c)] = \mathbb{E}[\tilde{\chi}_{n,1,i}(t)]$$

and that $\mathbb{E}[\tilde{\chi}_{n,1,i}(t)]$ is a polynomial in $\pi = e^{-t}$. Thus, there exist $B_{n,i}, C_{n,i} > 0$ such that

$$B_{n,i} < \max_{t\in[0,c]} \frac{\partial}{\partial t}\mathbb{E}[\tilde{\chi}_{n,1,i}(t)] < C_{n,i}.$$

In addition, $\mathbb{E}[A^{(n,1)}(t)A^{(n,1)}(c)]$ may also be expressed as a polynomial in $\pi = e^{-t}$ and therefore, by (ii) of Lemma 3.1.1, there exist $D_n, E_n > 0$ such that

$$D_n < \max_{t\in[0,c]} \frac{\partial}{\partial t}\mathbb{E}[A_{n,1}(t)A_{n,1}(c)] < E_n.$$

Condition (iv') follows for both $\tilde{\boldsymbol{R}}^*_\bullet(t)$ and $A_\bullet(t)$ since n_{max} is finite.

3.2 Parameter Estimation

We discuss the application of the central limit theorem established in Section 3.1 to parameter estimation from final size data and consider some of the properties of these estimators. This section follows the work of Ball and Lyne [2016] although considerable detail has been added here to justify the central limit theorem that we derive for $\hat{\boldsymbol{\theta}}$.

3.2.1 Application of Theorem 3.1.3

We now discuss estimation of the epidemic model parameters given by $\boldsymbol{\theta}$, as defined at the beginning of this chapter, and make use of the central limit theorem given in Section 3.1.3. First, recall that $\boldsymbol{\theta}$ is a vector of length n_{max} and let $P_n(j|\boldsymbol{\theta})$ be the probability that j individuals are ultimately infected in a given household of size n in an epidemic with infectious contact parameters given by $\boldsymbol{\theta}$ (c.f. Section 2.3). For an observed epidemic among a population structured and labelled in the same manner as E in Section 3.1, let $y_{n,k}$ $(n = 1, 2, ..., n_{max},\ k = 0, 1, ..., m_n)$ denote the number of susceptibles ultimately infected in household k of size n. Let $\boldsymbol{y}_D = \{y_{n,k} :\ n = 1, 2, ..., n_{max},\ k = 0, 1, ..., m_n,\ \delta_{n,k} = 1\}$. An unrestricted maximum pseudolikelihood estimator (MpLE) $\hat{\boldsymbol{\theta}}$ from observed data $\boldsymbol{y}_D$ is obtained by maximising the log-pseudolikelihood function

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{n=1}^{n_{max}} \sum_{k=1}^{m_n} \delta_{n,k} \log P_n(y_{n,k}|\boldsymbol{\theta}). \tag{3.2.1}$$

Let $\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y})$ be the row vector of the score statistic of the pseudolikelihood function with respect to $\boldsymbol{\theta}$, with i^{th} component $(\partial/\partial\theta_i)l(\boldsymbol{\theta}|\boldsymbol{y}_D)$ $(i, = 1, 2, ..., n_{max})$ and assume in the usual manner that $\hat{\boldsymbol{\theta}}$ is given by the solution to $\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) = \mathbf{0}$. Also, let $\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y})$ be the Fisher information matrix with respect to $\boldsymbol{\theta}$ with components $I_{ij}(\boldsymbol{\theta}|\boldsymbol{y}) = -(\partial^2/\partial\theta_i\partial\theta_j)l(\boldsymbol{\theta}|\boldsymbol{y})$ $(i, j = 1, 2, ..., n_{max})$. It follows using the usual Taylor series method that

$$m^{-1/2}\boldsymbol{U}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \approx m^{-1/2}\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) + m^{-1/2}\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

and hence,

$$0 \approx m^{-1/2}\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) - m^{-1}\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y})m^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

Thus,

$$m^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx \left\{m^{-1/2}\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y})\right\}\left\{m^{-1}\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y})\right\}^{-1}. \tag{3.2.2}$$

For the sequence of epidemics $E^{(\nu)}$ discussed in Section 3.1, assume that the infectious parameters, $\boldsymbol{\theta}$ and T_I, and the population structure, observation and initial infectivity parameters, $m^{(\nu)}$, $\boldsymbol{\alpha}^{(\nu)}$, $\boldsymbol{\beta}^{(\nu)}$ and $T_0^{(\nu)}$ $(\nu = 1, 2, ...)$, are such that the conditions required for Theorem 3.1.3 are satisfied. Let $\boldsymbol{U}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}_D^{(\nu)})$ and $\boldsymbol{I}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}_D^{(\nu)})$ be defined in the obvious manner and note that both are vector-valued functions of the final size data ($\boldsymbol{I}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}_D^{(\nu)})$ can be easily made into a vector of length n_{max}^2), summed across all households in the epidemic and therefore are both suitable choices for the function $\boldsymbol{R}_\bullet(t)$, as defined in Section 3.1. Note that there is a one-to-one correspondence between values of θ_1 and t so $\boldsymbol{R}_\bullet$ is a function of θ_1 with $t = \lambda_G\tau$ corresponding to $\theta_1 = \pi$.

First, let $R^{(\nu)}_{ni,k} = \delta_{n,k}\partial/\partial\theta_i[\log(P_n(y^{(\nu)}_{n,k}|\boldsymbol{\theta}))]$ where $y^{(\nu)}_{n,k}$ is the final number of infectives in household k of size n in $E^{(\nu)}$ and $R^{(\nu)}_{ni,k}$ is the i^{th} component of $\boldsymbol{R}_{n,k}$ $(n, i = 1, ..., n_{max};\ k = 1, 2, ..., m^{(\nu)}_n,\ \nu = 1, 2, ...)$ so that

$$\boldsymbol{R}_{\bullet} = \boldsymbol{U}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}^{(\nu)}_D) = \sum_{n=1}^{n_{max}}\sum_{k=1}^{m_n}\delta_{n,k}\boldsymbol{R}_{n,k}$$

is of the form given in (3.1.1). Then,

$$\begin{aligned} r_{ni}(\lambda_G\tau) &= \mathbb{E}[R_{ni,1}(\lambda_G\tau)] \\ &= \sum_{l=0}^{n} P_n(l|\boldsymbol{\theta})\frac{\partial}{\partial\theta_i}[\log(P_n(l|\boldsymbol{\theta}))] \\ &= 0 \end{aligned} \tag{3.2.3}$$

as is always the case when finding the expectation of a score statistic. (Recall that λ_G, z and hence τ are determined explicitly by $\boldsymbol{\theta}$ if the distribution of T_I is assumed to be known.) Hence, by application of Theorem 3.1.3,

$$(m^{(\nu)})^{-1/2}\boldsymbol{U}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}^{(\nu)}_D) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \text{ as } \nu \to \infty. \tag{3.2.4}$$

The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is discussed in more detail in Section 3.4.

Now consider $\boldsymbol{R}^{(\nu)}_{\bullet}(\lambda_G\tau) = m^{-1}\boldsymbol{I}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}^{(\nu)}_D)$ with components given by,

$$R^{(\nu)}_{nij,k} = -m^{-1}\frac{\partial^2}{\partial\theta_i\partial\theta_j}[\log(P_n(y^{(\nu)}_{n,k}|\boldsymbol{\theta}))]$$

where the index nij refers to the $(in_{max} + j)^{th})$ component of the vector $\boldsymbol{R}^{(\nu)}_n$ $(n, i, j = 1, ..., n_{max};\ k = 1, 2, ..., m^{(\nu)}_n,\ \nu = 1, 2, ...)$. Thus

$$\begin{aligned} mr_{nij}(\lambda_G\tau) &= m\mathbb{E}[m^{-1}R_{ni,1}(\lambda_G\tau)] \\ &= -\sum_{l=0}^{n} P_n(l|\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}[\log(P_n(l|\boldsymbol{\theta}))] \\ &= \boldsymbol{I}_n(\boldsymbol{\theta}), \text{ say.} \end{aligned}$$

Letting $\boldsymbol{I}_{\boldsymbol{\theta}} = \sum_{n=1}^{n_{max}}\beta_n\boldsymbol{I}_n(\boldsymbol{\theta})$ and applying Theorem 3.1.3 gives

$$m^{-1}\boldsymbol{I}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}^{(\nu)}_D) \xrightarrow{p} \boldsymbol{I}_{\boldsymbol{\theta}} \text{ as } \nu \to \infty. \tag{3.2.5}$$

The exact form of $\boldsymbol{I}_{\boldsymbol{\theta}}$ is discussed in Section 3.4. From Equations (3.2.2), (3.2.4) and (3.2.5) and noting that $\boldsymbol{I}_{\boldsymbol{\theta}}$ is symmetric, it follows that

$$(m^{(\nu)})^{1/2}(\hat{\boldsymbol{\theta}}^{(\nu)} - \boldsymbol{\theta}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}) \text{ as } \nu \to \infty. \tag{3.2.6}$$

Note that the omission of higher order terms in the Taylor expansion is justified using standard results form approximation theory, as in Section 4.2.2 of Serfling [1980].

3.2.2 Properties of covariance matrices

We now investigate properties of the values $\boldsymbol{B}$, $\boldsymbol{C}_{RR}(\tau)$, $\boldsymbol{C}_{RA}(\tau)$, $\boldsymbol{C}_{AR}(\tau)$ and $C_{AA}(\tau)$ from Theorem 3.1.3, which determine $\boldsymbol{\Sigma_\theta}$. For convenience we now denote $(\partial/\partial x)$ by ∂_x. Thus $R_{ni,k}^{(v)} = \partial_{\theta_i}[\log(P_n(y_{n,k}^{(v)}|\boldsymbol{\theta}))]$. Then, for $i, j = 1, 2, ..., n$ and recalling that τ is determined by $\boldsymbol{\theta}$, the $(i,j)^{th}$ component of $\boldsymbol{C}_{RR}^n(\tau)$ is given by

$$\begin{aligned}
c_{ij}^n(\tau) &= \sum_{k=0}^{n} \left\{ P_n(k|\boldsymbol{\theta}) \left[\partial_{\theta_i} \log(P_n(k|\boldsymbol{\theta}))\right] \left[\partial_{\theta_j} \log(P_n(k|\boldsymbol{\theta}))\right] \right\} - r_{ni}(\lambda_G\tau) r_{nj}(\lambda_G\tau) \\
&= \sum_{k=0}^{n} P_n(k|\boldsymbol{\theta}) \frac{\partial_{\theta_i} P_n(k|\boldsymbol{\theta})}{P_n(k|\boldsymbol{\theta})} \frac{\partial_{\theta_j} P_n(k|\boldsymbol{\theta})}{P_n(k|\boldsymbol{\theta})} \\
&= \sum_{k=0}^{n} [\partial_{\theta_i} P_n(k|\boldsymbol{\theta})][\partial_{\theta_j} P_n(k|\boldsymbol{\theta})] P_n(k|\boldsymbol{\theta})^{-1},
\end{aligned}$$

since $\sum_{k=0}^{n} r_{ni}(\lambda_G\tau) = 0$ as described in Section 3.2.1. Now, the $(i,j)^{th}$ component of $\boldsymbol{I}_n(\boldsymbol{\theta})$ is given by

$$\begin{aligned}
I_{nij}(\boldsymbol{\theta}) &= -\sum_{k=0}^{n} P_n(y_{n,k}|\boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} [\log(P_n(y_{n,k}|\boldsymbol{\theta}))] \\
&= -\sum_{k=0}^{n} P_n(y_{n,k}|\boldsymbol{\theta}) \left[\partial_{\theta_j} \frac{\partial_{\theta_i} P_n(k|\boldsymbol{\theta})}{P_n(k|\boldsymbol{\theta})}\right] \\
&= \sum_{k=0}^{n} \left\{ [\partial_{\theta_i} P_n(k|\boldsymbol{\theta})][\partial_{\theta_j} P_n(k|\boldsymbol{\theta})] P_n(k|\boldsymbol{\theta})^{-1} - \frac{\partial^2}{\partial\theta_i \partial\theta_j} P_n(k|\boldsymbol{\theta}) \right\} \\
&= c_{ij}^n(\tau)
\end{aligned}$$

since

$$\sum_{k=0}^{n} \frac{\partial^2}{\partial\theta_i \partial\theta_j} P_n(k|\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \sum_{k=0}^{n} P_n(k|\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} 1 = 0.$$

Thus $\boldsymbol{I_\theta} = \boldsymbol{C}_{RR}(\tau)$ since both are a sum of their matrices for specific values of n, weighted by $\boldsymbol{\beta}$.

The matrices $\boldsymbol{I_\theta}$ and $\boldsymbol{\Sigma_\theta}$ are dependent on both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Suppose a stratified sample of households in the population is taken such that, for $n = 1, 2, ..., n_{max}$

and $\beta \in (0,1]$, $\beta_n = \beta\alpha_n$ and hence the sampled households represent exactly $100\beta\%$ of the population. (We shall refer to this as a $100\beta\%$ stratified sample of the population.) Then, letting $\boldsymbol{C}_{RR}(\tau,\beta)$ be defined in the obvious manner, it follows from the definitions given in Section 3.1.3 that $\boldsymbol{C}_{RR}(\tau,\beta) = \beta\boldsymbol{C}_{RR}(\tau,1)$, $\boldsymbol{C}_{RA}(\tau,\beta) = \beta\boldsymbol{C}_{RA}(\tau,1)$, $\boldsymbol{C}_{AR}(\tau,\beta) = \beta\boldsymbol{C}_{AR}(\tau,1)$, $C_{AA}(\tau,\beta) = C_{AA}(\tau,1)$ (since $C_{AA}(\tau)$ does not depend on $\boldsymbol{\beta}$) and $\boldsymbol{B}_\beta = \beta\boldsymbol{B}_1$.

Clearly $\boldsymbol{I}_{\boldsymbol{\theta},\beta} = \beta\boldsymbol{I}_{\boldsymbol{\theta},1}$ due to the relationship with $\boldsymbol{C}_{RR}(\tau)$ outlined above. Therefore,

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\beta} =& \boldsymbol{C}_{RR}(\tau,\beta) + \boldsymbol{B}_\beta\boldsymbol{C}_{AR}(\tau,\beta) + \boldsymbol{C}_{RA}(\tau,\beta)\boldsymbol{B}_\beta^\top + \boldsymbol{B}_\beta C_{AA}(\tau,\beta)\boldsymbol{B}_\beta^\top \\ =& \beta\boldsymbol{C}_{RR}(\tau,1) + \beta^2\boldsymbol{B}_1\boldsymbol{C}_{AR}(\tau,1) + \beta^2\boldsymbol{C}_{RA}(\tau,1)\boldsymbol{B}_1^\top + \beta^2\boldsymbol{B}_1 C_{AA}(\tau,1)\boldsymbol{B}_1^\top \\ =& (1-\beta)\beta\boldsymbol{C}_{RR}(\tau,1) + \beta^2[\boldsymbol{C}_{RR}(\tau,1) + \boldsymbol{B}_1\boldsymbol{C}_{AR}(\tau,1) + \boldsymbol{C}_{RA}(\tau,1)\boldsymbol{B}_1^\top + \\ & \boldsymbol{B}_1 C_{AA}(\tau,1)\boldsymbol{B}_1^\top] \\ =& (1-\beta)\beta\boldsymbol{I}_{\boldsymbol{\theta},1} + \beta^2\boldsymbol{\Sigma}_{\boldsymbol{\theta},1}\end{aligned}$$

and thus,

$$\begin{aligned}\boldsymbol{I}_{\boldsymbol{\theta},\beta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\beta}\boldsymbol{I}_{\boldsymbol{\theta},\beta}^{-1} =& \beta^{-2}\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}[(1-\beta)\beta\boldsymbol{I}_{\boldsymbol{\theta},1} + \beta^2\boldsymbol{\Sigma}_{\boldsymbol{\theta},1}]\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1} \\ =& \beta^{-1}(1-\beta)\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1} + \boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},1}\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}.\end{aligned}$$

It now follows, from (3.2.6), that

$$(\beta m^{(\nu)})^{-1/2}(\hat{\boldsymbol{\theta}}^{(\nu)} - \boldsymbol{\theta}) \xrightarrow{D} N(\mathbf{0}, \tilde{\Sigma}_{\boldsymbol{\theta},\beta}) \text{ as } \nu \to \infty$$

where

$$\begin{aligned}\tilde{\Sigma}_{\boldsymbol{\theta},\beta} =& \beta^{1/2}[\beta^{-1}(1-\beta)\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1} + \boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},1}\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}]\beta^{1/2} \\ =& (1-\beta)\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1} + \beta\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},1}\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}.\end{aligned}$$

Note that in the limit as $\beta \to 0$, this yields the usual Fisher information matrix obtained by ignoring dependence between outcomes in different households and treating (3.2.1) as a log-likelihood. The relationship between $\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}$ and $\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}$ (and thus the importance of $\boldsymbol{\beta}$ on hypothesis testing on the unknown parameters $\boldsymbol{\theta}$) is investigated in Section 3.5.

3.3 Hypothesis Tests

3.3.1 Hypotheses and preliminaries

We wish to test whether observed final size data could have come from a given epidemic model with infectious parameters $\boldsymbol{\theta}$. Specifically we wish to test our current model, in which local contact rates depend on the household size n, against the more traditional households epidemic model for which there is only one local contact rate parameter which is the same for all household sizes. The test may be written using the nested hypotheses

$$\begin{aligned} &H_0 : \ \lambda_L^{(2)} = \lambda_L^{(3)} = ... = \lambda_L^{(n_{max})} \\ &H_1 : \ \lambda_L^{(i)} \neq \lambda_L^{(j)} \quad \text{for some } i \neq j \quad i, j = 2, 3, ..., n_{max}. \end{aligned} \tag{3.3.1}$$

or, to give the test in terms of $\boldsymbol{\theta}$,

$$\begin{aligned} &H_0 : \ \theta_2 = \theta_3 = ... = \theta_{n_{max}} \\ &H_1 : \ \theta_i \neq \theta_j \quad \text{for some } i \neq j \quad i, j = 2, 3, ..., n_{max}. \end{aligned}$$

Let $\boldsymbol{h}(\boldsymbol{\theta})$ be a vector of length $n_{max} - 2$ such that, for $i = 1, 2, ..., n_{max} - 2$, $h_i(\boldsymbol{\theta}) = \theta_{i+2} - \theta_2$. Then the above test may be re-written as

$$\begin{aligned} &H_0 : \ \boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0} \\ &H_1 : \ h_i(\boldsymbol{\theta}) \neq 0 \quad \text{for some } i = 1, 2, ..., n_{max} - 2. \end{aligned}$$

Let $\dot{\boldsymbol{\theta}}$ denote the restricted maximum pseudolikelihood estimator under H_0 and recall that $\hat{\boldsymbol{\theta}}$ is the unrestricted MpLE under H_1. The asymptotic distribution of $\hat{\boldsymbol{\theta}}$ under H_1 is given by (3.2.6). We now look to determine the asymptotic distribution of $\dot{\boldsymbol{\theta}}$ under H_0.

Let $\boldsymbol{\lambda}$ be a column vector of Lagrange multipliers and $\boldsymbol{H_\theta}$ be the $(n_{max}) \times (n_{max} - 2)$ matrix with elements given by $(H_{\boldsymbol{\theta}})_{ij} = \partial h_j / \partial \theta_i$ $(i = 1, 2, ..., n_{max};\ j =$

$1, 2, ..., n_{max} - 2)$, so that

$$\boldsymbol{H_\theta} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 1 \\ -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Note in particular that the first row of $\boldsymbol{H_\theta}$ is a row of zeros due to θ_1 not appearing in our constraint vector $\boldsymbol{h}(\boldsymbol{\theta})$. Then

$$\begin{aligned} \boldsymbol{U}(\dot{\boldsymbol{\theta}}|\boldsymbol{y}) - \boldsymbol{H}_{\dot{\boldsymbol{\theta}}}\dot{\boldsymbol{\lambda}} &= \boldsymbol{0} \\ \boldsymbol{h}(\dot{\boldsymbol{\theta}}) &= \boldsymbol{0}, \end{aligned}$$

where $\boldsymbol{0}$ denotes the zero column vector of length $n_{max} - 2$ and $\dot{\boldsymbol{\lambda}}$ is the appropriate vector of Lagrangian multipliers such that the equation above holds. Making use of Taylor's theorem yields the approximations

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) + \boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y})(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \boldsymbol{H_\theta}\dot{\boldsymbol{\lambda}} &\approx \boldsymbol{0} \\ \boldsymbol{H}_{\boldsymbol{\theta}}^\top(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}) &\approx \boldsymbol{0}. \end{aligned} \tag{3.3.2}$$

Page 80 of Silvey [1975], explains the presence of $\boldsymbol{H_\theta}\dot{\boldsymbol{\lambda}}$ rather than $\boldsymbol{H}_{\dot{\boldsymbol{\theta}}}$ here by noting that if $\dot{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}$ then it is also close to $\hat{\boldsymbol{\theta}}$ where $\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) = \boldsymbol{0}$ and hence $\dot{\boldsymbol{\lambda}}$ is small. Expanding $\boldsymbol{H}_{\dot{\boldsymbol{\theta}}}$ about $\boldsymbol{\theta}$ gives first order terms containing $\dot{\boldsymbol{\lambda}}$ and $\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}$ meaning that the order of the terms are small enough to ignore. Simple manipulation of (3.3.2) gives

$$\begin{aligned} \left[-m^{-1}\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y})\right] m^{1/2}(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}) + m^{-1/2}\boldsymbol{H_\theta}\dot{\boldsymbol{\lambda}} &\approx m^{-1/2}\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) \\ \boldsymbol{H}_{\boldsymbol{\theta}}^\top m^{1/2}(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}) &\approx \boldsymbol{0}. \end{aligned}$$

Therefore, recalling (3.2.5) and considering the sequence of epidemics $E^{(\nu)}$,

$$\begin{pmatrix} (m^{(\nu)})^{-1/2}\boldsymbol{U}^{(\nu)}(\boldsymbol{\theta}^{(\nu)}|\boldsymbol{y}^{(\nu)}) \\ \boldsymbol{0} \end{pmatrix} \approx \begin{pmatrix} \boldsymbol{I_\theta} & \boldsymbol{H_\theta} \\ \boldsymbol{H}_{\boldsymbol{\theta}}^\top & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} m^{1/2}(\dot{\boldsymbol{\theta}}^{(\nu)} - \boldsymbol{\theta}) \\ (m^{(\nu)})^{1/2}\dot{\boldsymbol{\lambda}}^{(\nu)} \end{pmatrix}.$$

Now let

$$\begin{pmatrix} \boldsymbol{I_\theta} & \boldsymbol{H_\theta} \\ \boldsymbol{H}_{\boldsymbol{\theta}}^\top & \boldsymbol{0} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{Q}^\top & \boldsymbol{R} \end{pmatrix}. \tag{3.3.3}$$

It is clear from their definitions that $\boldsymbol{I_\theta}$ is a symmetric, positive-definite matrix and that $\boldsymbol{H_\theta}$ has rank $n_{max} - 2$. Thus, by Appendix A.8 of Silvey [1975], the left hand side of (3.3.3) is non-singular, its inverse does indeed take the form given on the right hand side of (3.3.3) and, specifically, $\boldsymbol{P}$ is a symmetric matrix given by

$$\boldsymbol{P} = \boldsymbol{I}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}. \tag{3.3.4}$$

Then, by letting $\boldsymbol{Y} = (m^{(\nu)})^{-1/2}\boldsymbol{U}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}^{(\nu)})$,

$$(m^{(\nu)})^{1/2}(\dot{\boldsymbol{\theta}}^{(\nu)} - \boldsymbol{\theta}) \approx \boldsymbol{PY} \tag{3.3.5}$$

and thus, using (3.2.4),

$$(m^{(\nu)})^{1/2}(\dot{\boldsymbol{\theta}}^{(\nu)} - \boldsymbol{\theta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{P\Sigma_\theta P}) \tag{3.3.6}$$

as $\nu \to \infty$ (c.f. Serfling [1980], Section 4.4.4, Lemma C).

We can now consider a pseudolikelihood-ratio test, a pseudo-Wald's test and a pseudoscore test on the hypotheses given above. However, we should note that further hypothesis tests are possible under the same framework and using very minor modifications of the theory in this section. In particular, the tests given in this section can be generalised to any other hypotheses as long as the null hypothesis can be characterised by $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$ for some set of constraints $\boldsymbol{h}$.

For example, we may with to consider the model of Cauchemez et al. [2004] in which the local contact parameter takes the form $\lambda_L^{(n)} = \lambda_L / n^\eta$, where η may take any real value. To consider this as the alternative hypothesis against the local contact parameter being independent of household size is straightforward in that we let $\boldsymbol{\theta} = (\pi, \lambda_L, \eta)$ and $h(\boldsymbol{\theta}) = \theta_3$ (i.e $\eta = 0$ under the null hypothesis). Testing this model against the new model in which there is no relationship between the local contact rate of households of different sizes can be achieved in the following manner. The unknown parameters of the alternative hypothesis can be described, as before, by $\boldsymbol{\theta} = (\pi, \lambda_L^{(2)}, ..., \lambda_L^{(n_{max})})$. For $i = 3, 4, ..., n_{max}$ we require

$$2^\eta \theta_2 = i^\eta \theta_i \; (= \lambda_L) \tag{3.3.7}$$

for some value η. The constraint function $\boldsymbol{h}(\boldsymbol{\theta})$ cannot contain η however since $\boldsymbol{h}(\boldsymbol{\theta})$ needs to be a function of $\boldsymbol{\theta}$. By taking logarithms on each side of (3.3.7), note that for $i = 3, 4, .., n_{max}$

$$\frac{\log\left(\frac{\theta_i}{\theta_2}\right)}{\log\left(\frac{2}{i}\right)} = \eta$$

and thus $\boldsymbol{h}(\boldsymbol{\theta})$ may be given by

$$h_i(\boldsymbol{\theta}) = \frac{\log\left(\frac{\theta_3}{\theta_2}\right)}{\log\left(\frac{2}{3}\right)} - \frac{\log\left(\frac{\theta_{i+3}}{\theta_2}\right)}{\log\left(\frac{2}{i+3}\right)} \quad (i = 1, 2, ..., n_{max} - 3). \tag{3.3.8}$$

Details relating to tests on the local contact parameters only, such as those under the hypotheses outlined in (3.3.1) and those given above are considered in Section 3.5. However, we consider the more general case for the remainder of this section. Note also that if we assume that the observed households represent a $100\beta\%$ stratified sample of all households in the population, for some $\beta \in (0, 1]$, we have already established in Section 3.2.2 that $\boldsymbol{I}_{\boldsymbol{\theta},\beta} = \beta \boldsymbol{I}_{\boldsymbol{\theta},1}$. It is therefore clear from the definition of $\boldsymbol{P}$ given in (3.3.4) that $\boldsymbol{P}_\beta = \beta^{-1}\boldsymbol{P}_1$, where $\boldsymbol{P}_\beta$ is defined in the obvious manner, since $\boldsymbol{H}_{\boldsymbol{\theta}}$ is not dependent on $\boldsymbol{\beta}$. The following tests were originally established in Ball and Lyne [2016].

3.3.2 Pseudolikelihood ratio test

The pseudolikelihood ratio test is based upon the test statistic

$$2\log\lambda = 2\{l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - l(\dot{\boldsymbol{\theta}}|\boldsymbol{y})\}$$

with H_0 being rejected if $2\log\lambda$ is too large. Using a Taylor expansion,

$$l(\dot{\boldsymbol{\theta}}|\boldsymbol{y}) \approx l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + \boldsymbol{U}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})^\top(\dot{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + (\dot{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\dot{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})/2$$

and hence, by considering the sequence of epidemics $E^{(\nu)}$ and since $\boldsymbol{U}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = \boldsymbol{0}$,

$$\begin{aligned} 2\log\lambda^{(\nu)} \approx& -(\dot{\boldsymbol{\theta}}^{(\nu)} - \hat{\boldsymbol{\theta}}^{(\nu)})^\top \boldsymbol{I}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}^{(\nu)})(\dot{\boldsymbol{\theta}}^{(\nu)} - \hat{\boldsymbol{\theta}}^{(\nu)}) \\ =&(m^{(\nu)})^{-1/2}(\dot{\boldsymbol{\theta}}^{(\nu)} - \hat{\boldsymbol{\theta}}^{(\nu)})^\top\{-(m^{(\nu)})^{-1}\boldsymbol{I}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}^{(\nu)})\}(m^{(\nu)})^{-1/2}(\dot{\boldsymbol{\theta}}^{(\nu)} - \hat{\boldsymbol{\theta}}^{(\nu)}). \end{aligned}$$

Now, using equations (3.2.2), (3.3.3) and (3.3.5),

$$\begin{aligned} (m^{(\nu)})^{1/2}(\dot{\boldsymbol{\theta}}^{(\nu)} - \hat{\boldsymbol{\theta}}^{(\nu)}) &= (m^{(\nu)})^{1/2}(\dot{\boldsymbol{\theta}}^{(\nu)} - \boldsymbol{\theta}) - (m^{(\nu)})^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(\nu)}) \\ &\approx (\boldsymbol{P} - \boldsymbol{I}_{\boldsymbol{\theta}}^{-1})\boldsymbol{Y}, \end{aligned}$$

so

$$\begin{aligned}2\log\lambda^{(\nu)} \approx& \boldsymbol{Y}^\top(\boldsymbol{P}-\boldsymbol{I}_{\boldsymbol{\theta}}^{-1})\boldsymbol{I}_{\boldsymbol{\theta}}(\boldsymbol{P}-\boldsymbol{I}_{\boldsymbol{\theta}}^{-1})\boldsymbol{Y}\\ =&\boldsymbol{Y}^\top[\boldsymbol{P}\boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P}-2\boldsymbol{P}+\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}]\boldsymbol{Y}.\end{aligned}$$

However, by equation (3.3.4)

$$\begin{aligned}&\boldsymbol{P}\boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P} =\\ &(\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}-\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1})\boldsymbol{I}_{\boldsymbol{\theta}}(\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}-\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1})\\ &=\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}-\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1})\boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\\ &\quad-\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1})\boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\\ &\quad+\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\\ &=\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}-\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}=\boldsymbol{P},\end{aligned}$$

since $\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}$ gives the identity matrix. Thus,

$$\begin{aligned}2\log\lambda^{(\nu)} \approx& \boldsymbol{Y}^\top[\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}-\boldsymbol{P}]\boldsymbol{Y}\\ =&\boldsymbol{Y}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Y}\\ =&\boldsymbol{Y}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{Y}.\end{aligned}$$

Note that

$$\begin{aligned}&(\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2})(\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2})\\ =&\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\\ =&\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2},\end{aligned}$$

so $\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}$ is an idempotent matric of rank $n_{max}-2$ (the rank of $\boldsymbol{H}_{\boldsymbol{\theta}}$). Since $\boldsymbol{I}_{\boldsymbol{\theta}}$ is symmetric and positive semi-definite, $\boldsymbol{I}_{\boldsymbol{\theta}}^{1/2}$ is symmetric. Thus, $\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}$, $\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}$ and hence $\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}$ are also symmetric since the inverse of a symmetric matrix is symmetric.

Recall from (3.2.4) that $\boldsymbol{Y}\sim N(\boldsymbol{0},\boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ and let $\boldsymbol{X}=\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{Y}\xrightarrow{D}N(\boldsymbol{0},\boldsymbol{I})$ as $\nu\to\infty$ (where $\boldsymbol{I}$ is the identity matrix of order n_{max}) so that now,

$$2\log\lambda^{(\nu)}\approx\boldsymbol{X}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^\top\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}\boldsymbol{X}.$$

Now $\boldsymbol{\Sigma}^{1/2}$ is symmetric since $\boldsymbol{\Sigma}$ is symmetric and hence $\boldsymbol{\Sigma}^{1/2}=(\boldsymbol{\Sigma}^{1/2})^\top$ and hence we can once again appeal to Appendix A.8 of Silvey [1975] to show that the matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}$ is symmetric. Hence there exists orthogonal $\boldsymbol{P}^*$ such that

$$(\boldsymbol{P}^*)^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}\boldsymbol{P}^* = \boldsymbol{\Lambda}^*$$

where $\boldsymbol{\Lambda}^*$ is diagonal. Thus, by Section 29.2 of Johnson and Kotz [1970],

$$2\log\lambda^{(\nu)} \xrightarrow{D} \sum_{i=1}^{n_{max}-2} \lambda_i^* U_i \tag{3.3.9}$$

as $\nu \to \infty$, where $U_1, U_2, ..., U_{n_{max}-2}$ are independent and identically distributed χ_1^2 random variables and $\lambda_1, \lambda_2, ..., \lambda_{n_{max}-2}$ are the non-zero eigenvalues of

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}$$

which, by matrix similarity, are the same as the non-zero eigenvalues of

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1/2} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{I}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{P}). \tag{3.3.10}$$

In the case of a $100\beta\%$ stratified sample of households, note that

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\beta}(\boldsymbol{I}_{\boldsymbol{\theta},\beta}^{-1} - \boldsymbol{P}_\beta) = (\beta\boldsymbol{\Sigma}_{\boldsymbol{\theta},1} + (1-\beta)\boldsymbol{I}_{\boldsymbol{\theta},1})(\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1} - \boldsymbol{P}_1)$$

3.3.3 Pseudo-Wald's W test

The idea for this test stems from the notion that under the null hypothesis, $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$ and hence, if H_0 is true, applying $\boldsymbol{h}$ to the unrestricted maximum pseudolikelihood estimator should find that $\boldsymbol{h}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{0}$. Taylor's theorem gives $\boldsymbol{h}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{h}(\boldsymbol{\theta}) + \boldsymbol{H}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, so, under H_0,

$$\boldsymbol{h}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{H}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

Hence, using Equation (3.2.6) and considering the sequence of epidemics $E^{(\nu)}$,

$$(m^{(\nu)})^{1/2}\boldsymbol{h}(\hat{\boldsymbol{\theta}}^{(\nu)}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}) \tag{3.3.11}$$

as $\nu \to \infty$ (see the theorem of Serfling [1980], Section 4.4.4). Hypothesis tests may be carried out from here but it may be convenient to note from the above that

$$(m^{(\nu)})^{1/2}\boldsymbol{h}(\hat{\boldsymbol{\theta}}^{(\nu)})(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1/2} \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I})$$

and hence that

$$m\boldsymbol{h}(\hat{\boldsymbol{\theta}}^{(\nu)})^\top(\boldsymbol{H}_{\boldsymbol{\theta}}^\top \boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{h}(\hat{\boldsymbol{\theta}}^{(\nu)}) \xrightarrow{D} \chi^2_{n_{max}-2}$$

as $\nu \to \infty$, since $\boldsymbol{H}_{\boldsymbol{\theta}}^\top \boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}$ is symmetric. This test is easy to adapt to the case of observing a $100\beta\%$ stratified sample of households since we have already established in Section 3.2.2 that

$$\boldsymbol{I}_{\boldsymbol{\theta},\beta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\beta}\boldsymbol{I}_{\boldsymbol{\theta},\beta}^{-1} = \beta(1-\beta)\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1} + \beta^2\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},1}\boldsymbol{I}_{\boldsymbol{\theta},1}^{-1}.$$

3.3.4 Pseudoscore statistic test

Under H_0, $\boldsymbol{U}(\dot{\boldsymbol{\theta}}|\boldsymbol{y})$ should be close to $\boldsymbol{0}$. Expanding $\boldsymbol{U}(\dot{\boldsymbol{\theta}}|\boldsymbol{y})$ about the unrestricted MpLE gives

$$\begin{aligned}\boldsymbol{U}(\dot{\boldsymbol{\theta}}|\boldsymbol{y}) &\approx \boldsymbol{U}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + \boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\dot{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \\ &= \boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\dot{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}).\end{aligned}$$

since $\boldsymbol{U}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = \boldsymbol{0}$. Thus

$$\begin{aligned}(m^{(\nu)})^{-1/2}\boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)}) &= [-(m^{(\nu)})^{-1}\boldsymbol{I}^{(\nu)}(\hat{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)})](m^{(\nu)})^{1/2}(\hat{\boldsymbol{\theta}}^{(\nu)} - \dot{\boldsymbol{\theta}}^{(\nu)}) \\ &\approx \boldsymbol{I}_{\boldsymbol{\theta}}(\boldsymbol{P} - \boldsymbol{I}_{\boldsymbol{\theta}}^{-1})\boldsymbol{Y} \\ &= -(\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P})\boldsymbol{Y}\end{aligned}$$

where $\boldsymbol{I}$ again denotes the identity matrix of size n_{max}.

As such, under H_0,

$$(m^{(\nu)})^{-1/2}\boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)}) \xrightarrow{D} N(\boldsymbol{0}, (\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P})^\top) \tag{3.3.12}$$

as $\nu \to \infty$ (again, see the theorem of Serfling [1980], Section 4.4.4). Now, $(\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P})(\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P}) = \boldsymbol{I} - 2\boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P} + \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P}\boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P}$ since we have already established that $\boldsymbol{P}\boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P} = \boldsymbol{P}$ in Section 3.3.2. Hence $\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P}$ is idempotent and has rank $n_{max} - 2$, the rank of $\boldsymbol{H}_{\boldsymbol{\theta}}$. Letting $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P})^\top$, noting that $\hat{\boldsymbol{\Sigma}}$ also has rank $n_{max} - 2$ and following a similar approach to Section 3.3.2, this implies that there exists an orthogonal matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\hat{\boldsymbol{\Sigma}}\boldsymbol{A}^\top = \boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix of size n_{max} whose first $n_{max} - 2$ diagonal elements are the non-zero eigenvalues of $\hat{\boldsymbol{\Sigma}}$ and remaining diagonal elements are 0.

Thus

$$(m^{(\nu)})^{-1/2}\boldsymbol{A}\boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)}) \xrightarrow{D} N(\boldsymbol{0},\boldsymbol{D}) \text{ as } \nu \to \infty.$$

Let $\boldsymbol{C}$ be the $(n_{max}-2)\times n_{max}$ matrix given by $\boldsymbol{C} = [\boldsymbol{I}\ \boldsymbol{0}]\boldsymbol{A}$, where $[\boldsymbol{I}\ \boldsymbol{0}]$ is the matrix formed by binding the identity matrix of order $(n_{max}-2)$ with the $(n_{max}-2)\times 2$ zero matrix. Then

$$(m^{(\nu)})^{-1/2}\boldsymbol{C}\boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)}) \xrightarrow{D} N(\boldsymbol{0},\tilde{\boldsymbol{D}}) \text{ as } \nu \to \infty,$$

where $\tilde{\boldsymbol{D}}$ is the diagonal matrix $\boldsymbol{D}$ reduced to size $n_{max}-2$. Therefore,

$$(m^{(\nu)})^{-1}\boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)})^{\top}\boldsymbol{C}^{\top}\tilde{\boldsymbol{D}}^{-1}\boldsymbol{C}\boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)}) \xrightarrow{D} \chi^2_{n_{max}-2} \text{ as } \nu \to \infty.$$

Note that $\tilde{\boldsymbol{D}} = \boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^{\top}$ and thus

$$(m^{(\nu)})^{-1}\boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)})^{\top}\boldsymbol{C}^{\top}(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^{\top})^{-1}\boldsymbol{C}\boldsymbol{U}(\dot{\boldsymbol{\theta}}^{(\nu)}|\boldsymbol{y}^{(\nu)}) \xrightarrow{D} \chi^2_{n_{max}-2} \text{ as } \nu \to \infty.$$

Returning to the scenario in which $\beta_n = \beta\alpha_n$ for some $\beta \in [0,1]$, note that $\boldsymbol{I}_{\boldsymbol{\theta},\beta}\boldsymbol{P}_{\beta} = \boldsymbol{I}_{\boldsymbol{\theta},1}\boldsymbol{P}_1$ and hence $\hat{\boldsymbol{\Sigma}}_{\beta} = (\boldsymbol{I}-\boldsymbol{I}_{\boldsymbol{\theta},1}\boldsymbol{P}_1)\boldsymbol{\Sigma}_{\boldsymbol{\theta},\beta}(\boldsymbol{I}-\boldsymbol{I}_{\boldsymbol{\theta},1}\boldsymbol{P}_1)^{\top}$.

3.4 Calculation of covariance matrices

We give a method for calculating the values of $\boldsymbol{B}$, $\boldsymbol{C}_{RR}(\tau)$, $\boldsymbol{C}_{RA}(\tau)$, $\boldsymbol{C}_{AR}(\tau)$ and $C_{AA}(\tau)$ for $\boldsymbol{R}_{\bullet}^{(\nu)} = \boldsymbol{U}^{(\nu)}(\boldsymbol{\theta}|\boldsymbol{y}^{(\nu)})$, and hence the matrices $\boldsymbol{I}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. All calculations given below involving n are defined for $n = 1, 2, ..., n_{max}$. It is assumed that the β_n (the proportion of households in the population that are observed and of size n) are known during these calculations. On a practical level, this may be as a result of knowing the total distribution of household sizes $\boldsymbol{\alpha}$ and assuming that the observed households represent a $100\beta\%$ stratified sample of these households as discussed Section 3.2.2 and at various points in Section 3.3.

The section begins by calculating $C^n_{RR}(\tau)$ using results from Section 3.2.2 and manipulating the final size probabilities given in Section 2.3 using differentiation. We then introduce a joint moment generating function for the final size and severity of a single household epidemic. The first moment of the severity is calculated using a simple Wald's identity formula. However, the second moment is trickier to calculate and thus we introduce a method of manipulating the joint moment generating function using Gontcharoff polynomials. The first and second moments of the severity of a single household epidemic can then be used to calculate $C_{AA}(\tau)$.

Further suitable differentiation of our joint moment generating function, aided by Gontcharoff polynomials is then used to calculate $\boldsymbol{C}_{AR}(\tau)$ and the component parts of $\boldsymbol{B}$. Finally, we utilise our joint moment generating function and Gontcharoff polynomials to show that $\boldsymbol{B}$ is a well-defined vector.

A formula for $C^n_{RR}(\tau)$ has already been established in Section 3.2.2, specifically,

$$c^n_{ij}(\tau) = \sum_{k=0}^{n} [\partial_{\theta_i} P_n(k|\boldsymbol{\theta})][\partial_{\theta_j} P_n(k|\boldsymbol{\theta})] P_n(k|\boldsymbol{\theta})^{-1}. \tag{3.4.1}$$

We therefore need values for the $P_n(k|\boldsymbol{\theta})$ $(k = 0, 1, ..., n)$ and their first derivatives with respect to θ_i $(i = 1, 2, ..., n_{max})$. It has already been established in Section 2.3 that $P_n(k|\boldsymbol{\theta})$ can be determined from the triangular system of linear equations

$$\sum_{j=0}^{k} \binom{n-j}{k-j} \frac{P_n(j|\boldsymbol{\theta})}{\phi((n-k)\lambda_L^{(n)})^j \pi^{n-k}} = \binom{n}{k}, \quad k = 0, 1, ..., n. \tag{3.4.2}$$

Clearly $\partial_{\theta_i} P_n(k|\boldsymbol{\theta}) = 0$ for all j if $i \neq 1, n$ so we need only focus on these two components of $\boldsymbol{\theta}$. Differentiating (3.4.2) with respect to $\theta_1 = \pi$ yields

$$\sum_{j=0}^{k} \binom{n-j}{k-j} \frac{\partial_{\theta_1} P_n(j|\boldsymbol{\theta})}{\phi((n-k)\lambda_L^{(n)})^j \pi^{n-k}} = \sum_{j=0}^{k} \binom{n-j}{k-j} \frac{(n-k) P_n(j|\boldsymbol{\theta})}{\phi((n-k)\lambda_L^{(n)})^j \pi^{n-k+1}} \tag{3.4.3}$$

and differentiating (3.4.2) with respect to $\theta_n = \lambda_L^{(n)}$ yields

$$\sum_{j=0}^{k} \binom{n-j}{k-j} \frac{\partial_{\theta_n} P_n(j|\boldsymbol{\theta})}{\phi((n-k)\lambda_L^{(n)})^j \pi^{n-k}} = \sum_{j=0}^{k} \binom{n-j}{k-j} \frac{j[(n-k)\phi'((n-k)\lambda_L^{(n)})] P_n(j|\boldsymbol{\theta})}{\phi((n-k)\lambda_L^{(n)})^{j+1} \pi^{n-k}} \tag{3.4.4}$$

for $k = 0, 1, ..., n$, where $\phi'(x) = \partial\phi(x)/\partial x$ (which depends entirely upon the distribution of T_I). Values for the $\partial_{\theta_i} P_n(k|\boldsymbol{\theta})$ $(n = 1, 2, ..., n_{max}; k = 0, 1, ..., n)$ can therefore be determined by solving (3.4.2) and, subsequently, (3.4.3) and (3.4.4) (all of which are triangular systems of linear equations). Calculation of $\boldsymbol{C}_{RR}(\tau)$ follows easily by inserting the results into (3.4.1). Note that $\partial_{\theta_n}\phi((n-k)\lambda_L^{(n)})$ values depend on the distribution of T_I but should be calculable if T_I is assumed to take a standard distribution. For example, Gamma distributed T_I are discussed in Section 3.6.

To calculate the remaining values, more information is required about the severity of a single-household epidemic in an initially fully susceptible household of size n in which each individual avoids global infection with probability π. Let Y be the final size of such an epidemic, A be its severity and define

$$\Phi_n(s,\vartheta) = \mathbb{E}[s^{n-Y}\exp(-A\vartheta)] \quad (0 \leq s \leq 1,\ \varphi \geq 0). \tag{3.4.5}$$

Note that moments of A are equivalent to moments of $A_{n,1}(\tau)$ and hence $\mathbb{E}[A]$ and $\mathbb{E}[A^2]$ are of interest. The expected severity $\mathbb{E}[A] = a_n(\tau)$ can be found using the Wald's identity $\mathbb{E}[A] = \mathbb{E}[Y]\mathbb{E}[T_I]$ (cf. Corollary 2.2 of Ball [1986]), where $\mathbb{E}[Y] = \sum_{k=0}^{n} kP_n(k|\boldsymbol{\theta})$ which can be found using Equation (3.4.2). To see that $\mathbb{E}[A] = \mathbb{E}[Y]\mathbb{E}[T_I]$ we follow the proof of Ball and Shaw [2016]. Label individuals in the household $1, 2, ..., n$ and, for $i = 1, 2, ..., n$, let I_i be the length of individual i's infectious period should they become infected and let $\chi_i = 1$ if individual i becomes infected and $\chi_i = 0$ otherwise. Then $A = \sum_{i=1}^{n} \chi_i I_i$ but, for given i, χ_i and I_i are independent. Thus

$$\mathbb{E}[A] = \sum_{i=1}^{n} \mathbb{E}[\chi_i]\mathbb{E}[I_i] = \mathbb{E}[T_I]\mathbb{E}\left[\sum_{i=1}^{n} \chi_i\right] = \mathbb{E}[Y]\mathbb{E}[T_I].$$

Finding $\mathbb{E}[A^2]$ however, does require manipulation of $\Phi_n(s,\vartheta)$ and this requires use of Gontcharoff polynomials. Before introducing these polynomials formally, we note that they could have been exploited earlier in this thesis to find final size probabilities for an epidemic. However, House et al. [2013] note that the matrix type methods as given in Section 2.3 are more efficient numerically. They also point out that Gontcharoff polynomial methods are numerically unstable for large n and whilst this is not an issue in most practical circumstances, it should be borne in mind if one wishes to remove the assumption used in this thesis of there being a maximum possible household size n_{max} when performing the calculations given below.

Let $\mathcal{U} = u_0, u_1, ...$ be a sequence of real numbers. The Gontcharoff polynomials associated with $\mathcal{U}$ are defined recursively by the system of equations

$$\sum_{w=0}^{v} \frac{u_w^{v-w}}{(v-w)!} G_w(s|\mathcal{U}) = \frac{s^v}{v!} \quad v = 0, 1, \ldots . \tag{3.4.6}$$

(See Gontcharoff [1937] for the introduction of Gontcharoff polynomials and Picard and Lefèvre [1990] for their use in a similar epidemiological context.)

Then, by Equation (3.9) of Ball et al. [1997], for $s \in \mathbb{R}$ and $\vartheta \in \mathbb{R}_+$

$$\Phi_n(s,\vartheta) = \sum_{w=0}^{n} \frac{n!}{(n-w)!} \phi(\vartheta + \lambda_L^{(n)} w)^{n-w} \pi^w G_w(s|\mathcal{U}), \tag{3.4.7}$$

where $\mathcal{U}$ is given by $u_w = \phi(\vartheta + \lambda_L^{(n)} w)$ $(w = 0, 1, ...)$.

From the definition of $\Phi_n(s,\vartheta)$, $\mathbb{E}[A^2] = \partial_\vartheta^{(2)} \Phi_n(s,\vartheta)$ evaluated at $s = 1$, $\vartheta = 0$, where $\partial_\vartheta^{(i)}$ denotes the partial derivative $\partial^i / \partial \vartheta^i$ $(i = 1, 2, ...)$. Differentiating (3.4.7) with $s = 1$ gives,

$$\begin{aligned}
&\partial_\vartheta^{(2)} \Phi_n(1,\vartheta) \\
&= \partial_\vartheta \sum_{w=0}^{n-1} \frac{n!}{(n-w-1)!} \pi^w \phi'(\vartheta + \lambda_L^{(n)} w) \phi(\vartheta + \lambda_L^{(n)} w)^{n-w-1} G_w(1|\mathcal{U}) \\
&\quad + \partial_\vartheta \sum_{w=0}^{n} \frac{n!}{(n-w)!} \pi^w \phi(\vartheta + \lambda_L^{(n)} w)^{n-w} [\partial_\vartheta G_w(1|\mathcal{U})] \\
&= \sum_{w=0}^{n-2} \frac{n!}{(n-w-2)!} \pi^w [\phi'(\vartheta + \lambda_L^{(n)} w)]^2 \phi(\vartheta + \lambda_L^{(n)} w)^{n-w-2} G_w(1|\mathcal{U}) \\
&\quad + 2 \sum_{w=0}^{n-1} \frac{n!}{(n-w-1)!} \pi^w \phi'(\vartheta + \lambda_L^{(n)} w) \phi(\vartheta + \lambda_L^{(n)} w)^{n-w-1} [\partial_\vartheta G_w(1|\mathcal{U})] \\
&\quad + \sum_{w=0}^{n} \frac{n!}{(n-w)!} \pi^w \phi(\vartheta + \lambda_L^{(n)} w)^{n-w} [\partial_\vartheta^{(2)} G_w(1|\mathcal{U})].
\end{aligned} \tag{3.4.8}$$

A recursive formula for the derivatives of the Gontcharoff polynomials with respect to ϑ can be found by differentiating (3.4.6). In our case, for $v = 1, 2, ...n$,

$$\begin{aligned}
&\sum_{w=0}^{v} \frac{\phi(\vartheta + \lambda_L^{(n)} w)^{v-w}}{(v-w)!} \partial_\vartheta G_w(s|\mathcal{U}) \\
&\qquad = - \sum_{w=0}^{v-1} \frac{\phi'(\vartheta + \lambda_L^{(n)} w) \phi(\vartheta + \lambda_L^{(n)} w)^{v-w-1}}{(v-w-1)!} G_w(s|\mathcal{U})
\end{aligned} \tag{3.4.9}$$

(for $v = 0$ note that $G_0(s|\mathcal{U}) \equiv 1$ and hence all of its derivatives are 0) and thus

$$\begin{aligned}\sum_{w=0}^{v} &\frac{\phi(\vartheta+\lambda_L^{(n)}w)^{v-w}}{(v-w)!}\partial_\vartheta^{(2)}G_w(s|\mathcal{U})\\ &=-2\sum_{w=0}^{v-1}\frac{\phi'(\vartheta+\lambda_L^{(n)}w)\phi(\vartheta+\lambda_L^{(n)}w)^{v-w-1}}{(v-w-1)!}\partial_\vartheta G_w(s|\mathcal{U})\\ &\quad-\sum_{w=0}^{v-1}\frac{\phi''(\vartheta+\lambda_L^{(n)}w)\phi(\vartheta+\lambda_L^{(n)}w)^{v-w-1}}{(v-w-1)!}G_w(s|\mathcal{U})\\ &\quad-\sum_{w=0}^{v-2}\frac{[\phi'(\vartheta+\lambda_L^{(n)}w)]^2\phi(\vartheta+\lambda_L^{(n)}w)^{v-w-2}}{(v-w-2)!}G_w(s|\mathcal{U})\end{aligned} \tag{3.4.10}$$

provide numerically calculable formulae for $\partial_\vartheta G_w(1|\mathcal{U})$ and $\partial_\vartheta^{(2)}G_w(1|\mathcal{U})$ respectively. If $\phi'(x)$ and $\phi''(x)$ can be calculated (from knowing the distribution of T_I) then $\mathbb{E}[A^2]$ follows from evaluating (3.4.6), (3.4.8), (3.4.9) and (3.4.10) at $\vartheta = 0$ (again noting that all of these are systems of linear equations), thus $C_{AA}^{(n)} = \mathbb{E}[A_{n,1}(\tau)^2] - \mathbb{E}[A_{n,1}(\tau)]^2$ may be evaluated and C_{AA} follows.

We now turn our attention to the vector $\boldsymbol{C}_{RA}^n(\tau)$ whose i^{th} component is given by $\text{cov}(R_{ni,1}(\tau), A_{n,1}(\tau)) = \mathbb{E}[R_{ni,1}(\tau)A_{n,1}(\tau)]$, since $\mathbb{E}[R_{ni,1}(\tau)] = r_{ni}(\tau) = 0$ (c.f. (3.2.3)) and hence $\mathbb{E}[R_{ni,1}(\tau)]\mathbb{E}[A_{n,1}(\tau)] = 0$. Now, $\boldsymbol{R}_{n,1}(\tau)$ is a vector-valued function of the final size of a single-household epidemic with parameters $\boldsymbol{\theta}$, specifically the score statistic, and as such can be written as $\boldsymbol{R}_{n,1}(\tau) = \sum_{k=0}^n \boldsymbol{U}(\theta|y_{n,1} = k)\mathbb{1}_{\{y_{n,1}=k\}}$. (Here $\boldsymbol{U}(\theta|y_{n,1} = k)$ refers to the score statistic based on observing a single household of size n with k recovered individuals at the end of the epidemic.) Thus,

$$\begin{aligned}\boldsymbol{C}_{RA}^n(\tau) =&\mathbb{E}\left[A_{n,1}(\tau)\sum_{k=0}^n \boldsymbol{U}(\theta|y_{n,1}=k)\mathbb{1}_{\{y_{n,1}=k\}}\right]\\ =&\sum_{k=0}^n \boldsymbol{U}(\theta|y_{n,1}=k)\mathbb{E}\left[A_{n,1}\mathbb{1}_{\{y_{n,1}=k\}}\right].\end{aligned} \tag{3.4.11}$$

We now look to manipulate $\Phi_n(s,\vartheta)$ to obtain $\boldsymbol{C}_{RA}^n(\tau)$, using (3.4.11).

Let $X = n - Y$ in the definition of $\Phi_n(s,\vartheta)$ given in (3.4.5) and, for $i = 1, 2, ...n$, let $\Phi_n^{(i)}(s,\vartheta)$ denote the i^{th} derivative of $\Phi_n(s,\vartheta)$ with respect to s. Then,

$$\begin{aligned}\Phi_n(s,\vartheta) =&\mathbb{E}[s^X e^{-A\vartheta}]\\ =&\sum_{k=0}^n \mathbb{P}(X=k)s^k\mathbb{E}[e^{-A\vartheta}|X=k],\end{aligned}$$

so, for $i = 0, 1, ..., n$,

$$\Phi_n^{(i)}(s, \vartheta) = \sum_{k=i}^{n} \frac{k!}{(k-i)!} \mathbb{P}(X = k) s^{k-i} \mathbb{E}[e^{-A\vartheta} | X = k]$$

and

$$\Phi_n^{(i)}(0, \vartheta) = i! \mathbb{P}(X = i) \mathbb{E}[e^{-A\vartheta} | X = i],$$

since all terms other than the first in the sum disappear as a result of setting $s = 0$. Consequently, for $k = 0, 1, ..., n$,

$$\mathbb{E}[e^{-A\vartheta} \mathbb{1}_{\{X=k\}}] = \Phi_n^{(k)}(0, \vartheta)/k!$$

and thus,

$$\begin{aligned}
\mathbb{E}[A \mathbb{1}_{\{X=k\}}] &= [\partial_\vartheta \Phi_n^{(k)}(0, \vartheta)]/k! \;|_{\vartheta=0} \\
&= -\frac{1}{k!} \Bigg\{ \sum_{w=0}^{n-1} \frac{n!}{(n-w-1)!} \pi^w \phi'(\vartheta + \lambda_L^{(n)} w) \phi(\vartheta + \lambda_L^{(n)} w)^{n-w-1} G_w^{(k)}(0|\mathcal{U}) \\
&\quad + \sum_{w=0}^{n} \frac{n!}{(n-w)!} w \pi^{w-1} \phi(\vartheta + \lambda_L^{(n)} w)^{n-w} [\partial_\vartheta G_w^{(k)}(0|\mathcal{U})] \Bigg\} \Bigg|_{\vartheta=0}, \qquad (3.4.12)
\end{aligned}$$

where $G_w^{(k)}(s|\mathcal{U})$ denotes the k^{th} derivative of $G_w(s|\mathcal{U})$ with respect to s. Equation (2.7) of Picard and Lefèvre [1990] shows that $G_w^{(k)}(s|\mathcal{U}) = G_{w-k}(s|\mathcal{U}^{(k)})$ (where $\mathcal{U}^{(k)}$ is the sequence $u_k, u_{k+1}, ...$) if $k \leq w$ and $G_w^{(k)}(s|\mathcal{U}) = 0$ otherwise. Hence the Gontcharoff polynomials and their derivatives with respect to s, evaluated at $s = 0$, can be calculated easily using (3.4.6) and all of their first derivatives with respect to ϑ can be found using the same technique as was used in (3.4.9). The vector $\boldsymbol{C}_{RA}^n(\tau)$ can thus be calculated using (3.4.11) and (3.4.12) and calculation of $\boldsymbol{C}_{RA}(\tau)$ follows. Note also that $\boldsymbol{C}_{AR}(\tau)$ is simply given by $\boldsymbol{C}_{RA}(\tau)^\top$ (see the respective definitions in Section 3.1.3).

Recall from Section 3.1.3 that $\boldsymbol{B} = D_{\boldsymbol{R}}(m_H - D_A)^{-1}$ where D_f denotes the first derivative of a continuous vector-valued function f with respect to τ. Since $\pi = e^{-\lambda_G \tau}$ and hence $\partial\pi/\partial\tau = -\lambda_G \pi$, it follows that $D_f = -\lambda_G \pi \tilde{D}_f$ where $\tilde{D}_f$ denotes the first derivative of f with respect to π. Therefore $\boldsymbol{B} = -\lambda_G \pi \tilde{D}_{\boldsymbol{R}}(m_H + \lambda_G \pi \tilde{D}_A)^{-1}$. It is clear from the definition of $\boldsymbol{I_\theta}$ that $\tilde{D}_{\boldsymbol{R}}$ is equal to the first column of $\boldsymbol{I_\theta}$ (or $\boldsymbol{C}_{RR}(\tau)$). This leaves only $\tilde{D}_A$ to be calculated. Observe, by using a similar method to the derivation of (3.4.8), that

$$\begin{aligned}a(\lambda_G\tau) = &\sum_{w=0}^{n-1}\frac{n!}{(n-w-1)!}\pi^w\phi'(\vartheta+\lambda_L^{(n)}w)\phi(\vartheta+\lambda_L^{(n)}w)^{n-w-1}G_w(1|\mathcal{U})\\ &+\sum_{w=0}^{n}\frac{n!}{(n-w)!}\pi^w\phi(\vartheta+\lambda_L^{(n)}w)^{n-w}[d_\vartheta G_w(1|\mathcal{U})]\end{aligned}$$

and hence

$$\begin{aligned}\tilde{D}_A = &\sum_{w=0}^{n-1}\frac{n!}{(n-w-1)!}w\pi^{w-1}\phi'(\vartheta+\lambda_L^{(n)}w)\phi(\vartheta+\lambda_L^{(n)}w)^{n-w-1}G_w(1|\mathcal{U})\\ &+\sum_{w=0}^{n}\frac{n!}{(n-w)!}w\pi^{w-1}\phi(\vartheta+\lambda_L^{(n)}w)^{n-w}[d_\vartheta G_w(1|\mathcal{U})].\end{aligned} \tag{3.4.13}$$

Equations (3.4.6), (3.4.9), (3.4.13) can thus be used to evaluate $\boldsymbol{B}$. It is now possible to calculate the matrices $\boldsymbol{I_\theta}$ and $\boldsymbol{\Sigma_\theta}$ from the above. In most practical situations, $\boldsymbol{\theta}$ is unknown but the matrices can be estimated by evaluating at the MpLE $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Observe that for $\boldsymbol{B}$ to be well-defined, we require $m_H - D_A \neq 0$. Now, we have already established in Section 3.1.2 that 0 and τ are roots of $m_H t = a(t)$. If $a(t)$ is strictly concave then these are the only two roots of the equation and both are proper crossing points (not tangent), meaning that $m_H \neq a'(\tau) = D_A$ as required. The function $a(t)$ is strictly concave if and only if its second derivative $a''(t) < 0$ for all t. First note that $a''(t) = \sum_{n=1}^{n_{max}} \alpha_n a_n''(t)$ which is strictly negative if $a_n''(t)$ is strictly negative for each n . Appealing to the notation used above when discussing $\Phi_n(s,\vartheta)$, recall that $\mathbb{E}[A] = \mathbb{E}[Y]\mathbb{E}[T_I]$ and that

$$\begin{aligned}\Phi_n(s,\vartheta) &= \mathbb{E}[s^{n-Y}\exp(-A\vartheta)],\\ \Phi_n^{(1)}(s,\vartheta) &= \mathbb{E}[(n-Y)s^{n-Y-1}\exp(-A\vartheta)],\\ \Phi_n^{(1)}(1,0) &= n-\mathbb{E}[Y],\\ \mathbb{E}[Y] &= n-\Phi_n^{(1)}(1,0).\end{aligned}$$

Thus,

$$\begin{aligned}a_n(t) =& n\mathbb{E}[T_I] - \mathbb{E}[T_I]\sum_{w=0}^{n}\frac{n!}{(n-w)!}(e^{-\lambda_G t})^w\phi(\lambda_L^{(n)}w)^{n-w}G_{w-1}(1|\mathcal{U}^{(1)}) \text{ and}\\ a_n''(t) =& -\mathbb{E}[T_I]\sum_{w=0}^{n}\frac{n!}{(n-w)!}(w\lambda_G)^2e^{-w\lambda_G t}\phi(\lambda_L^{(n)}w)^{n-w}G_{w-1}(1|\mathcal{U}^{(1)}).\end{aligned} \tag{3.4.14}$$

Now, $\phi(x)$ is positive for any $x \in \mathcal{R}$ since ϕ is a moment generating function and $\lambda_G > 0$ by definition. Hence $a_n''(t) < 0$ if $G_{w-1}(1|\mathcal{U}^{(1)}) > 0$ for $w = 1, 2, ...$ Now $\mathcal{U}^{(1)}$ is given by the sequence $u_w = \mathbb{E}[\exp(-(\lambda_L^{(n)}(w+1))T_I)]$, which is monotone non-increasing in w but strictly positive. For $w = 0, 1, ..$ the integral representation of $G_w(|\mathcal{U}^{(1)})$ is given by

$$G_w(|\mathcal{U}^{(1)}) = \int_{u_0}^{1} \int_{u_1}^{\xi_0} \int_{u_2}^{\xi_1} ... \int_{u_{w-1}}^{\xi_{w-2}} d\xi_0 d\xi_1 ... d\xi_{w-1}$$

(see Equation (2.5) of Picard and Lefèvre [1990]). Since $u_0 < 1$, it follows immediately that the $G_w(1|\mathcal{U}^{(1)})$ are strictly positive (c.f. Section 3.2 of Ball et al. [1997]). Therefore $a(t)$ is indeed concave and hence $\boldsymbol{B}$ is well-defined.

3.5 Results relating to the dependence between outcomes in different households

No knowledge of the parameter $\boldsymbol{\beta}$ is necessary to determine the MpLE $\hat{\boldsymbol{\theta}}$, since (3.2.1) shows that only observed data are included in the log-pseudolikelihood function and, for $n = 1, 2, ..., n_{max}$; $k = 0, 1, ..., n$ calculation of $P_n(k|\boldsymbol{\theta})$ does not even rely on knowledge of the related population structure parameter $\boldsymbol{\alpha}$. However, the asymptotic covariances of the parameter estimator given by $\hat{\boldsymbol{\theta}}$, and thus asymptotic properties of hypothesis tests relating to these estimators, are affected by the matrices $\boldsymbol{I_\theta}$ and $\boldsymbol{\Sigma_\theta}$ which are shown to depend on $\boldsymbol{\beta}$ in Section 3.4. Since $\boldsymbol{I_\theta} = \sum_{n=1}^{n_{max}} \beta_n \boldsymbol{I}_n(\boldsymbol{\theta})$ and, for $n = 1, 2, ..., n_{max}$, the number of observed households of size n is given by $\beta_n m_n$ it is clear that hypothesis tests using only $\boldsymbol{I_\theta}$ (i.e. by assuming that all observed household outcomes are mutually independent in the manner discussed at the end of Section 3.2.2, thus making $\boldsymbol{\Sigma_\theta}$ irrelevant) are only affected by observed households and, as such, do not depend on $\boldsymbol{\beta}$. However, $\boldsymbol{\Sigma_\theta}$ cannot generally be ignored, since the relationship between $\boldsymbol{\Sigma_\theta}$ and $\boldsymbol{\beta}$ is more complicated than that between $\boldsymbol{I_\theta}$ and $\boldsymbol{\beta}$, and hence such assumptions regarding the lack of dependence on $\boldsymbol{\beta}$ cannot be made.

In this section it is shown that hypothesis tests of the form outlined in Section 3.3 do not require knowledge of $\boldsymbol{\Sigma_\theta}$ (and hence any knowledge of $\boldsymbol{\beta}$, including the overall population structure $\boldsymbol{\alpha}$) if none of the constraints given by the vector

$\boldsymbol{h}(\boldsymbol{\theta})$ contain $\theta_1 = \pi$, such as in the tests suggested in Section 3.3.1. We consider the relationship between $\boldsymbol{I_\theta}$ and $\boldsymbol{\Sigma_\theta}$ but must first introduce some notation. For a given matrix $\boldsymbol{A}$, let $\boldsymbol{A}_{[n,]}$ be the row vector given by the n^{th} row of $\boldsymbol{A}$, $\boldsymbol{A}_{[,n]}$ be the column vector given by the n^{th} column of $\boldsymbol{A}$ and, for $b < c$, $\boldsymbol{A}_{[b:c]}$ be the square matrix of size $c - b$ formed using only elements a_{ij} of $\boldsymbol{A}$ for which $b \leq i, j \leq c$.

Recall from (3.2.6) that, for large m, the covariance matrix of $\hat{\boldsymbol{\theta}}$ is approximately given by $m^{-1}\boldsymbol{I_\theta}^{-1}\boldsymbol{\Sigma_\theta}\boldsymbol{I_\theta}^{-1}$ and that $\boldsymbol{\Sigma_\theta} = \boldsymbol{C}_{RR}(\tau) + \boldsymbol{B}\boldsymbol{C}_{AR}(\tau) + \boldsymbol{C}_{RA}(\tau)\boldsymbol{B}^\top + \boldsymbol{B}C_{AA}(\tau)\boldsymbol{B}^\top$. Then, using the fact that $\boldsymbol{I_\theta} = \boldsymbol{C}_{RR}(\tau)$,

$$\begin{aligned}&\boldsymbol{I_\theta}^{-1}\boldsymbol{\Sigma_\theta}\boldsymbol{I_\theta}^{-1}\\ =&\boldsymbol{I_\theta}^{-1} + \boldsymbol{I_\theta}^{-1}\boldsymbol{B}\boldsymbol{C}_{AR}(\tau)\boldsymbol{I_\theta}^{-1} + \boldsymbol{I_\theta}^{-1}\boldsymbol{C}_{RA}(\tau)\boldsymbol{B}^\top\boldsymbol{I_\theta}^{-1} + \boldsymbol{I_\theta}^{-1}\boldsymbol{B}C_{AA}(\tau)\boldsymbol{B}\boldsymbol{I_\theta}^{-1}.\end{aligned}$$

Now, from the discussion in Section 3.4, $\boldsymbol{B} = K(\boldsymbol{I_\theta})_{[,1]}$, where $K = -\lambda_G\pi(m_H - D_A)^{-1}$ is a scalar, and thus,

$$\boldsymbol{I_\theta}^{-1}\boldsymbol{B} = K\boldsymbol{I_\theta}^{-1}(\boldsymbol{I_\theta})_{[,1]} = (K, 0, 0, ..., 0)^\top \tag{3.5.1}$$

and, similarly,

$$\boldsymbol{B}^\top\boldsymbol{I_\theta}^{-1} = (K, 0, 0, ..., 0).$$

It follows easily that

$$\begin{aligned}&(\boldsymbol{I_\theta}^{-1}\boldsymbol{B}\boldsymbol{C}_{AR}(\tau)\boldsymbol{I_\theta}^{-1})_{[2:n_{max}]} = (\boldsymbol{I_\theta}^{-1}\boldsymbol{C}_{RA}(\tau)\boldsymbol{B}^\top\boldsymbol{I_\theta}^{-1})_{[2:n_{max}]}\\ =&(\boldsymbol{I_\theta}^{-1}\boldsymbol{B}C_{AA}(\tau)\boldsymbol{B}\boldsymbol{I_\theta}^{-1})_{[2:n_{max}]} = \boldsymbol{0}\end{aligned}$$

and thus that

$$(\boldsymbol{I_\theta}^{-1}\boldsymbol{\Sigma_\theta}\boldsymbol{I_\theta}^{-1})_{[2:n_{max}]} = (\boldsymbol{I_\theta}^{-1})_{[2:n_{max}]}. \tag{3.5.2}$$

Recalling from Section 3.3.1 that $(H_\theta)_{ij} = \partial h_j/\partial\theta_i$, note that $(\boldsymbol{H_\theta})_{[1,]} = \boldsymbol{0}^\top$ and that $(\boldsymbol{H_\theta}^\top)_{[,1]} = \boldsymbol{0}$ (where $\boldsymbol{0}$ now represents the zero column vector) if the condition that none of the constraints comprising $\boldsymbol{h}(\boldsymbol{\theta})$ contains $\theta_1 = \pi$ is satisfied. Hence, using (3.5.2),

$$\boldsymbol{H_\theta}^\top\boldsymbol{I_\theta}^{-1}\boldsymbol{\Sigma_\theta}\boldsymbol{I_\theta}^{-1}\boldsymbol{H} = \boldsymbol{H_\theta}^\top\boldsymbol{I_\theta}^{-1}\boldsymbol{H}. \tag{3.5.3}$$

Using (3.3.11) and (3.5.3) and the discussion at the beginning of this section, it is immediately clear that the pseudo-Wald's test does not require calculation of $\boldsymbol{\Sigma_\theta}$ if $(\boldsymbol{H_\theta})_{[1,]} = \boldsymbol{0}^\top$.

By splitting $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ into its component parts and recalling (3.3.12), it is also clear that the pseudoscore statistic test does not need $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ to be calculated if

$$(\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P})\boldsymbol{B} = \mathbf{0}. \tag{3.5.4}$$

Now, using the definition of $\boldsymbol{P}$ given in (3.3.4),

$$(\boldsymbol{I} - \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{P})\boldsymbol{B} = \boldsymbol{B} - \boldsymbol{B} + \boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{B}$$

and hence we need only show that

$$\boldsymbol{I}_{\boldsymbol{\theta}}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{B} = \mathbf{0}.$$

However, this follows immediately from (3.5.1) if we have the condition that our constraints do not include θ_1 (and hence that $(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top})_{[,1]} = \mathbf{0}$).

Finally, we consider the pseudolikelihood ratio test which, recalling (3.3.10), depends upon the non-zero eigenvalues of

$$\mathbf{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}. \tag{3.5.5}$$

Considering the component parts of $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ as above and noting that $\boldsymbol{B}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{\mathbf{-1}}\boldsymbol{H}_{\boldsymbol{\theta}} = \mathbf{0}^{\top}$, (3.5.5) becomes

$$(\boldsymbol{I}_{\boldsymbol{\theta}} + \boldsymbol{B}\boldsymbol{C}_{AR}(\tau))\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}.$$

Now,

$$\begin{aligned}
&\big[(\boldsymbol{I}_{\boldsymbol{\theta}} + \boldsymbol{B}\boldsymbol{C}_{AR}(\tau))\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\big]^2\\
=&\big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{B}\boldsymbol{C}_{AR}(\tau)\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\big]^2\\
=&\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\\
&+\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{B}\boldsymbol{C}_{AR}(\tau)\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\\
&+\boldsymbol{B}\boldsymbol{C}_{AR}(\tau)\bigg(\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\\
&+\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{B}\boldsymbol{C}_{AR}(\tau)\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\bigg)\\
=&\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{B}\boldsymbol{C}_{AR}(\tau)\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\\
=&(\boldsymbol{I}_{\boldsymbol{\theta}} + \boldsymbol{B}\boldsymbol{C}_{AR}(\tau))\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{H}_{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}
\end{aligned}$$

since $\boldsymbol{H}_{\boldsymbol{\theta}}^{\top}\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}\boldsymbol{B} = \mathbf{0}$. Hence the matrix given in (3.5.5) is idempotent under our condition on the constraints and therefore its only non-zero eigenvalues are 1 and the pseudolikelihood test does not require any knowledge of $\boldsymbol{\beta}$. Moreover, the test given by (3.3.9) may now be simplified to say that, as $\nu \to \infty$, $2\log\lambda^{(\nu)}$ converges in distribution to a random variable taking a χ^2_r distribution, where r is the rank of $\boldsymbol{H_\theta}$.

3.6 Applications

We seek to illustrate the tests outlined in Section 3.3 using real data and simulation studies. In particular, we look to test for the dependence of local contact rate on household size using influenza data from Tecumseh, Michigan and Seattle, Washington, which have been studied extensively within the mathematical epidemiology field. These data are used since any results obtained can be compared to the work of previous authors. In general, we consider tests on the following pairs of hypotheses, as discussed in Section 3.3.1,

$$\begin{aligned} H_0 :\ & \lambda_L^{(2)} = \lambda_L^{(3)} = ... = \lambda_L^{(n_{max})} \\ \text{vs } H_1 :\ & \lambda_L^{(i)} \neq \lambda_L^{(j)} \quad \text{for some } i \neq j \quad i,j = 2,3,...,n_{max} \end{aligned} \tag{3.6.1}$$

and

$$\begin{aligned} H_0 :\ & \frac{\log\left(\frac{\lambda_L^{(3)}}{\lambda_L^{(2)}}\right)}{\log\left(\frac{2}{3}\right)} = \frac{\log\left(\frac{\lambda_L^{(4)}}{\lambda_L^{(2)}}\right)}{\log\left(\frac{2}{4}\right)} = ... = \frac{\log\left(\frac{\lambda_L^{(n_{max})}}{\lambda_L^{(2)}}\right)}{\log\left(\frac{2}{n_{max}}\right)} \quad (i = 1,2,...,n_{max}-3), \\ \text{vs } H_1 :\ & \frac{\log\left(\frac{\lambda_L^{(i)}}{\lambda_L^{(2)}}\right)}{\log\left(\frac{2}{i}\right)} \neq \frac{\log\left(\frac{\lambda_L^{(j)}}{\lambda_L^{(2)}}\right)}{\log\left(\frac{2}{j}\right)} \quad \text{for some } i \neq j. \end{aligned} \tag{3.6.2}$$

Recall from Section 3.3.1 that H_0 in (3.6.2) refers to the model of Cauchemez et al. [2004] in which, for $n = 2,3,...,n_{max}$, $\lambda_L^{(n)} = n^{-\eta}\lambda_L$ for some λ_L, η. We shall refer to H_0 of (3.6.2) as the *Cauchemez model*, H_0 of (3.6.1) as the *basic model* and the alternative hypotheses as the *unrestricted model* on the local contact parameters. We perform formal goodness-of-fit tests in Section 3.6.3 to show that all three of these models provide a reasonable fit to our data, thus validating the use of the above hypothesis tests as tools for model selection.

3.6.1 Testing against the unrestricted model

We begin by using real data to test the mathematically less complex basic model and Cauchemez model against the unrestricted model for local contacts rates. The data comprise two influenza outbreaks in Seattle, Washington, reported in Fox and Hall [1980], from 1975-76 and 1978-79, and two outbreaks in Tecumseh, Michigan from 1977–78 and 1980-81, reported in Monto et al. [1985]. Ball et al. [1997] offers the original lead for considering the Tecumseh data to be taken from a households epidemic model similar to that used here and both sets of data have been studied under household epidemic models in Clancy and O′Neill [2007], Neal [2012] and Neal and Kypraios [2015]. References within those papers cite a considerable amount of further literature using one or both data sets.

The data for the Tecumseh and Seattle outbreaks are given in Tables 3.1 and 3.2 respectively. The Tecumseh data refer to outbreaks of the same influenza A strain (H3N2 virus) and consist of an approximately 10% sample of households in the population, however we have already seen in Section 3.5 that knowledge of the population structure beyond what is observed is unnecessary for the tests on local contact parameters that we wish to perform. We treat these data separately and also consider combined data, assuming that both share common local contact parameters and the same global contact parameter λ_G.

Note that if the population structure α differed for the two epidemics then the global infectious escape probability π would differ for the two epidemics under the assumptions made above. For the sake of simplicity when considering the combined data, we assume that the population structure of Tecumseh did not change between the two outbreaks. This is a reasonable assertion given that the epidemics take place a mere three years apart in the same town and the greater prevalence of larger households in the 1980-81 data set can possibly be explained by assuming that the 1977-78 data represent a cross-sectional sample of households and that the 1980-81 set deliberately recruited more of the larger households, since minimal data are available on them in the 1977-78 sample. Under these assumptions and considering the arguments of Section 3.5, we combine the data sets in the most obvious manner (by simply adding one to the other) when performing hypothesis tests relating to local contact pa-

Table 3.1: Observed final size data from two influenza A epidemics (H3N2 virus) in Tecumseh, Michigan

No. infected	Household size (1977-78)							Household size (1980-81)						
per household	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	66	87	25	22	4	0	0	44	62	47	38	9	3	2
1	13	14	15	9	4	0	0	10	13	8	11	5	3	0
2		4	4	9	2	1	0		9	2	7	3	0	0
3			4	3	1	1	1			3	5	1	0	0
4				1	1	0	0				1	0	0	0
5					0	0	0					1	0	0
6						0	0						0	0
7							0							0
Total	79	105	48	44	12	2	1	54	84	60	62	19	6	2

rameter values on the combined data.

The Seattle data are taken from outbreaks of different influenza strains, namely the influenza B outbreak of 1975-76 and an influenza A (H1N1) outbreak of 1978-79, so it is only appropriate to treat these as separate data sets. Note also from Table 3.2 that we only have information for households up to size-3 for the 1978-79 outbreak. Therefore, the test given in (3.6.2) is not applicable to these data since households of at least three different sizes (ignoring size-1 households) are needed to constrain the Cauchemez model in such a way that the restricted MpLE $\dot{\boldsymbol{\theta}}$ is not equal to the unrestricted MpLE $\hat{\boldsymbol{\theta}}$.

Following the lead of Addy et al. [1991], we consider the infectious period to take a gamma distribution with a mean of 4.1 days and shape parameter 2 for the Tecumseh data. Returning briefly to Section 3.4, it is noted that calculation of the covariance matrices $\boldsymbol{I_\theta}$ and $\boldsymbol{\Sigma_\theta}$ relies on knowing the derivatives of the moment generating function of T_I. For gamma distributed T_I with shape parameter a and scale parameter b,

$$\phi(t) = \left(1 + \frac{t}{b}\right)^{-a},$$

so

$$\phi'(t) = -\frac{a}{b}\left(1 + \frac{t}{b}\right)^{-(a+1)} \quad \text{and} \quad \phi''(t) = \frac{a(a+1)}{b^2}\left(1 + \frac{t}{b}\right)^{-(a+2)},$$

Table 3.2: Observed final size data from the 1975-76 influenza B outbreak and the 1978-9 influenza A (H1N1) outbreak, both in Seattle, Washington

No. infected	Household size (1975-76)					Household size (1978-79)		
per household	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**
0	9	12	18	9	4	15	12	4
1	1	6	6	3	3	11	17	4
2		2	3	4	0		21	4
3			1	3	2			5
4				0	0			
5					0			
Total	10	20	28	20	9	26	50	17

where the derivatives given above are with respect to t. Such a precedent has not been set for the Seattle data with Clancy and O'Neill [2007], for example, suggesting both constant and exponential infectious periods for these data. As such, we use the same gamma distribution, with a mean of 4.1 days, as the Tecumseh data for the Seattle data (recalling that only the shape of the distribution is important since the estimates of $\boldsymbol{\theta}$ will adjust for scale accordingly). This allows for easier comparison between the data sets since these data are simply being used to illustrate the methods for hypothesis testing discussed in this chapter.

Tables 3.3 and 3.4 give the maximum pseudolikelihood estimates for the unknown parameters for each of the three models for all for outbreaks and the combined Tecumseh data as well as p-values for each of the three hypothesis tests outlined in Section 3.3 for testing the basic and Cauchemez models against the unrestricted model. Note that for the Tecumseh 1980-81 data we would obtain $\hat{\theta}_6 = 0$, since none of the three globally contacted households of size-6 in these data experienced any local contact. As such, our unrestricted MpLE, $\hat{\boldsymbol{\theta}}$, lies at the edge of our parameter space and the theory of Section 3.3 breaks down. To account for this we follow the precedent of Addy et al. [1991] and Ball et al. [1997], who omit households of 6 and 7 individuals in their studies, when looking at the Tecumseh 1980-81 data. We re-introduce these households into the combined Tecumseh data.

Immediate observations from Table 3.3 are that the unrestricted model does

Table 3.3: Parameter estimates and hypothesis test results on the Tecumseh and Seattle data for the pseudolikelihood-ratio test (LRT), pseudo-Wald's test (Wald) and pseudoscore statistic test (Score). Estimators and hypothesis tests using the basic and Cauchemez (Cauch) models as the null hypothesis are shown. The number of households in each population is given in brackets next to the epidemic location and date

	Model	Parameter estimate			Hypothesis test p-value		
		π	λ_L	η	LRT	Wald	Score
Tecumseh 1977-78 (289)	Basic	0.8542	0.0361		0.8642	0.9240	0.7374
	Cauch	0.8544	0.1075	0.8050	0.9878	0.9877	0.9868
Tecumseh 1980-81 (279)	Basic	0.8792	0.0513		0.1630	0.3223	0.0027
	Cauch	0.8798	0.3633	1.5203	0.9985	0.9982	0.9987
Tecumseh comb. (576)	Basic	0.8699	0.0417		0.1205	0.0681	0.0887
	Cauch	0.8703	0.2360	1.3072	0.9495	0.9857	0.9905
Seattle 1975-76 (87)	Basic	0.8319	0.0333		0.9104	0.8964	0.9254
	Cauch	0.8324	0.0967	0.8080	0.9809	0.9831	0.9827
Seattle 1978-79 (93)	Basic	0.5383	0.0987		0.5146	0.5174	0.5141
	Cauch	0.5401	0.2726	1.1487	N/A	N/A	N/A

Table 3.4: Parameter estimates for the Tecumseh and Seattle data using the unrestricted model

	Parameter estimate						
	π	$\lambda_L^{(2)}$	$\lambda_L^{(3)}$	$\lambda_L^{(4)}$	$\lambda_L^{(5)}$	$\lambda_L^{(6)}$	$\lambda_L^{(7)}$
Tec. 1977-78	0.8543	0.0431	0.0553	0.0362	0.0258	0.0212	0.0208
Tec. 1980-81	0.8797	0.1261	0.0715	0.0425	0.0330		
Tec. comb.	0.8702	0.0850	0.0620	0.0398	0.0299	0.0080	0.0225
Seat. 1975-76	0.8324	0.0491	0.0391	0.0366	0.0203		
Seat. 1978-79	0.5401	0.1230	0.0772				

not offer any advantage over the Cauchemez model, to the extent that such consistently high p-values suggest that the Cauchemez model behaves almost identically to the unrestricted model. Also, it is only when the Tecumseh 1980-81 data are included that there is any case for rejecting the basic model. Note also that although the three hypothesis tests appear to offer broad agreement (as one would hope), it seems that there can be some difference between the results of these tests, especially for lower p-values. In particular, the p-value of 0.0027 for the pseudoscore test is rather lower than those of the other two tests, under which one would be unlikely to reject the basic model.

The estimate $\eta = 1.5203$ under the Cauchemez model for the Tecumseh 1980-81 data suggests an abnormally high level of dependence on household size for the local contact rate in this outbreak. This explains the lower p-values for the testing the basic model against the unrestricted model for these and the combined Tecumseh data sets. It would be foolish however to dismiss the Tecumseh 1980-81 data as anomalous with such a small number of other data sets to compare to. Even if the Tecumseh 1980-81 data are unusual, it may still be perfectly reasonable for the local contact rate to have different levels of dependency of household size for the same strain of influenza if other conditions change. The manner in which households were recruited into the Tecumseh 1980-81 data (some households dropped out of the study between outbreaks and had to be replaced) may also provide a reason for this apparent change. Whatever the explanation, the results of Table 3.3 suggest that combining data from the two Tecumseh outbreaks to form a single data set should be done with caution when modelling these influenza outbreaks using a households SIR model.

A key strength of the Cauchemez model is its robustness to such changes, as evidenced by the extremely high p-values under all three tests comparing the Cauchemez model to the unrestricted model for all of the available data. The Cauchemez model also maintains a large amount of the simplicity of the basic model, in that it only introduces one extra parameter no matter how many different households sizes there are in the data. Both the basic and Cauchemez models benefit from this simplicity and do not suffer from having highly unreliable parameter estimates which can occur in the unrestricted model if there are very few households of a given size. Specifically, no estimator was available for θ_7 of the Tecumseh 1980-81 data since no individuals in size-7 households were infected and $\hat{\theta}_6 = 0$ for this data set since no individuals were infected by local contact in size-6 households. Such issues can be averted by simply ignoring data from households whose size is rare in the data, as was done here, but this is unnecessary under the basic and Cauchemez models. Thus, for these data, it appears that there is generally little to gain from using the unrestricted model for influenza, since the basic and Cauchemez alternatives provide far greater simplicity (both mathematically and in terms of considering the reliability of parameter estimates when minimal data are available for certain household sizes) without significantly reducing the goodness-of-fit to the data.

3.6.2 Testing the basic model vs the Cauchemez model

We now look to test the basic model against the Cauchemez model for our influenza data. That is to say that we let $\boldsymbol{\theta} = (\pi, \lambda_L, \eta)$ and wish to test

$$\begin{aligned} H_0 &: \ \eta = 0 \\ \text{vs } H_1 &: \ \eta \neq 0. \end{aligned} \tag{3.6.3}$$

Under this new definition of $\boldsymbol{\theta}$ and new hypotheses from (3.6.3) we now have $\boldsymbol{h}(\boldsymbol{\theta}) = \theta_3$ and $\boldsymbol{H_\theta} = (0, 0, 1)^\top$. It is straightforward to see that all of the theory of Sections 3.3 and 3.5 still holds, since our null hypothesis still places no restrictions on $\theta_1 = \pi$, but some amendment is needed to the calculations of $\boldsymbol{I_\theta}$ and $\boldsymbol{\Sigma_\theta}$ given in Section 3.4. Observe that the only changes required are to equations such as (3.4.4), which are obtained by taking derivatives with respect to $\lambda_L^{(n)}$ $(n = 2, 3, .., n_{max})$ under the unrestricted model. Under the Cauchemez model, we now require derivatives with respect to λ_L and η. However, since $\lambda_L^{(n)} = \lambda_L / n^\eta$, we can make use of the chain rule to acquire these derivatives. Specifically, for $n = 2, 3, ..., n_{max}$ and any given function f,

$$\begin{aligned} \frac{\partial}{\partial \lambda_L} f(\lambda_L^{(n)}) &= n^{-\eta} f'(\lambda_L^{(n)}) \quad \text{and} \\ \frac{\partial}{\partial \eta} f(\lambda_L^{(n)}) &= -\lambda_L \log{(n)} n^{-\eta} f'(\lambda_L^{(n)}), \end{aligned}$$

where f' denotes the first derivative of f with respect to $\lambda_L^{(n)}$. Calculation of $\boldsymbol{I_\theta}$ and $\boldsymbol{\Sigma_\theta}$ now follows easily using the methods of Section 3.4.

Since $\boldsymbol{H_\theta}$ now has rank 1, all three of our tests now involve comparison to a χ_1^2 distribution and thus we reject H_0 at the 95% significance level if the relevant

Table 3.5: p-values from testing the basic model against the Cauchemez model using all three tests for the influenza data

Epidemic	Basic vs Cauchemez test p-value		
	LRT	Wald	Score
Tecumseh 1977-78	0.2569	0.2473	0.2817
Tecumseh 1980-81	0.0058	0.0038	0.0165
Tecumseh combined	0.0059	0.0032	0.0158
Seattle 1975-76	0.5488	0.5525	0.5679
Seattle 1978-79	0.5146	0.5116	0.5191

test statistic exceeds 3.8415. Table 3.5 gives the p-values for each of the tests on our new hypotheses for the influenza data. Again there is good agreement between the pseudolikelihood ratio test, pseudo-Wald's test and pseudoscore test for each data set and this agreement is clearly far stronger than for tests relating to the unrestricted model, although there is a general trend for the pseudo-Wald's test to give the lowest p-value and the pseudoscore test to give the highest p-value. The pseudoscore test also relies on the likelihood function $L(\boldsymbol{\theta})$ having a derivative close to 0 at the MpLE under the null hypothesis. As such it is particularly prone to erroneous results when parameter values are close to their boundary (see the discussion at the end of Section 3.6.1).

As with the hypothesis tests against the unrestricted model, Table 3.5 displays far smaller p-values when the Tecumseh 1980-81 data are included. However, unlike before, there is now clear evidence to reject the basic model in favour of the alternative model for both the Tecumseh 1980-81 data and the combined data. This is particularly evident under the pseudolikelihood ratio and pseudo-Wald's test which both give p-values less than 0.01. The p-values for the remaining data do not fall near any realistic rejection region, although this may be due to a lack of data or, in the case of the Seattle outbreaks, not having data for a wide enough variety of household sizes. (Note that we are now able to use the full Tecumseh 1980-81 data set, including households of sizes 6 and 7, following the discussion at the end of Section 3.6.1. This explains the difference in estimates of η for the Tecumseh 1980-81 data between Table 3.3 and Table 3.6 given below.)

Table 3.6 gives 95% confidence intervals for η, as obtained under the Cauchemez

Table 3.6: 95% confidence intervals for η from the influenza data

Epidemic	MpLE of η (95% confidence interval)
Tecumseh 1977-78	0.8050 (-0.5587, 2.1687)
Tecumseh 1980-81	1.7542 (0.5674, 2.9411)
Tecumseh combined	1.3072 (0.4382, 2.1761)
Seattle 1975-76	0.8080 (-1.8581, 3.4742)
Seattle 1978-79	1.1482 (-2.2804, 4.5769)

model. These intervals were calculated using the asymptotic distribution of MpLEs given by (3.2.6). The confidence intervals are generally wide, particularly for the Seattle data, which confirms the suggestion that more data, on a greater number of household sizes is needed to determine if the basic model should generally be rejected in favour of the Cauchemez model for influenza. However, our MpLEs for η are consistently closer to 1 than 0. This corresponds with the estimate of $\eta = 0.84$ given in Cauchemez et al. [2004] for influenza data from Epigrippe in France and there is some evidence to suggest that use of the Cauchemez model with $\eta \approx 1$ as an alternative to the basic model should be investigated further.

3.6.3 Goodness of fit

In this section we have considered which of our three models (basic, Cauchemez or unrestricted) provide a "best fit" to our influenza data using hypothesis testing. However, we have not considered how well any of our models fit the data in the wider sense and not just in comparison to each other. This may be achieved by using the usual Pearson chi-squared goodness-of-fit statistic and following the methods of Ball and Lyne [2016].

Let $N_* = \{n \in \{1, 2, ..., n_{max}\} \ : \ \sum_{i=1}^{m_n} \delta_{n,i} \geq 1\}$ denote the set of household sizes for which we have observed data from a given epidemic. For $n \in N_*$ and $0 \leq k \leq n$ let $O_{n,k}$ and $E_{n,k}(\hat{\boldsymbol{\theta}}) = (\sum_{i=1}^{m_n} \delta_{n,i}) P_n(k|\hat{\boldsymbol{\theta}})$ be the observed and expected number of households of size n respectively in which k individuals are ultimately infected by the epidemic. Let

$$X^2 = \sum_{n \in N_*} \sum_{k=0}^{n} \left(O_{n,k} - E_{n,k}(\hat{\boldsymbol{\theta}})\right)^2 / E_{n,k}(\hat{\boldsymbol{\theta}}).$$

For the sequence of epidemics $E^{(\nu)}$ described in Section 3.1, Ball and Lyne [2016] show that in our single-type epidemic setting

$$X^2_{(v)} \xrightarrow{D} \chi^2_{n^*}$$

where $X^2_{(v)}$ is defined in the obvious manner for $E^{(\nu)}$ and n^* is the degrees of freedom for the standard Pearson test assuming independent households.

Table 3.7 shows the results of this goodness-of fit test applied to each of our models when applied to the Tecumseh and Seattle influenza data sets. In each case we provide the number of degrees of freedom, n^* for the Pearson chi-squared test, the test statistic X^2 and the p-value for the goodness-of-fit test under the null hypothesis that the observed data is drawn from a distribution described by our model. Since the number of degrees of freedom for a Person chi-squared is equal to the number of categories in our data minus the number of parameters in our model, we find that $n^* = n_{max}(n_{max} + 3)/2 - p$, where p is the number of model parameters and $p = 2, 3$ and n_{max} for the basic, Cauchemez and unrestricted models respectively. Note that we once again use a restricted version of the Tecumseh 1980-81 data set, ignoring households of size 6 and 7 for the reasons outlined in Section 3.6.1.

We observe from the high p-values in Table 3.7 that each of our three models appear to provide a reasonable fit to each of the real data sets used within this chapter. Therefore, selecting one of these models to apply to each of our data sets is sensible and hence the work in this section is a useful application of the hypothesis testing method for model selection that is presented in this chapter.

3.7 Discussion

We have derived a central limit theorem for final size data under the households epidemic model outlined in Chapter 2 based on the theorem derived by Ball and Lyne [2002a]. This central limit theorem was used to present a general theory for performing three types of hypothesis test (pseudoLRT, pseudo-Wald's and pseudoscore) based on maximum pseudolikelihood estimates of epidemic parameters. In particular, we have focussed upon hypotheses concerning local contact parameters $\lambda_L^{(n)}$ $(n = 2, 3, ..., n_{max})$, giving specific calculations of covariance matrices for this model and showing that hypothesis tests that only

Table 3.7: Goodness-of-fit test results on the Tecumseh and Seattle data for the basic, Cauchemez and unrestricted models for local infectiousness. DoF refers to the degrees of freedom of the Pearson chi-squared test.

	Model	DoF n^*	Test statistic X^2	p-value
Tecumseh 1977-78	Basic	33	30.3109	0.6017
	Cauchemez	32	27.3596	0.7006
	Unrestricted	28	26.7493	0.5319
Tecumseh 1980-81	Basic	18	15.3074	0.6408
	Cauchemez	17	10.4738	0.8826
	Unrestricted	15	10.2511	0.8037
Tecumseh combined	Basic	33	21.7988	0.9320
	Cauchemez	32	14.1415	0.9973
	Unrestricted	28	13.4078	0.9909
Seattle 1975-76	Basic	18	8.2309	0.9750
	Cauchemez	17	8.1850	0.9624
	Unrestricted	15	8.0386	0.9222
Seattle 1978-79	Basic	7	2.0988	0.9542
	Cauchemez	6	1.6582	0.9483
	Unrestricted	6	1.6582	0.9483

consider local contact parameters do not require knowledge of what proportion of the population the observed data represents (although a knowledge of the population structure $\boldsymbol{\alpha}$ is required).

The effects of including a maximum household size n_{max} in our epidemic model have also been considered. In Section 3.1 we note that, with a minor change to condition (ii) of Lemma 3.1.1, the central limit theorem derived in this chapter still holds if a maximum household size is not imposed. However, the discussion at the end of Chapter 2 and surrounding the use of Gontcharoff polynomials in Section 3.4 point out that the numerical methods needed to compute theoretic final size probabilities and the covariance matrices needed to apply our central limit theorem become intensive and potentially unstable if n_{max} is too large. As such, imposing a maximum household size on our epidemic model is sensible until such time as these numerical methods can be improved upon.

Previously studied influenza data sets from Seattle, Washington and Tecumseh, Michigan were used to illustrate the theory of this chapter. Hypothesis tests were performed on the data in an attempt to decide between three nested household epidemic models. These were the basic model, in which local contact rates are independent of household size, the Cauchemez model, in which local contact rates depend on household size in a set manner according to a parameter η; and the unrestricted model, defined in Chapter 2 in which local contact rates depend on household size in an arbitrary way. The hypothesis tests showed no evidence that the unrestricted model was superior to the basic or Cauchemez models for these data but was less conclusive in determining whether the basic model should be rejected in favour of the Cauchemez model. Of particular interest is a specific case of the Cauchemez model in which $\eta = 1$ (which is as simple as the basic model in terms of the number of parameters that need to be estimated). Wide confidence intervals for η, derived using the central limit theorem of Ball and Lyne [2001], suggested that further investigation into whether η may take a specific value for influenza epidemics may provide a fruitful area for future research.

We have also established that our three models all provide a reasonable to observed data from Seattle and Tecumseh using a standard goodness-of-fit test which is applicable due to results in Ball and Lyne [2016]. Therefore, deciding between these particular models for fitting to the observed data is a sensible approach to statistical inference. However, one issue that has not been discussed is whether the hypothesis testing approach outlined in this chapter is the best tool available for model selection. The theory presented in this chapter shows that our hypothesis testing approach is perfectly valid but other methods should also be considered.

Information-theoretic criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are perhaps the most popular tools for model selection within the statistical community. Another model selection tool that is becoming increasing popular is cross-validation. However, these methods all rely on having independent data or else their asymptotic properties are unknown. Thus it would seem unwise to use any of these tools for model selection with households epidemic data unless further research establishes their asymptotic properties for dependent data, augmented versions of these methods can be found for such data or the number of households, m, in the population sampled for a given data set is known to be large enough that dependence between outcomes in different households is extremely weak. As such, the hypothesis testing methods developed in this chapter appear to be the best available model selection method for households epidemic data at present.

The real data presented in this chapter are both relatively small and are only for influenza outbreaks and as such, our general unrestricted model cannot be dismissed altogether. Thus, for the remainder of this thesis we continue to present theory in terms of the unrestricted model but will use the basic and Cauchemez models (with emphasis on the $\eta = 1$ Cauchemez model) when providing applications and illustrations.

CHAPTER 4

Estimating within-household infection rates in emerging epidemics

Thus far we have focussed upon inference from completed epidemics. We now look to estimate the parameters of outbreaks which are still in their initial stages, specifically, the time in which an epidemic replicates the branching process set out in Section 2.2. Previous literature on emerging epidemics has largely focused on the exponential growth rate and this is defined and reviewed in Section 4.1. Our key focus for this chapter however is on the local dynamics in the early stages of an outbreak. Section 4.2 suggests an intuitive method for estimating the local contact parameters of an emerging epidemic but establishes that the resulting estimators are biased. An asymptotically unbiased estimator of local contact rates for our epidemic model is derived in Section 4.3 by utilising branching process theory and is then adapted to the discrete-time Reed-Frost model in Section 4.4. In Section 4.5 we illustrate how the new estimator may be used in practice and to assess factors affecting the bias of the intuitive estimators outlined in Section 4.2. A proof of the strong consistency of estimators using the new method is outlined in Section 4.6. The chapter closes with a brief discussion in Section 4.7.

This chapter is based upon the paper of Ball and Shaw [2015]. Permission has been obtained from the publisher to reproduce this work, in particular the figures, within this thesis.

4.1 Review of emerging epidemics and their growth rate

In Section 2.2 we introduced the threshold parameter R_* which determines the expected number of fully susceptible households that become newly infected as the result of a typical single-household epidemic, with one initial infective, in the early stages of a global outbreak. Thus R_* gives the rate at which infected households multiply on a generation by generation basis, where initially infected households represent the 0^{th} generation and any household newly infected by global contact from an individual in a k^{th} generation household belongs to the $k+1^{th}$ generation $(k = 0, 1, ...)$. We have seen in Chapter 2 that R_* provides useful information as to whether there is a positive probability of an epidemic taking off. However, since generations of infected households overlap in time, R_* does not correspond to any observable growth rate. Therefore, it is useful to consider the rate at which the number of infected households increases in real time during the early stages of a global epidemic.

Diekmann and Heesterbeek [2000] p.9 note that for a deterministic epidemic model with a homogeneously mixing population (i.e. $\lambda_L^{(n)} = 0$ for all n), incidence of disease increases at an exponential rate, r, in the early stages of an epidemic. That is to say that if $Y(t)$ is the number of individuals that have been infected up to time t, then

$$Y(t) \approx Ke^{-rt}$$

for some constant K. They also conclude that $r > 0$ if and only if $R_0 > 1$. Recall from Chapter 1 that R_0 is an individual reproduction number obtained by treating the proliferation of infected individuals in the initial stages of an epidemic in a homogeneously mixing population as a branching process. Diekmann and Heesterbeek [2000] p.103 extend this point to deterministic models with heterogeneity and show that an exponential growth rate r still exists in such a population. Thus it is natural to ask whether the proliferation of infected households under our stochastic households model also increases at an exponential rate, r, in the early part of a global outbreak and if it is possible to calculate r given the parameters of an epidemic.

In Section 2.2, basic theory of discrete-time branching processes (specifically the

Galton-Watson process) is exploited to yield the households reproduction number R_*. Given that we are now interested in real time dynamics of an epidemic, we now need to consider the theory of branching processes in continuous time and hence we turn our attention towards the Crump-Mode-Jagers branching process (CMJBP), see Jagers [1975] p.123 and in particular, results relating to the generalised CMJBP given by Nerman [1981]. Individuals in the generalised CMJBP are associated with a random variable denoting their life length and a point process denoting their reproduction times. We can use a generalised CMJBP to approximate an epidemic among households considering infected households in the epidemic as alive individuals in a CMJBP. Crucially, CMJBPs are associated with a Malthusian parameter, r, (Jagers [1975] p.132) which gives the rate at which the process grows exponentially. Further details on the approximating CMJBP for a households epidemic are given in Section 4.3 which also details the relationship between r and the infectious rate parameters of an epidemic. Pellis et al. [2011] have previously shown how r can be calculated using the other parameters of a households epidemic however their formula is only practical in the Markovian case in which infectious periods are exponentially distributed. We encounter similar issues in this chapter which are discussed in Section 4.3.2.

Before moving on to use theory associated with CMJBPs in order to understand the real time dynamics of epidemics among households, one should ascertain whether the theoretic exponential growth rate, r, discussed above may bear any resemblance to data which may be observed in real life. Figure 4.1 shows the number of households infected over time in a single simulation of a global outbreak among 1 million households of size 4. The infectious period was chosen to be exponentially distributed, the infectious parameters were $\lambda_G = 1$ and $\lambda_L^{(4)} = 1$ and the epidemic was initiated by a single individual chosen uniformly at random. A large population was used to ensure that the epidemic approximately mimicked a branching process for a reasonable period of time. The left hand plot appears to show an exponential growth during the time $t < 13$, after which the epidemic is no longer in its initial stages and thus grows at an increasingly slower rate prior to termination.

The right hand plot shows the number of infected households on a logarithmic scale and is significantly more informative since it appears to show a "burn in" period up to time $t = 5$, during which the epidemic becomes established. This is followed by linear growth (on the logarithmic scale) up to approximately time $t = 13$ before the epidemic growth slows and eventually terminates. Specifically, the right hand plot shows that the number of infected households appears to grow exponentially from approximately e^4 at time $t = 5$ to e^{10} at time $t = 10$, suggesting an exponential growth rate of $r \approx 6/5 = 1.2$. The theoretic exponential growth rate for this epidemic (calculated using the formula of Pellis et al. [2011]) is $r = 1.2095$.

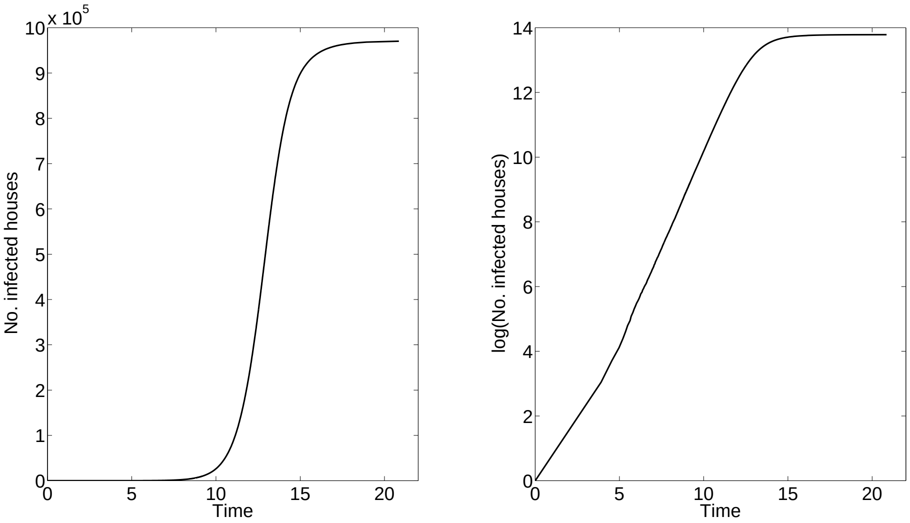

Figure 4.1: Plots showing the number of infected households over time in a single global epidemic among 1 million households. The right hand plot displays the same information as the left hand plot but is on a logarithmic for ease of observing exponential growth

Note from the above that an estimate of r should be one of the most readily available pieces of information from an emerging epidemic (see also Riley et al. [2003]). Pellis *et al.* also prove that the proliferation of infected households and individuals occurs at the same exponential rate under the stochastic households model. A final and important point from the formula of Pellis *et al.* for the calculation of r is that it indicates that there is a one-to-one correspondence between r and λ_G. Hence, if an estimate of r is available for an epidemic and the distribution of T_I is known, only the local contact rates need to be estimated complete parameter estimation. Therefore, assuming that an estimate of r is available, establishing local contact rate estimators is our aim for the rest of this chapter.

4.2 Basic approach to estimating local contact rates

For $n = 2, 3, ..., n_{max}$, suppose one wishes to estimate $\lambda_L^{(n)}$ for an epidemic that is observed whilst it is still in its initial stages, as described in Section 4.1. For $x = 0, 1, ..., n-1$, let $p_{basic}^{(n)}(x|\lambda_L^{(n)})$ be the probability that a single-household epidemic (without global infection) in a household of size n, started by one initial infective, finishes with x susceptibles remaining. It is clear that $p_{basic}^{(n)}(x|\lambda_L^{(n)}) = P_{n,1}(n-x|\lambda_L^{(n)})$ and thus may be determined using the triangular system of equations given by (2.3.1) in Section 2.3.

Let $a_{x,y}^{(n)}$ be the number of households of size n containing x susceptibles and y infectives at the time when the epidemic is observed. By considering only the households in which the single-household epidemic has ceased (i.e. where $x < n$ and $y = 0$), one can attempt to estimate $\lambda_L^{(n)}$ by maximising the pseudolikelihood function

$$L_{basic}^{(n)}(\lambda_L^{(n)}|\boldsymbol{a}) = \prod_{x=0}^{n-1} p_{basic}^{(n)}(x|\lambda_L^{(n)})^{a_{x,0}^{(n)}}. \tag{4.2.1}$$

Recall that (4.2.1) is not a true likelihood function as it assumes independence between households. This method of estimation, which we call *basic MpLE*, is simple but does not use all of the information available since households in which infectives are still present are ignored. A similar approach using more of the information available is to use maximum pseudolikelihood estimation but with censoring on households in which there are still infectives remaining. For $n = 2, 3, ...n_{max}$ and $x = 0, 1, ..., n-1$, let $q_{basic}^{(n)}(x|\lambda_L^{(n)}) = \sum_{i=0}^{x} p_{basic}^{(n)}(i|\lambda_L^{(n)})$ be the probability that a household of size n has *at most* x survivors from a single household epidemic and let $b_x^{(n)} = \sum_{y=1}^{n-x} a_{x,y}^{(n)}$ be the number of observed households of size n containing at least one infective and exactly x susceptibles. Such households will have at most x survivors once the single-household epidemic is completed. We can now use what is referred to as the *censored MpLE* approach for estimating $\lambda_L^{(n)}$, with left-censoring for the number of survivors (i.e. right-censoring for the total size), by maximising

$$L_{censor}^{(n)}(\lambda_L^{(n)}|\boldsymbol{a}, \boldsymbol{b}) = \prod_{x=0}^{n-1} p_{basic}^{(n)}(x|\lambda_L^{(n)})^{a_{x,0}^{(n)}} q_{basic}^{(n)}(x|\lambda_L^{(n)})^{b_x^{(n)}}.$$

Figure 4.2 shows how well the basic and censored MpLE methods perform in practice. For these histograms, epidemics were simulated using the same population and parameters as those used for Figure 4.1, with estimates of $\lambda_L^{(4)}$ taking place after the 1000^{th} recovery has occurred. Any epidemic not reaching 1000 recoveries was considered not to have taken off and was ignored. Estimates of $\lambda_L^{(4)}$ were made for the first 1000 epidemics to reach the 1000 recovery milestone. As before, a large population was used to ensure that the simulated epidemics were still approximately mimicking a branching process at the time of estimation.

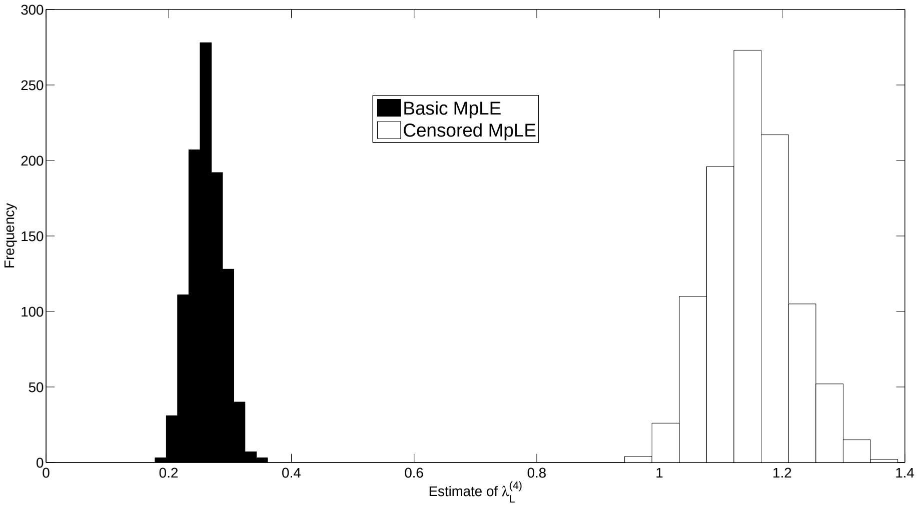

Figure 4.2: Estimates of $\lambda_L^{(4)}$, with a true value of 1, from 1000 epidemic simulations using the basic and censored MpLE methods

It is clear from Figure 4.2 that the basic MpLE method severely underestimates $\lambda_L^{(4)}$. This can be attributed to small local epidemics being more likely to have been completed at the time of estimation than larger local epidemics. Consequently, households that contain less severe local epidemics are more likely to be included in the basic MpLE estimate, causing the observed underestimate of $\lambda_L^{(4)}$. The censored MpLE approach appears to offer an improvement but repeated simulations with different parameters showed that this method generally overestimates $\lambda_L^{(4)}$, as is observed in Figure 4.2. Repeating the simulation for populations with different household sizes reveals the same trend. (The effect of household size on this observed bias is considered in more detail later in this chapter in the discussion surrounding Figure 4.9.)

In order to obtain a more accurate estimate of $\lambda_L^{(n)}$ $(n = 2, 3, ..., n_{max})$ one must understand the infected households branching process in more detail. The basic idea is the following. If the approximating branching process does not go extinct, then it grows exponentially at a rate r, which depends on the parameters of the households epidemic model, and as time $t \to \infty$ the fraction of completed single household epidemics (in the branching process), in households of size n, that leave x members susceptible, converges to a limit $\tilde{p}_{x,0}^{(n)}(r|\lambda_L^{(n)})$ $(x = 0, 1, ..., n-1)$. Thus we assume that each observed household in the data has final size that comes from that distribution and estimate $\lambda_L^{(n)}$ by maximising the pseudolikelihood obtained by replacing $p_{basic}^{(n)}(x|\lambda_L^{(n)})$ by $\tilde{p}_{x,0}^{(n)}(\hat{r}|\lambda_L^{(n)})$ in (4.2.1), where $\hat{r}$ is an estimate of the growth rate r; see (4.3.5) in the Section 4.3, where calculation of $\tilde{p}_{x,0}^{(n)}(r|\lambda_L^{(n)})$ is explained.

4.3 A new method

4.3.1 A more accurate estimator

We begin by formalising the approximation of a households epidemic to a CMJBP, as suggested in Section 4.1. Consider the approximating branching process introduced in Section 2.2, in which individuals correspond to infected households and an individual has one offspring whenever a global contact emanates from the corresponding single-household epidemic. For $n = 1, 2, ..., n_{max}$, let $E_H^{(n)}$ denote a typical size-n single-household epidemic, started by one member of the household being infected at time $t = 0$. For $t \geq 0$, let $X_H^{(n)}(t)$ and $Y_H^{(n)}(t)$ be respectively the numbers of susceptibles and infectives in $E_H^{(n)}$ at time t. Let $\mathcal{T}^{(n)} = \{(x, y) : \ x = 0, 1, ..., n-1; \ y = 0, 1, ..., n-x\}$ and, for $(x, y) \in \mathcal{T}^{(n)}$, let $p_{x,y}^{(n)}(t|\lambda_L^{(n)}) = \mathbb{P}(X_H^{(n)}(t) = x, \ Y_H^{(n)}(t) = y)$ $(t \geq 0)$ and $\tilde{p}_{x,y}^{(n)}(r|\lambda_L^{(n)}) = \int_0^\infty e^{-rt} p_{x,y}^{(n)}(t|\lambda_L^{(n)}) \ \mathrm{d}t$ $(r \geq 0)$. Note that $\mathcal{T}^{(n)}$ covers all possibilities for the numbers of susceptibles and infectives in a household once it has become infected (including its state at the end of the single-household epidemic).

Further, let $\xi_H^{(n)}$ be the point process describing times that global contacts emanate from $E_H^{(n)}$, so, for $t \geq 0$, $\xi_H^{(n)}([0, t])$ is the number of global contacts that emanate from $E_H^{(n)}$ during $[0, t]$. For $t \geq 0$ let $\mu^{(n)}(t) = \mathbb{E}[\xi_H^{(n)}([0, t])]$ and note that

$$\mu^{(n)}(\mathrm{d}t) = \lambda_G \sum_{(x,y)\in\mathcal{T}^{(n)}} y p_{x,y}^{(n)}(t|\lambda_L^{(n)})\ \mathrm{d}t. \tag{4.3.1}$$

Let ξ_H be a mixture of $\xi_H^{(1)}, \xi_H^{(2)}, ..., \xi_H^{(n_{max})}$ with mixing probabilities $\tilde{\alpha}_1, \tilde{\alpha}_2, ...,$ $\tilde{\alpha}_{n_{max}}$. (For $n = 1, 2, ..., n_{max}$, $\tilde{\alpha}_n$ is the probability of a global contact being with an individual in a size-n household in the early stages of an epidemic.) Then ξ_H is a point process which describes the ages at which a typical individual reproduces in the approximating branching process. This branching process is a general CMJBP since we have a point process associated with the reproduction of infected households and we let the life length of an infected household be infinite. Thus we have a random variable denoting life length that is actually constant. It is convenient to assume that individuals live forever in the branching process, though of course an individual ceases to reproduce as soon as there is no infective in the corresponding single-household epidemic. (Recall that individuals in the branching process are equated to infected households in the epidemic.)

Thus we may now exploit CMJBP theory. For $t \geq 0$, let

$$\mu(t) = \mathbb{E}[\xi_H([0,t])] = \sum_{n=1}^{n_{max}} \tilde{\alpha}_n \mu^{(n)}(t). \tag{4.3.2}$$

The branching process has a Malthusian parameter, $r \in (0, \infty)$, given by the unique solution of the equation

$$\int_0^\infty e^{-rt}\mu(\mathrm{d}t) = 1.$$

(See, for example, Jagers [1975] p.132.) Note, from (4.3.1) and (4.3.2), that r satisfies

$$\lambda_G \sum_{n=1}^{n_{max}} \tilde{\alpha}_n \sum_{(x,y)\in\mathcal{T}^{(n)}} y\tilde{p}_{x,y}^{(n)}(r|\lambda_L^{(n)}) = 1. \tag{4.3.3}$$

For $n = 1, 2, ..., n_{max}$ and $(x, y) \in \mathcal{T}^{(n)}$, an individual in the branching process is said to be in state (n, x, y) if it corresponds to a single size-n household epidemic and there are x susceptibles and y infectives in that epidemic. Let $\mathcal{T} = \{(n, x, y) : n = 1, 2, ..., n_{max} \text{ and } (x, y) \in \mathcal{T}^{(n)}\}$. For $t \geq 0$ and $(n, x, y) \in \mathcal{T}$, let $Y_{n,x,y}(t)$ be the number of individuals in state (n, x, y) at time t in the branching process.

Suppose that the Malthusian parameter r is strictly positive. We wish to verify that the conditions of Theorem 5.4 of Nerman [1981] are satisfied. Condition 5.1 of Nerman sates that there exists on $[0,\infty)$, a non-increasing, bounded, positive integrable function g such that

$$\mathbb{E}\left[\sup_{t\in[0,\infty)}\frac{\xi(\infty)-\xi(t)}{g(t)}\right]<\infty.$$

This follows from the remark given afterwards which states that the condition is satisfied if there exists a non-increasing, positive integrable function g such that

$$\int_0^\infty \frac{1}{g(t)}e^{-rt}\mu(\mathrm{d}t)<\infty.$$

Following Nerman's suggestion, we let $g(t)=e^{-rt}$ to satisfy this remark since $\mu(\infty)$ if r is finite.

For some $(n,x,y)\in\mathcal{T}$ let $\varphi(t)$ denote the indicator function on whether a given household is in state (n,x,y) at time t. Condition 5.2 of Nerman [1981] states that there exists on $[0,\infty)$, a non-increasing, bounded, positive integrable function h such that

$$U=\sup_{t\in[0,\infty)}\left(\frac{e^{-rt}\varphi(t)}{h(t)}\right)$$

has finite expectation. Then setting $h(t)=e^{-rt}$ clearly satisfies Condition 5.2.

Thus, we may apply Theorem 5.4 of Nerman [1981]. Applied to our setting, this shows that there exists a random variable $W\geq 0$, where $W=0$ if and only if the branching process goes extinct, such that for all $(n,x,y)\in\mathcal{T}$,

$$e^{-rt}Y_{n,x,y}(t)\xrightarrow{\text{a.s.}}\tilde{\alpha}_n\tilde{p}^{(n)}_{x,y}(r|\lambda^{(n)}_L)W \quad \text{as } t\to\infty, \tag{4.3.4}$$

where $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence (i.e. convergence with probability 1).

Note that $\sum_{(x,y)\in\mathcal{T}^{(n)}}p^{(n)}_{x,y}(t|\lambda^{(n)}_L)=1$, so $\sum_{(x,y)\in\mathcal{T}^{(n)}}\tilde{p}^{(n)}_{x,y}(r|\lambda^{(n)}_L)=1/r$ $(n=1,2,...,n_{max})$. Thus, if the branching process does not go extinct, as $t\to\infty$ the proportion of individuals that are in state (n,x,y) converges almost surely to $\tilde{\alpha}_n r\tilde{p}^{(n)}_{x,y}(r|\lambda^{(n)}_L)$.

Return to the households epidemic model. Recall that for $(n,x,y)\in\mathcal{T}$, the number of households of size n that contain x susceptibles and y infectives when the epidemic is observed is denoted by $a^{(n)}_{x,y}$. Suppose that an estimate,

$\hat{r}$ say, of the growth rate r is available. Then, provided the epidemic has taken off and it has been running for a sufficiently short period of time so that the branching process provides a good approximation but a sufficiently long time so that the above asymptotic composition of the branching process is applicable, the $\lambda_L^{(n)}$ can be estimated by maximising the normalised pseudolikelihood function

$$L_{full}(\lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})} | \boldsymbol{a}, \hat{r}) = \prod_{n=2}^{n_{max}} \prod_{(x,y) \in \mathcal{T}^{(n)}} \tilde{p}_{x,y}^{(n)}(\hat{r}|\lambda_L^{(n)})^{a_{x,y}^{(n)}}. \tag{4.3.5}$$

In Section 4.6 we prove that, under suitable conditions, the estimator

$$(\hat{\lambda}_L^{(2)}, \hat{\lambda}_L^{(3)}, ..., \hat{\lambda}_L^{(n_{max})}) = \text{argmax } L_{full}(\lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})} | \boldsymbol{a}, \hat{r})$$

is strongly consistent as the number of households $m \to \infty$, i.e. that $\hat{\lambda}_L^{(n)}$ converges almost surely to the true value $\lambda_L^{(n)}$ as $m \to \infty$.

Suppose, as in the basic MpLE method, that estimation is based only on completed single-household epidemics. Then the $\lambda_L^{(n)}$ may be estimated by maximising

$$L_{final}(\lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})} | \boldsymbol{a}, \hat{r}) = \prod_{n=2}^{n_{max}} \prod_{x=0}^{n-1} \tilde{p}_{x,0}^{(n)}(\hat{r}|\lambda_L^{(n)})^{a_{x,0}^{(n)}}.$$

Observe that subject to mild conditions,

$$p_{basic}^{(n)}(x|\lambda_L^{(n)}) = \lim_{t \to \infty} p_{x,0}^{(n)}(t|\lambda_L^{(n)}) = \lim_{r \to 0^+} r\tilde{p}_{x,0}^{(n)}(r|\lambda_L^{(n)}).$$

(Since $e^{-rt} \to 1$ as $r \to 0$.) It follows that, under appropriate conditions, the basic MpLE method becomes asymptotically unbiased as the growth rate tends down to zero.

A key assumption of the estimator based on L_{full} is that the exact state of a household is observable but this is unlikely to be realised in practice. Suppose that only recoveries are observed. For $n = 2, 3, ..., n_{max}$ and $j = 1, 2, ..., n$ let $c_j^{(n)}$ be the observed number of households of size n with j recoveries, let $\mathcal{A}_j^{(n)} = \{(x, y) \in \mathcal{T}^{(n)} : \; x + y = n - j\}$ and let

$$\tilde{q}_j^{(n)}(r|\lambda_L^{(n)}) = \sum_{(x,y) \in \mathcal{A}_j^{(n)}} \tilde{p}_{x,y}^{(n)}(r|\lambda_L^{(n)}) / (\frac{1}{r} - \tilde{q}_0^{(n)}(r|\lambda_L^{(n)})),$$

where $\tilde{q}_0^{(n)}(r|\lambda_L^{(n)}) = \sum_{y=1}^{n} \tilde{p}_{n-y,y}^{(n)}(r|\lambda_L^{(n)})$. Then the $\lambda_L^{(n)}$ may be estimated by maximising

$$L_{rec}(\lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})}|\boldsymbol{c}, \hat{r}) = \prod_{n=2}^{n_{max}} \prod_{j=1}^{n} \tilde{q}_j^{(n)}(\hat{r}|\lambda_L^{(n)})^{c_j^{(n)}}. \qquad (4.3.6)$$

4.3.2 Practicalities and extensions

Estimates of $\lambda_L^{(n)}$ based upon L_{full} and L_{rec} are both dependent on knowing $\tilde{p}_{x,y}^{(n)}(r|\lambda_L^{(n)})$ for $(n, x, y) \in \mathcal{T}$, which is not practical in many circumstances. However, it is possible if we restrict ourselves to the Markovian case, in which the infectious period T_I is exponentially distributed, by following a similar argument to that used in Section 4 of Pellis et al. [2011] to calculate real-time growth rates. Under these circumstances, the single-household epidemic $E_H^{(n)} = \{(X_H^{(n)}(t), Y_H^{(n)}(t)) : t \geq 0\}$ is a continuous-time Markov chain (CTMC). Figure 4.3 shows the transition rates of $E_H^{(3)}$ as a CTMC and also assigns labels to each state $(x, y) \in \mathcal{T}^{(3)}$. The exact assignment of these state labels is unimportant, however it is convenient for the initial state $(n-1, 1)$ to be assigned as state 1 for a size-n household. Note that the state space $\mathcal{T}^{(n)}$ of $E_H^{(n)}$ has size $s^{(n)} = |\mathcal{T}^{(n)}| = n(n+3)/2$. Let $Q^{(n)}(\lambda_L^{(n)}) = [q_{ij}^{(n)}(\lambda_L^{(n)})]$ be the $s^{(n)} \times s^{(n)}$ transition-rate matrix of $E_H^{(n)}$, using the assigned labelling. Thus, if $i \neq j$ then $q_{ij}^{(n)}(\lambda_L)$ is the transition rate of $E_H^{(n)}$ from the state having label i to the state having label j, and $q_{ii}^{(n)}(\lambda_L^{(n)}) = -\sum_{j \neq i} q_{ij}^{(n)}(\lambda_L^{(n)})$. Note that if a label i corresponds

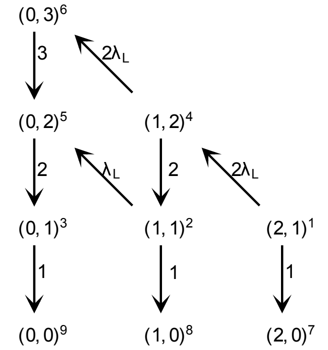

Figure 4.3: Graphical representation of a single-household epidemic for households of size 3 as a CTMC, where (x, y) denotes the household state and state labels (shown as superfixes) for the CTMC are assigned as described. The values on the arrows represent positive transition rates between states in the single-household epidemic

to a household state $(x, 0)$, then $q_{ij}^{(n)}(\lambda_L^{(n)}) = 0$ for all j. If k is the label assigned to state $(x, y) \in \mathcal{T}^{(n)}$ then $p_{x,y}^{(n)}(t|\lambda_L^{(n)}) = (e^{tQ^{(n)}(\lambda_L^{(n)})})_{1k}$, where $e^{tQ^{(n)}(\lambda_L^{(n)})} = \sum_{l=0}^{\infty}(tQ^{(n)}(\lambda_L^{(n)}))^l/l!$ denotes the usual matrix exponential. Hence,

$$\tilde{p}_{x,y}^{(n)}(r|\lambda_L^{(n)}) = \int_0^\infty e^{-rt}(e^{tQ^{(n)}(\lambda_L^{(n)})})_{1k}\,\mathrm{d}t = ([rI_{s^{(n)}} - Q^{(n)}(\lambda_L^{(n)})]^{-1})_{1k},$$

where $I_{s^{(n)}}$ is the $s^{(n)} \times s^{(n)}$ identity matrix (cf. Equation (12) on p.78 of Grimmett and Stirzaker [2001]).

The estimating procedure described in Section 4.3.1 assumes that the distribution of the infectious period is known. The theory may be extended easily to the setting where a parametric form is assumed for the infectious period distribution, with unknown parameters that need to be estimated from the data. E.g. if the infectious period is assumed to follow an exponential distribution with rate γ, then the preceding theory goes through with $p_{x,y}^{(n)}(t|\lambda_L^{(n)})$ replaced in an obvious fashion by $p_{x,y}^{(n)}(t|\lambda_L^{(n)}, \gamma)$ and $(\lambda_L^{(n)}, \gamma)$ being estimated by maximising the appropriate normalised pseudolikelihood function (which should be adjusted to include households of size 1 since their dynamics are affected by T_I). Note that for final outcome data it is impossible to estimate both the various $\lambda_L^{(n)}$ and γ, since the final outcome distribution is invariant to rescaling of time. However, that is not the case in an emerging epidemic setting, as the exponential growth rate is clearly time-scale dependent.

The assumption of exponentially distributed infectious periods can be relaxed by using the phase method (e.g. Asmussen [1987] p.71-78). For example, a J-stage Erlang distribution for the infectious period can be accommodated by splitting the infectious period into J stages having independent exponentially distributed durations. The Markov property is maintained by expanding the state space of a single-household epidemic to include the number of infectives in each of the J stages. This can lead to an appreciable increase in the size of $\mathcal{T}^{(n)}$. One can also extend the model to an SEIR (susceptible → exposed → infectious → recovered) model by introducing a latent period. In the simplest case, both infectious and latent periods follow exponential distributions, in which case the state space of a single-household epidemic is extended to include the number of exposed (i.e. latent) individuals, but again the phase method can be used to accommodate more general distributions.

A further extension would be to assume that infectious cases are observed with some (un)known, probability $\delta \in [0, 1]$ which is independent of all other processes in the epidemic and independent of whether previous infectious cases have been correctly observed. (Gamado et al. [2014], for example, consider a similar idea.) This could be easily incorporated into the model by replacing the $p_{x,y}^{(n)}(t|\lambda_L^{(n)})$ with $p_{x,y}^{(n)}(t|\lambda_L^{(n)}, \delta)$. Note that if $\delta = 0$ (whether known or unknown) only recovered individuals are observed and thus we are in the same situation as discussed at the end of Section 4.3.1.

Finally, note that the theory above has been generalised from that presented in Ball and Shaw [2015] in which the basic model of local contact is assumed ($\lambda_L^{(n)} = \lambda_L$ for all n). Alternatively, one can assume a specific form for $\lambda_L^{(n)}$, such as the Cauchemez model (see Section 3.3.1) under which one can estimate the unknown parameters λ_L and η in the obvious fashion.

4.4 Application to Reed-Frost Epidemics

We make a small diversion in this section to discuss how theory similar to the above can be applied to the discrete-time Reed-Frost epidemic model. The Reed-Frost model, in the context of a population of households, is briefly introduced before showing how a multitype branching process method can be used to estimate local person-to-person infectious probabilities. An adapted version of the method for Reed-Frost epidemics using continuous-time Markov processes can be used give an alternative unbiased estimator for our continuous-time epidemic model to that derived in Section 4.3 for epidemics with exponentially distributed infectious periods. This is briefly outlined at the end of this section.

The work presented in this section utilises theory presented in Ball and Shaw [2016] as well as that of Ball and Shaw [2015].

4.4.1 The Reed-Frost epidemic model

We consider a population structured in the same manner as described in Section 2.1, parameterised by α. Under the Reed-Frost model (see, for example, Abbey [1952]), susceptibles contacted by an infective experience a latent period of constant duration, which without loss of generality can be taken to be one unit of time, and the infectious period is reduced to a single point in time.

Consider an epidemic initiated by a small number of individuals being infected at time $t = 0$. For $t = 0, 1, \ldots$, individuals infected at time t become infectious at time $t + 1$. Different infectives behave independently of each other. Consider an individual in a household of size n that is infected at time t. At time $t + 1$ it makes global infectious contact with any given susceptible in the population with probability $p_G = 1 - \exp(-\mu_G/N)$ and, additionally and independently, local infectious contact with any given susceptible in its household with probability $p_L^{(n)}$. Moreover, contacts between this infectious individual and distinct susceptible individuals are mutually independent. Any susceptible individual that is contacted by at least one infective at time t is infected and becomes infectious at time $t + 1$. As in the continuous-time case, the process continues until there is no infective left in the population. For ease of notation let $\boldsymbol{p_L} = (p_L^{(2)}, p_L^{(3)}, ..., p_L^{(n_{max})})$ be a vector denoting the local contact probabilities for each household size.

Again, our focus is on emerging epidemics, so it is assumed that, when the epidemic is observed, the proliferation of infected households still mimics a discrete-time branching process. Note that in the limit as the population size $N \to \infty$, the mean number of global contacts made by a typical infective is μ_G. Note also that upon infection a household of size n is in state $(n, n-1, 1)$ and that in subsequent generations that household contains at least one recovered individual. We assume that it is possible to observe the geometric growth rate $\rho(\boldsymbol{p_L}, \mu_G)$ of the approximating branching process. (That being the rate at which the number of infectives multiplies with each generation. In the initial stages of an epidemic, generation $t + 1$ will have approximately $\rho(\boldsymbol{p_L}, \mu_G)$ times as many infectives as generation t.) The parameter μ_G increases with $\rho(\boldsymbol{p_L}, \mu_G)$ for fixed $\boldsymbol{p_L}$, so for any estimate of $\boldsymbol{p_L}$, an estimate for μ_G is predetermined since it is assumed that $\rho(\boldsymbol{p_L}, \mu_G)$ can be observed directly. Note that even though we have moved to a discrete-time setting, there is still the potential for an overlap between generations in the approximating branching process since households may contain infectives for more than one generation. Thus a threshold parameter R_* is still not readily available from real data and therefore it is sensible to consider methods utilising the observable geometric growth rate $\rho(\boldsymbol{p_L}, \mu_G)$. It is important to note however that $R_* > 1$ if and only if $\rho(\boldsymbol{p_L}, \mu_G) > 1$.

4.4.2 Estimating $\boldsymbol{p_L}$ using a multitype branching process approximation

A multitype branching process is process in which individuals take one of a number of possible forms and thus can assume different behaviours (see Athreya and Ney [1972] p.191). We consider a discrete-time multitype branching process S to approximate the early stages of a Reed-Frost epidemic. Define the type space of S as $\mathcal{T}_{RF} = \{(n, n-1, 1) : 1 \leq n \leq n_{max}\} \cup \bigcup_{n=1}^{n_{max}} \{(n, x, y) : x \geq 0, y \geq 1, x + y < n\}$ and label the elements of $\mathcal{T}_{RF}$ as $1, 2, ..., k$ where $k = |\mathcal{T}_{RF}| = n_{max} + \sum_{n=2}^{n_{max}} \frac{n(n-1)}{2} = n_{max}(n_{max}^2 + 5)/6$. Thus our "types" define the size of a household and the number of susceptibles and infectives present within it and the type space includes all possible household states where infection is still present.

Let $\boldsymbol{M}$ be the mean matrix of S on $\mathcal{T}_{RF}$, so the element m_{ij} is the expected number of type-j individuals that a typical type-i individual gives birth to upon death. Under the Reed-Frost model, a household in state (n, x, y) gives birth to an expected number of $\tilde{\alpha}_{n'}\mu_G$ households in state $(n', n' - 1, 1)$, for $n' = 1, 2, ..., n_{max}$, as a result of global infectious contacts, and to an expected number of $\binom{x}{z}(1 - (1 - p_L^{(n)})^y)^z(1 - p_L^{(n)})^{y(x-z)}$ households in state $(n, x - z, z)$, for $z = 0, 1, ..., x$, from local contacts. Let $\boldsymbol{Y}_t = (Y_{t1}, Y_{t2}, ..., Y_{tk})$ denote the number of individuals of each type from $\mathcal{T}_{RF}$ alive after t generations of S and let $\rho(\boldsymbol{p_L}, \mu_G)$ be the maximal eigenvalue of $\boldsymbol{M}$. Assume that $\rho(\boldsymbol{p_L}, \mu_G) > 1$, so the branching process is supercritical. Kesten and Stigum [1966] show that if $\boldsymbol{u}(\boldsymbol{p_L}, \mu_G)$ is the left-eigenvector associated with $\rho(\boldsymbol{p_L}, \mu_G)$, normalised so that its components sum to one, then

$$\rho(\boldsymbol{p_L}, \mu_G)^{-t}\boldsymbol{Y}_t \xrightarrow{\text{a.s.}} W\boldsymbol{u}(\boldsymbol{p_L}, \mu_G) \quad \text{as } t \to \infty, \tag{4.4.1}$$

where W is a non-negative random variable such that $W = 0$ if and only if S becomes extinct. The eigenvector $\boldsymbol{u}(\boldsymbol{p_L}, \mu_G)$ therefore gives the proportions of individuals of each type in S as $t \to \infty$, conditional upon S not going extinct. It follows from (4.4.1) that

$$\rho(\boldsymbol{p_L},\mu_G)^{-t}\sum_{t'=1}^{t}\boldsymbol{Y}_t' \xrightarrow{\text{a.s.}} \frac{\rho(\boldsymbol{p_L},\mu_G)}{\rho(\boldsymbol{p_L},\mu_G)-1}W\boldsymbol{u}(\boldsymbol{p_L},\mu_G) \quad \text{as} \quad t\to\infty. \tag{4.4.2}$$

Let $\boldsymbol{Z}_t = (Z_{t1}, Z_{t2}, ..., Z_{tk})$, where Z_{ti} denotes the number of single-household epidemics that terminate before t generations of the epidemic, for which the last active household state was $i \in \mathcal{T}_{RF}$. A household in state (n,x,y) at time t' has probability $(1-p_L^{(n)})^{xy}$ of containing no infectives at time $t'+1$. Hence, if (n,x,y) is the household state associated with a type-i individual in S, it follows from (4.4.2) and the strong law of large numbers that, for $i=1,2,...,k$,

$$\rho(\boldsymbol{p_L},\mu_G)^{-t}Z_{ti} \xrightarrow{a.s.} W\frac{(1-p_L^{(n)})^{xy}}{\rho(\boldsymbol{p_L},\mu_G)-1}u_i(\boldsymbol{p_L},\mu_G) \quad \text{as} \quad t\to\infty.$$

Let $u_{(n,x,y)} = u_i$ where i is the label of a type-(n,x,y) individual in S. Note that any single-household epidemic finishing the generation after it was in state (n,x,y) finishes with x susceptibles remaining. Thus, define the function $p_{RFfull}(n,x,y|p_L^{(n)},\mu_G)$ as follows:

$$p_{RFfull}(n,x,y|\boldsymbol{p_L},\mu_G) = \begin{cases} Ku_{(n,x,y)} & \text{if } y\geq 1, \\ K\displaystyle\sum_{y=1}^{n-x-1}(1-p_L^{(n)})^{xy}\frac{u_{(n,x,1)}(\boldsymbol{p_L},\mu_G)}{\rho(p_L,\mu_G)-1} & \text{if } y=0, \end{cases}$$

where K is chosen such that

$$\sum_{n=1}^{n_{max}}\left[\Big(\sum_{x=0}^{n-1}\sum_{y=0}^{n-x-1}p_{RFfull}(n,x,y|\boldsymbol{p_L},\mu_G)\Big)+\Big(p_{RFfull}(n,n-1,1|\boldsymbol{p_L},\mu_G)\Big)\right]=1.$$

One can then estimate $\boldsymbol{p_L}$ by performing maximum pseudolikelihood estimation in exactly the same manner as described using L_{full} in Section 4.3.1. Note that this estimation procedure can be adapted to the case where susceptibles and infectives are indistinguishable, using the same method as described for L_{rec} in Section 4.3.1.

4.4.3 An alternative estimator for the continuous-time model

The estimator derived for $\boldsymbol{p_L}$ above relies on the Markovian property of Reed-Frost epidemics. That is to say that at any given generation t, no knowledge is needed of previous generations to determine the probabilities relating to any infectious contact. In the continuous-time case, we have already alluded to the fact that an exponentially distributed infectious period T_I makes the epidemic Markovian (see Section 4.3.2). Suppose we have an epidemic in which T_I follows an exponential distribution with mean 1 (after rescaling of time). The approximating branching process for this epidemic may be described by a multitype birth-death (B-D) process, S_{BD}, a Markov process in which there are a number of individuals of each type, which may increase by one following a birth or decrease by one following a death.

The process S_{BD} is defined on active individual types (households in which infectives are present in our epidemic) and thus its type space is given by $\mathcal{T}_{RF}$. An individual in S_{BD} of type (n, x, y) has an exponentially distributed lifetime with rate $y(1 + x\lambda_L^{(n)})$, during which it gives birth to type-$(n, n-1, 1)$ individuals at rate $y\lambda_G$ as a result of infectives making global contacts with susceptibles in previously uninfected households. Upon death, a type-(n, x, y) individual produces a type-$(n, x-1, y+1)$ individual with probability $x\lambda_L^{(n)}/(x\lambda_L^{(n)} + 1)$. Otherwise it produces a type-$(n, x, y-1)$ individual if $y \geq 2$ or no individual if $y = 1$, since the recovery of the last remaining infective in a household causes a single-household epidemic to cease. Label the types in the manner described in Section 4.4.2 such that a type-$(n-i, 1)$ individual has label i.

Recall that $k = |\mathcal{T}_{RF}|$ and let $\mathbf{\Lambda}$ be the $k \times k$ birth-rate matrix of S, with element λ_{ij} being the rate at which a type-i individual gives birth to a type-j individual. Let $\text{diag}(\boldsymbol{\mu})$ be the diagonal death rate matrix of the process, with elements given by $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_s)$, where μ_i is the rate at which a type-i individual dies. (Thus if i corresponds to the state (n, x, y), $\mu_i = y(1 + x\lambda_L^{(n)})$, $\lambda_{ij} = y\tilde{\alpha}_n\lambda_G$ if j corresponds to the state $(n, n-1, 1)$, $\lambda_{ij} = xy\lambda_L^{(n)}$ if j corresponds to the state $(n, x-1, y+1)$, $\lambda_{ij} = y$ if j corresponds to the state $(n, x, y-1)$ and $\lambda_{ij} = 0$ for all other j.) For ease of notation, let $\boldsymbol{\theta} = (\lambda_G, \lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})})$. Let $r(\boldsymbol{\theta})$ be the maximal eigenvalue of $\boldsymbol{A} = \mathbf{\Lambda} - \text{diag}(\boldsymbol{\mu})$. Then $r(\boldsymbol{\theta})$ is the Malthusian parameter of S_{BD} and hence is also the real-time growth rate of the initial stages of the households epidemic. Let $\boldsymbol{v}(\boldsymbol{\theta}) = (v_1(\boldsymbol{\theta}), v_2(\boldsymbol{\theta}), ..., v_k(\boldsymbol{\theta}))$ be the left eigenvector of $\boldsymbol{A}$ associated with $r(\boldsymbol{\theta})$, normalised such that $\sum_{i=1}^k v_i(\boldsymbol{\theta}) = 1$.

Following Athreya and Ney [1972] p.206 yields a continuous-time equivalent of (4.4.1). Let $\tilde{Z}_{n,i}(t)$ denote the number of single households epidemics that terminate before time t in households of size n with i recoveries in the epidemic and define $Z_{n,i}(t)$ similarly for S_{BD}. All such households are in state $(n-i,1)$ immediately before the household epidemic ceases. Suppose i' is the label associated with state $(n-i,1)$ and recall that the recovery rate of infectives is 1. Then, for large t,

$$Z_{n,i}(t) \approx \int_0^t Y_{i'}(u)du \approx \int_0^t Wv_{i'}(\boldsymbol{\theta})e^{ur(\boldsymbol{\theta})}du = \frac{Wv_{i'}(\boldsymbol{\theta})}{r(\boldsymbol{\theta})}(e^{tr(\boldsymbol{\theta})}-1).$$

Moreover, see Jagers [1992]),

$$e^{-tr(\boldsymbol{\theta})}Z_{n,i}(t) \xrightarrow{a.s.} W\frac{v_{i'}(\boldsymbol{\theta})}{r(\boldsymbol{\theta})} \quad \text{as } t \to \infty.$$

Let $v_{n,x,y}(\boldsymbol{\theta}) = v_i(\boldsymbol{\theta})$ where i is the label of a type-(n,x,y) individual in S_{BD}. Then for, $(n,x,y) \in \mathcal{T}$, define

$$p^{(n)}_{multi}(x,y|\boldsymbol{\theta}) = \begin{cases} K(\boldsymbol{\theta})v_{(n,x,y)}(\boldsymbol{\theta}) & \text{if } y \geq 1 \\ K(\boldsymbol{\theta})\frac{v_{(n,x,1)}(\boldsymbol{\theta})}{r(\boldsymbol{\theta})} & \text{if } y = 0 \end{cases}$$

where $K(\boldsymbol{\theta})$ is chosen such that $\sum_{n=1}^{n_{max}}\sum_{x=0}^{n-1}\sum_{y=0}^{n-x} p^{(n)}_{multi}(\boldsymbol{\theta}) = 1$. Assuming the epidemic mimics the CMJBP outlined in Section 4.3 (as $t \to \infty$), it is clear that $p^{(n)}_{multi}(x,y|\boldsymbol{\theta})$ gives the asymptotic proportion of households in all possible single-household epidemic states. Thus estimation of local contact rates may now be carried out as described in Section 4.3 using this estimator.

4.5 Numerical illustrations

Applications of the preceding theory are presented in this section. For ease of illustration we revert to the *basic model*, as discussed in Section 3.6, in which $\lambda_L^{(n)} = \lambda_L$ for all n. Thus we need only estimate a single local contact parameter λ_L (or p_L under the Reed-Frost model), which may be achieved under the full-pseudolikelihood method by maximising the pseudolikelihood function

$$L_{full}(\lambda_L|\boldsymbol{a},\hat{r}) = \prod_{n=2}^{n_{max}} \prod_{(x,y)\in\mathcal{T}^{(n)}} \tilde{p}^{(n)}_{x,y}(\hat{r}|\lambda_L)^{a^{(n)}_{x,y}}.$$

Pseudolikelihood functions for the recovery-pseudolikelihood, basic MpLE and censored MpLE methods may be obtained in a similar fashion.

4.5.1 Simulation studies

We begin by performing a series of simulation studies. The parameter choices for our studies are loosely based on the Fraser [2007] analysis of varicella data. Simulations are performed on a population of $m = 10\,000$ households with size distribution $\boldsymbol{\alpha} = [0.13, 0.30, 0.23, 0.18, 0.09, 0.07]$, taken from 1961 UK census data (see Registrar General for England and Wales [1961]). Specifically, this population has a mean household size of 3.01 and thus our population has size $N = 30100$. This structure contains a higher proportion of larger households than those obtained using more recent censuses and is used to maximise the effect of local infectious contacts on the simulated epidemics. The population size is chosen so that it is small enough to represent a realistic population cluster (e.g. a town) but large enough so that there are sufficient data to estimate λ_L whilst the epidemic is still in its emerging phase. For the sake of simplicity, an exponentially distributed infectious period with rate 1 is used. Fraser suggests having a within-household susceptible-infectious escape probability of 0.39, as reported by Hope Simpson [1952], and that infected individuals be expected to infect 1.21 susceptibles outside of their household. This implies parameter values of $\lambda_G = 1.21$, $\lambda_L = 1.565$ (since $\phi(1.565) = 0.39$, where $\phi(\theta) = \mathbb{E}[\exp(-\theta T_I)] = (1+\theta)^{-1}$ and $r = 1.762$ (recall (4.3.3)) in the continuous-time case and $\mu_G = 1.21$, $p_L = 0.61$ $(= 1 - 0.39)$, $\rho(p_L, \mu_G) = 2.248$ under the Reed-Frost model.

Unless stated otherwise, growth rates are estimated by fitting a straight line to the logarithm of the number of recoveries, as a function of time, using the polyfit function in MATLAB. The first 20 recoveries are ignored when estimating r, to enable the exponential growing phase of the epidemic to settle in. Note that, while this is the most common method to estimate r, other methods are also considered in the literature; see, for example, Ma et al. [2014]. Further to this, King et al. [2015] show that this method showed severe bias when applied to real life Ebola data from West Africa in 2014. Thus we use this method to give an estimator of r purely out of convenience when dealing with simulated data and do not advocate using this method when analysing real life data.

For illustrative purposes, estimates of λ_L are given in terms of the secondary attack rate (SAR), as defined by Longini and Koopman [1982]. The SAR is the probability that an infective infects locally a given household member, expressed as a percentage, and is given by $100\%(1-\phi(\lambda_L))$. (Note that with the continuous-time and discrete-time models, matching the SAR and λ_G results in different growth rates.) The SAR is used since the variance of estimates of λ_L, under any of the methods outlined in this paper, increases greatly as the true value of λ_L increases, whereas the variance of the SAR estimates is closer to being constant whatever its true value. Note that for a given distribution of T_I, SAR strictly increases with λ_L.

It is shown in Section 4.3 that an emerging households epidemic can be approximated by a Crump-Mode-Jagers branching process (CMJBP), however there is no indication as to when an epidemic can still be considered to be in its emerging phase. Figure 4.4 shows estimates of the SAR throughout the lifetime of a single simulated SIR epidemic using the parameters outlined above. Estimations of λ_L (and hence of the SAR using the formula given above) were made at regular intervals throughout the epidemic using basic MpLE, censored MpLE, full- and recovery-pseudolikelihood estimation methods (using (4.3.5) and (4.3.6) respectively) and by considering the distribution of individuals at the *end* of an epidemic using the methods of Section 3.2. This is referred to as the *final-size* method of estimation.

For the basic MpLE method, it takes some time before the SAR is estimated to be non-zero. This can be explained by the reliance of this method on household epidemics being completed since the basic MpLE method will only pick up any trace of local infectivity when a completed single-household epidemic with more than one recovered individual is observed. As would be expected, the final-size method appears to tend to the true SAR value as $t \to \infty$. The initially large estimates from the final size data can be explained by noting that few households are infected at this time but that recoveries are clustered within households. The former point suggests a very low value of λ_G (considering that the estimator assumes that the epidemic is complete), so the estimate of the SAR is large to account for the clustering of recovered individuals.

Note that the recovery-pseudolikelihood method estimates the SAR to be 100% as the epidemic approaches completion. In the epidemic outlined above, with growth rate $r = 1.762$ but with an SAR of 100%, appreciably fewer than half of all infected households of size 3 and above are expected to contain only recovered individuals during the emerging phase. Once the true epidemic (with an SAR of 61%) is completed, appreciably more than 80% of households of size 3 and above in the entire population are expected to contain only recovered individuals. This suggests that there is a threshold, after the epidemic has stopped approximating a CMJBP, when the number of recovered individuals in infected households exceeds the expectations of even the maximum possible SAR in the recovery-pseudolikelihood estimation method, hence this method will continue to give an MpLE for the SAR as 100% for the remainder of the epidemic.

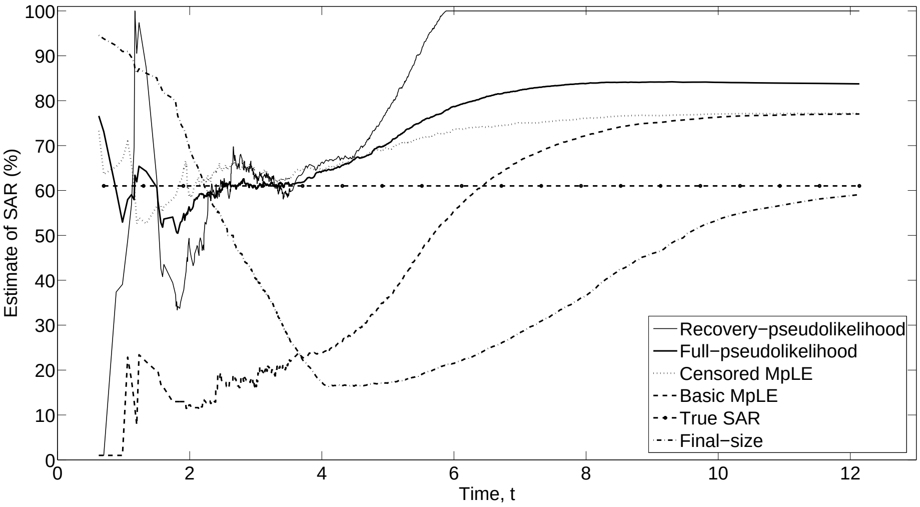

Figure 4.4: Estimates of the SAR (true value 61%) through time for a single SIR households epidemic. The four estimation methods outlined earlier in this chapter are shown along with estimates of the SAR using the final-size method

Figure 4.4 shows that once an epidemic has had sufficient time to establish itself, there is a window when both the full and recovery CMJBP methods appear to give a good estimate of the SAR. This corresponds to the time in Figure 4.1 when exponential growth is evident but is of a shorter length of time due to the smaller population size used in this simulation. Moreover, the length of this window in Figure 4.4 is roughly the same for both CMJBP methods, although the recovery method gives a less reliable estimate owing to it using less information. This is confirmed in Figure 4.5 which shows kernel density estimates of the distribution of the estimator of SAR for both CMJBP methods from 1000 simulations of the epidemic outlined above. The plots marked 'γ

known' use the methodology described in Section 4.3.1 and those marked 'γ unknown' assume that γ is also estimated, as described in Section 4.3.2. Estimates of the SAR were made from each simulation after 500 recoveries were observed for reasons outlined below. Irrespective of whether or not γ is also estimated, both the full and recovery methods yield estimates of the SAR that are centred broadly around the true value of 61% but the recovery method yields estimates having a far greater variance. The variance of the estimates is greater when γ is assumed unknown than when it is assumed known but the difference is appreciably smaller than that between the full and recovery methods. Kernel density estimates are used here for visual reasons as it allows us to view all four of our methods on the same plot. The inset of Figure 4.5 shows a scatter plot of the estimates of (SAR, γ) using the full-pseudolikelihood method, which indicates that the estimates of the SAR and γ are positively correlated.

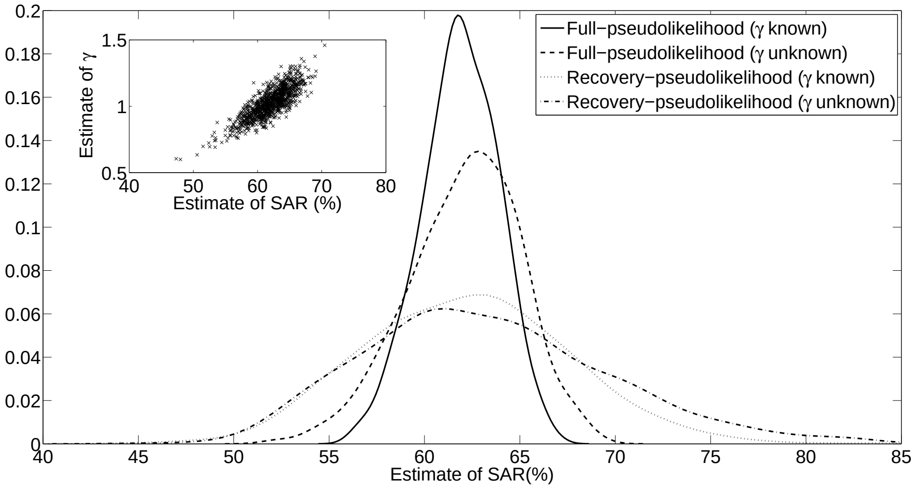

Figure 4.5: Kernel density estimates of the distribution of the estimator the SAR (true value 61%) based on 1000 simulations of the outlined epidemic using the full and recovery CMJBP estimation methods, both with and without the recovery rate γ (true value 1.00) being also estimated. Inset: Scatter plot of estimates of (SAR, γ) for the full-pseudolikelihood (γ unknown) method

Repeated simulations using different population sizes yielded very similar results to those seen in Figure 4.4, in that there appears to be a window once the epidemic has established itself when a households SIR epidemic can still be considered to be in its emerging phase and the full-pseudolikelihood estimate is relatively accurate. The start of this window corresponds to when the asymptotic behaviour of the approximating CMJBP kicks in, the timing of which is independent of the total population size N, provided N is sufficiently large. Further simulations suggested that this window ends when approximately $N^{2/3}$ recoveries have occurred, after which the CMJBP approximation of the households epidemic breaks down. The time taken for $N^{2/3}$ recoveries to take place depends on the severity of the epidemic and the population size. Note that Barbour and Utev [2004] prove that a homogeneously mixing Reed-Frost model can be closely approximated by a branching process up until order $N^{2/3}$ individuals have been infected.

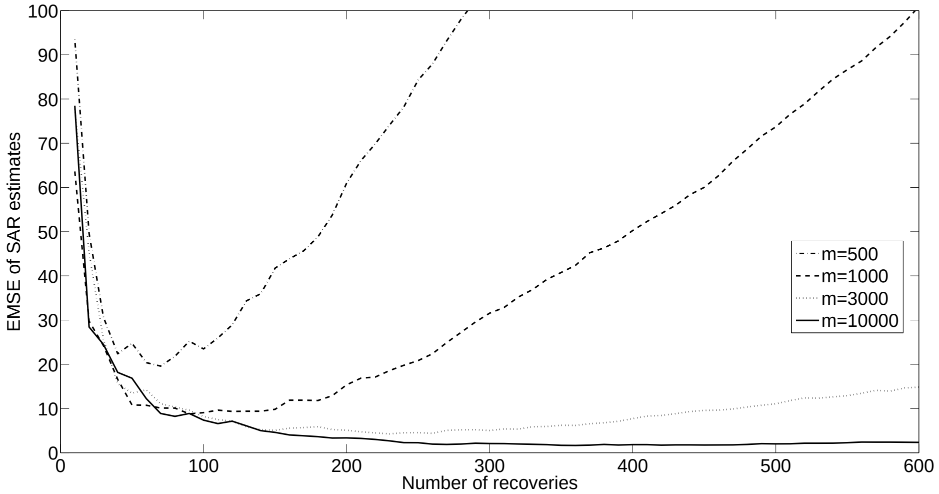

Figure 4.6: EMSE of estimates of the SAR using the full-pseudolikelihood method. See text for details

The above points are illustrated in Figure 4.6. This figure shows estimates of the mean squared error (EMSE) of estimates of the SAR, assuming that γ $(= 1)$ is known, using the full-pseudolikelihood method throughout the emerging stages of 1000 simulated epidemics and among populations with differing numbers of households. All simulated epidemics used in this figure are from a model with the same population structure $\boldsymbol{\alpha}$, growth-rate r and SAR as given above. If $SAR_1, SAR_2, ..., SAR_{1000}$ denote the estimates of the SAR (true value 61%) obtained from these 1000 simulated epidemics then EMSE $= 1000^{-1} \sum_{i=1}^{1000} (SAR_i - 61)^2$. It is assumed that the value r is known, in order that the figure illustrates only when the distribution of household states in an emerging epidemic conforms to its equivalent branching process. It can be seen that it takes approximately 50 recoveries to occur (regardless of population size) for the EMSE to stabilise and settle at a lower value due to the high variance of SAR estimates when too few households have been infected and the epidemic is yet to establish itself in the population. The length of this window then clearly increases with population size as a result of a higher percentage of fully susceptible households still being available at this stage of the epidemic. For the population considered in most of the numerical illustrations, i.e. consisting of 10 000 households, it appears appropriate to estimate the SAR after approximately 500 recoveries have occurred. This issue is discussed further in Section 4.5.3.

We now consider estimation of p_L in the Reed-Frost model. A single-household epidemic in a household of size n can last for at most n generations. Thus, under the assumption that all global contacts are with individuals in previously uninfected households, if the households epidemic is observed in the k^{th} generation, one can estimate p_L by using an adaptation of the basic MpLE method from the continuous time case as follows. If one wishes to make the estimate in the k^{th} generation then the single-household epidemics in all households with at least one recovery in the $(k - n_{max} + 1)^{th}$ generation are certain to have been completed. One can then estimate p_L by using only the latter households and considering the final-size distributions of single-household epidemics under the Reed-Frost model to perform the basic MpLE method of estimation in the same manner as before. This circumvents the problem of uncompleted epidemics in households but at the expense of ignoring the information about p_L contained in those single-household epidemics.

Figure 4.7 gives kernel density estimates of p_L (true value 0.61) for 1000 simulations of Reed-Frost epidemics with parameters as outlined at the beginning of this section. Estimates were made in the first generation at which 1000 recoveries were observed using the full- and recovery-pseudolikelihood methods (i.e. both with and without the ability to distinguish between susceptibles and infectives) and by using the adapted basic MpLE method outlined above. Note that all three methods appear to give estimates that are centred roughly around the true value of p_L, however, the adapted basic MpLE method estimates have a far larger variance than the other estimates, suggesting that the full- and recovery-pseudolikelihood methods are preferable, regardless of whether or not infectives are distinguishable. Estimates were made after 1000 recoveries had been observed rather than the 500 recoveries used in the continuous-time case, owing to the time it takes for 500 recoveries to occur potentially being $n_{max} - 1 = 5$ generations.

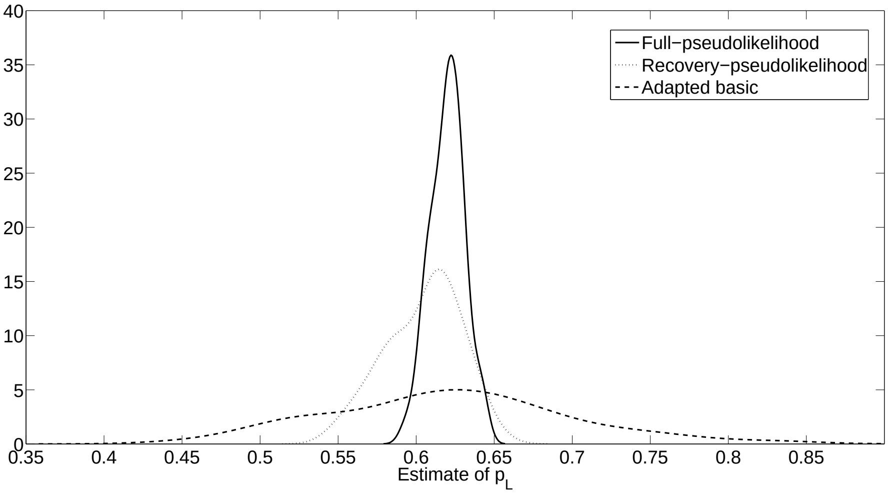

Figure 4.7: Kernel density estimates of the distribution of the estimator of p_L (true value 0.61) based on 1000 simulations of Reed-Frost type epidemics; see text for details

4.5.2 Relationship between parameters of the model and bias of the basic and censored MpLE methods

We examine the extent of the bias of the basic and censored MpLE methods and how the bias is affected by various parameters of an epidemic, by considering "perfect" household data, $\boldsymbol{a}$, from an emerging epidemic (as determined by its CMJBP or multitype branching process approximation) and using these data to estimate λ_L (or p_L if the model is Reed-Frost) using the basic and censored MpLE methods. Households data are considered to be perfect for an emerging epidemic in continuous-time with parameters λ_L and r, if the proportion of households in state (n, x, y) is exactly $\tilde{\alpha}_n r \tilde{p}^{(n)}_{x,y}(r|\lambda_L)$ for all $(n, x, y) \in \mathcal{T}$. (Note that with perfect data, $\hat{\lambda}_L = \text{argmax}\, \tilde{l}^{(\infty)}_{full}$, see equation (4.6.6) in Section 4.6.) Similarly, perfect data for an emerging Reed-Frost epidemic with parameters p_L and μ_G is achieved when the proportion of households in state (n, x, y) is exactly $p_{RFfull}(n, x, y|p_L, \mu_G)$ for all $(n, x, y) \in \mathcal{T}_{RF}$. Note that in both cases, the distribution of household states representing perfect data is also dependent on the population structure $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_{n_{max}})$. Note also that assuming perfect data is equivalent to assuming an infinite population, in which all households are observed, and that in this setting, estimates of the SAR have no illustrative advantage over those of λ_L, since all estimates have no variance.

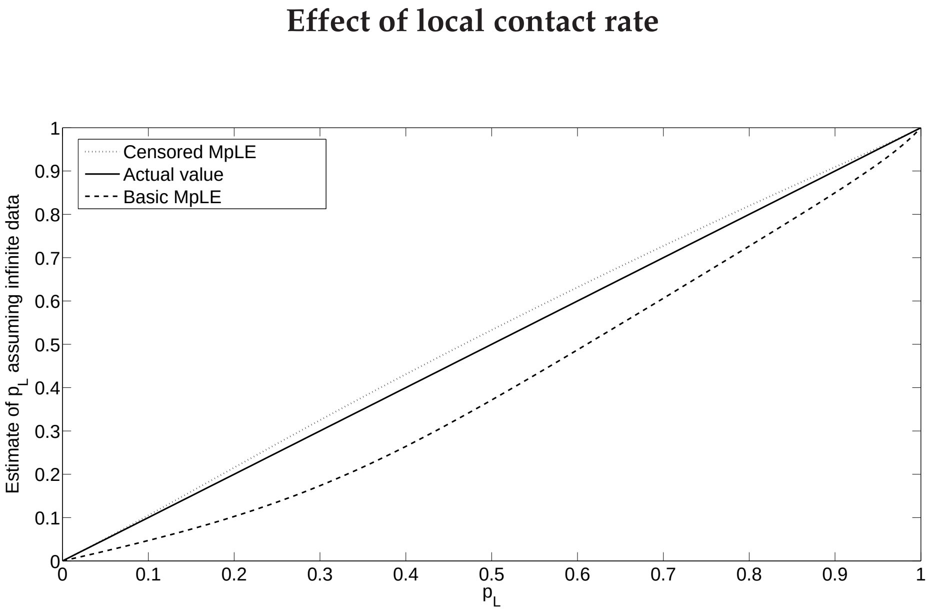

Figure 4.8: Estimates of different values p_L assuming perfect data in emerging Reed-Frost type epidemics, $\rho = 2.248$, using the basic and censored MpLE methods

Figure 4.8 illustrates the effect of the local contact rate on the bias of the basic and censored MpLE methods by considering estimates of p_L for emerging Reed-Frost epidemics with geometric growth rate $\rho = 2.248$ and household distribution $\boldsymbol{\alpha} = [0.13, 0.30, 0.23, 0.18, 0.09, 0.07]$, as given in Section 4.5.1 but with different local contact probabilities. Note that given perfect data, both estimates converge to the true value of p_L as p_L tends to 0 or 1. This can be easily explained by noting that all completed single-household epidemics in households of size n will have exactly 1 recovery if $p_L = 0$ and exactly n recoveries if $p_L = 1$, implying that the issue of less severe single-household epidemics being more likely to be included in the estimation data becomes irrelevant since all single-household epidemics are of the same severity. The basic and censored MpLE methods appear to be at their most biased in the region $0.3 < p_L < 0.6$ when the proportion of recoveries from single-household epidemics in households of sizes 3 and 4 (which make up a significant portion of the population) are distributed in a relatively uniform manner.

Effect of household size

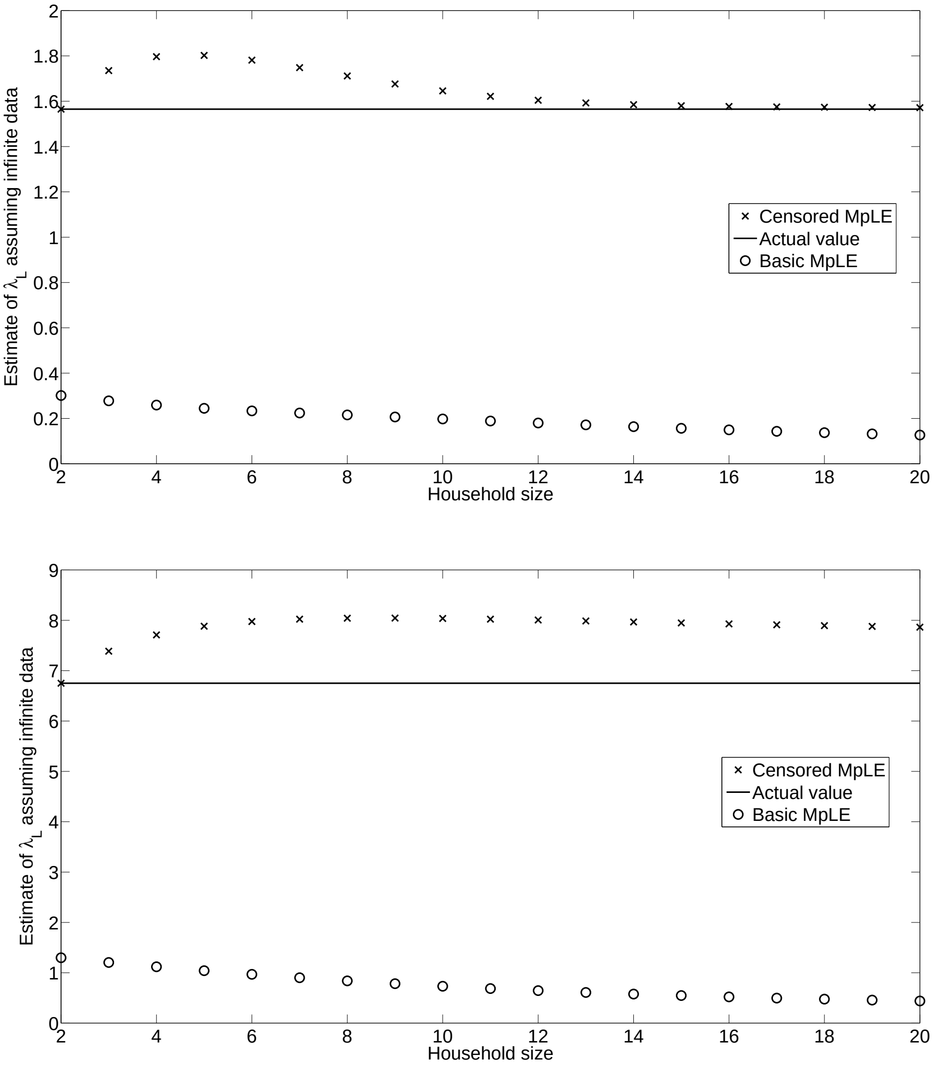

Figure 4.9: Estimates of λ_L assuming perfect data for emerging epidemics, with $r = 1.762$, among populations with equal household sizes using the basic and censored MpLE methods. The upper plot takes $\lambda_L = 1.565$ for all household sizes. The lower plot adopts the model $\lambda_L^{(n)} = \lambda_L/n$, where n is household size and $\lambda_L = 6.75$

Figure 4.9 gives two plots showing estimates of λ_L in continuous-time epidemics with real-time growth rate $r = 1.762$ assuming perfect data for populations of equal sized households from 2 to 20. The upper plot considers the case where $\lambda_L = 1.565$, independent of household size. In this plot the basic MpLE estimate considerably underestimates λ_L regardless of household size but the bias appears to get marginally worse as household size increases. This can be attributed to the most severe single-household epidemics taking longer in larger households and hence fewer of the more severe epidemics are completed by the time of estimation in larger households. The censored MpLE fares better however and appears to converge towards the true value of λ_L as household size increases. Since λ_L is a person-to-person contact rate, larger households are far more likely to have severe epidemics than smaller households with the same λ_L, since the number of local infectious contacts in a household increases quadratically with n. Therefore, as household size increases, the proportion of recoveries from single-household epidemics with the same local contact rate becomes less uniform, leading to less bias in the censored MpLE estimate (as observed in Figure 4.8).

The lower plot of Figure 4.9 uses the same real-time growth rate and population distributions but assumes that the local infection rate depends on household size, specifically that $\lambda_L^{(n)} = \lambda_L/n$ with $\lambda_L = 6.75$ (i.e. the Cauchemez model described in Section 4.3.2). This value was chosen as it gives a value of $\lambda_G = 1.21$ when $r = 1.762$ from the population distribution $\boldsymbol{\alpha}$ as used previously in this section. Here it can be seen that the basic MpLE approach again becomes more biased as household size increases while the censored MPLE approach does not appear to converge back towards the true value of λ_L. In the basic case this is for the same reasons as before, whereas in the censored case, the additional local contacts that come from an increased household size are compensated by the reduction of the local contact rate, leading to the relatively uniform distribution of recoveries in a single household-epidemic which causes bias. However, as in the upper plot, the censored MpLE does eventually converge back towards the true value of λ_L as household size increases beyond the scope of Figure 4.9 since the number of local contacts in a household increases with household size at a greater rate than the local contact rate decreases.

Effect of growth rate

Figure 4.10 shows estimates of λ_L in emerging epidemics with λ_L and $\boldsymbol{\alpha}$ as defined in Section 4.5.1. It is clear from the plot that both the basic and censored

MpLE estimates converge to the true value of λ_L as $r \to 0$, as is proved in Section 4.3.1.

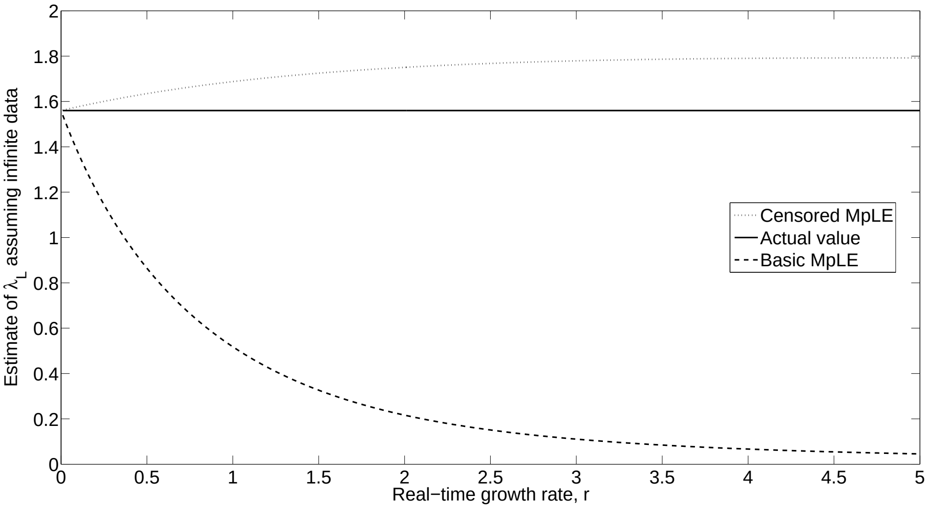

Figure 4.10: Estimates of λ_L assuming perfect data in emerging epidemics with different real-time growth rates r using the basic and censored MpLE methods

4.5.3 Accuracy of the new estimator

In Section 4.5.3 we discuss the reliability of the full-pseudolikelihood method (following Figure 4.6) and in particular when it is at its most accurate. We now attempt to provide some insight into how population and household sizes affect the accuracy of this estimator of λ_L (or the SAR). Let $\tilde{a}^{(n)}_{x,y}$ denote the proportion of infected households of size n that are in state (n, x, y), then the TV (total variation) distance between the observed epidemic and the limiting distribution of its approximating CMJ branching process is given by

$$D(\boldsymbol{a}, \lambda_L, r) = \sum_{(n,x,y)\in\mathcal{T}} \tilde{\alpha_n} |\tilde{a}^{(n)}_{x,y} - \tilde{p}^{(n)}_{x,y}(\hat{r}|\lambda^{(n)}_L)|$$

Figure 4.11 shows how $D(\boldsymbol{a}, \lambda_L, r)$ changes as epidemics progress. Specifically epidemics with parameters λ_G = 1.21, $\lambda_L = 0.64$ and a unit-mean exponential infectious period were simulated and 1000 that took off were used for each of the following population structures. In the left hand plot, populations of 21000 individuals are partitioned into equally sized households of 2, 4, 6 and 8 while the right hand plot uses populations of 5000, 10000, 20000 households of size 4 and a CMJBP made up of households of size 4 (representing an infinite population). During the simulations, the distance $D(\boldsymbol{a}, \lambda_L, r)$ was recorded at regular intervals based on the number of recovered individuals observed and the mean TV distance at each interval over the 1000 simulations provided the data points for the plots.

For a population of 21000 individuals, $D(\boldsymbol{a}, \lambda_L, r)$ is minimised after approximately 500 recoveries have occurred regardless of the household size. As stated in the discussion surrounding Figure 4.6, before this point the epidemics have not had long enough in general to settle into behaviour resembling the asymptotic behaviour of the CMJBP, whilst, after this point, global infectious contacts with susceptibles in previously infected households begin to make the CMJBP approximation break down.

It is also worth noting the general pattern of $D(\boldsymbol{a}, \lambda_L, r)$ increasing as household size increases. Initially this can be attributed to the smaller state space in epidemics with smaller households reducing the number of elements in the sum used to calculate $D(\boldsymbol{a}, \lambda_L, r)$ and allowing the epidemic to settle into its approximate CMJBP behaviour more quickly. As epidemics progress, the greater number of households in epidemics with smaller-sized households also means that global infectious contacts with susceptibles in previously infected households occur less frequently, so $D(\boldsymbol{a}, \lambda_L, r)$ remains small for longer in populations split into smaller sized households. The right hand plot shows that as population size increases, the number of recoveries needed before the CMJBP approximation begins to break down becomes increasingly large (again offering agreement with Barbour and Utev [2004] as discussed after Figure 4.6). For an infinite population, the mean TV distance converges towards zero as the number of recoveries increases, as predicted by theory; the mean TV distance drops quickly to about 0.05 but thereafter convergence is much slower.

4.6 Strong consistency of estimators

We consider the asymptotic behaviour of the estimators described in Section 4.3 as the number of households in the population tends to infinity. Specifically

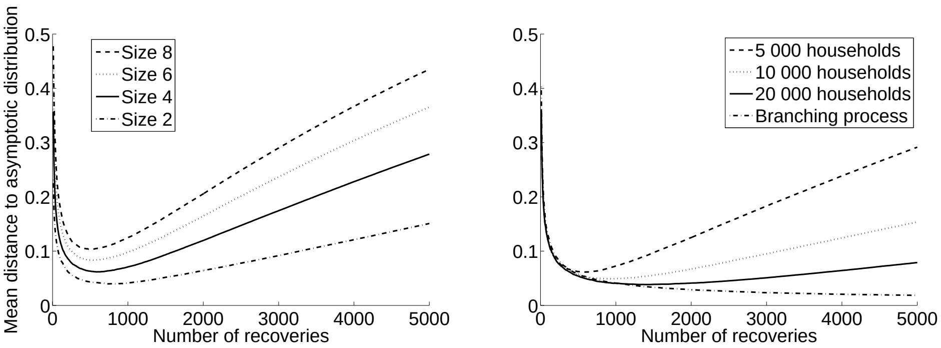

Figure 4.11: Mean TV distances between the observed and asymptotic distribution of household states as the number of recovered individuals increases, based on 1000 epidemic simulations. See main text of Section 4.5.3 for further details

we show that, under suitable conditions, the estimators are strongly consistent, conditional upon the epidemic taking off. This section extends the results of Professor Frank Ball that were originally published in Ball and Shaw [2015], which only considers estimators of λ_L under the basic epidemic model and assumes that the recovery rate γ is known. The proof is adapted here to incorporate estimators household size dependent $\lambda_L^{(n)}$ and an unknown infectious period/recovery rate γ.

The proof is structured as follows. Theorem 4.6.1 shows that if the approximating branching process takes off and if the number of households in the population is large enough, the epidemic does mimic the approximating branching process for some time at the start of the outbreak. (Recall from Chapter 2 that we are not interested in the case where the epidemic does not take off.) Specifically, we equate births in the approximating branching process to global contacts made by infectives in the epidemic and consider the number of such contacts in the epidemic until the first one with an individual not in a fully susceptible household. We then show that, if the epidemic occurs in a large enough population, there is a time frame (which tends to ∞ as $m \to \infty$) in which the approximating branching process contains strictly fewer births than this value, meaning that all births in the approximating branching process during this time frame correspond to a global infectious contact with an individual in a fully susceptible household in the epidemic.

Theorems 4.6.2 and 4.6.3 show the strong consistency of the estimators obtained using the full- and recovery-pseudolikelihood methods respectively. We adapt the standard methodology of considering the maximum of the pseudolikelihood function of our unknown parameters for epidemics with large m and the maximum of the limit of a sequence of these functions as $m \to \infty$. Exploration of the behaviour of the pseudolikelihood function towards the limits of its domain are required to complete the proof and it is here that the extension of the results in Ball and Shaw [2015] is particularly non-trivial. Completion of the proof of strong consistency is achieved for the full-pseudolikelihood estimator but is left as an open problem for the recovery-pseudolikelihood estimator.

Consider a sequence of epidemics $E^{(m)}$ $(m = 1, 2, ...)$, indexed by the number of households in the population. For $m = 1, 2, ...$ and $n = 1, 2, ..., n_{max}$, let $\alpha_n^{(m)}$ be the proportion of households in $E^{(m)}$ that have size n. The epidemic $E^{(m)}$ is as defined in Chapter 2 and has one initial infective, who is chosen uniformly at random from the population. The infection parameters $(\lambda_G, \lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})})$ and the infectious period distribution are all assumed to be independent of m, as is the maximum household size n_{max}. We assume in this proof that the infectious period T_I takes an exponential distribution with unknown rate γ, as suggested in Section 4.3.2. Adapting the proof to the case in which T_I takes an arbitrary but known distribution is trivial (and is the proof given in Ball and Shaw [2015]). Let $\boldsymbol{\theta} = (\gamma, \lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})})$ be a vector denoting the unknown parameters which we are estimating. It is assumed that $\alpha_n^{(m)} \to \alpha_n$ as $n \to \infty$ $(n = 1, 2, ..., n_{max})$.

Let $E^{(\infty)}$ denote the general branching process, analysed in Section 4.3, which approximates the epidemic $E^{(m)}$ for suitably large m. Recall that for $(n, x, y) \in \mathcal{T}$, the number of individuals in $E^{(\infty)}$ having state (n, x, y) at time t is denoted by $Y_{n,x,y}(t)$. For $m = 1, 2, ...$, $(n, x, y) \in \mathcal{T}$ and $t \geq 0$, let $Y_{n,x,y}^{(m)}(t)$ denote the number of size-n households in $E^{(m)}$ that have x susceptibles and y infectives at time t. Let $\mathcal{T}_L = \{(n, x, y) \in \mathcal{T} : \ y \geq 1\}$. For $t \geq 0$, let $Y(t) = \sum_{(n,x,y) \in \mathcal{T}_L} Y_{n,x,y}(t)$ denote the number of "live" individuals in $E^{(\infty)}$ at time t. Recall that r denotes the Malthusian parameter of $E^{(\infty)}$.

Theorem 4.6.1. *Suppose that $r > 0$. Then there is a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ on which are defined a sequence of epidemics $E^{(m)}$ $(m \geq 1)$ and the approximating branching process $E^{(\infty)}$ satisfying the following property. Let $A = \{\omega \in \Omega : \lim_{t \to \infty} Y(t, \omega) = 0\}$ denote the set on which the branching process $E^{(\infty)}$ goes extinct. Then for* P*-almost* *all $\omega \in A^c$ and any $c \in (0, \frac{1}{2}r^{-1})$,*

$$\sup_{0 \le t \le c \log m} \max_{(n,x,y) \in \mathcal{T}} |Y^{(m)}_{n,x,y}(t,\omega) - Y_{n,x,y}(t,\omega)| = 0 \tag{4.6.1}$$

for all sufficiently large m.

Proof. For $m = 1, 2, ...,$ let $N^{(m)} = m \sum_{n=1}^{n_{max}} n\alpha_n^{(m)}$ denote the total number of individuals in the population among which $E^{(m)}$ is spreading. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space on which are defined the following independent sets of random quantities: (i) a realisation of the branching process $E^{(\infty)}$; (ii) $\chi_k^{(m)}$ $(m = 1, 2, ...;\ k = 1, 2, ...)$, where for each m, $\chi_1^{(m)}, \chi_2^{(m)}, ...$ are independent and uniformly distributed on $\{1, 2, ..., N^{(m)}\}$.

For $m = 1, 2, ...,$ a realisation of the early stages of the epidemic $E^{(m)}$ can be defined on $(\Omega, \mathcal{F}, \mathrm{P})$ as follows. Label the individuals in the m^{th} population $1, 2, ..., N^{(m)}$. The initial infective in $E^{(m)}$ has a label given by $\chi_1^{(m)}$ and corresponds to the ancestor in the branching process $E^{(\infty)}$. Births of individuals in $E^{(\infty)}$ correspond to global infectious contacts being made in $E^{(m)}$. For $k = 1, 2, ..,$ the individual contacted in $E^{(m)}$ corresponding to the k^{th} birth in $E^{(\infty)}$ has a label given by $\chi_{k+1}^{(m)}$. If the household in which $\chi_{k+1}^{(m)}$ resides has not been infected previously, then $\chi_{k+1}^{(m)}$ becomes infected in $E^{(m)}$ and initiates a new single-household epidemic in $E^{(m)}$ whose course and subsequent global infectious contacts are given by the life-history of the $(k+1)^{th}$ individual in $E^{(\infty)}$. If the household in which $\chi_{k+1}^{(m)}$ resides has been infected previously then the construction of $E^{(m)}$ needs modifying but such detail is not required for the present proof. Note that local infectious contacts still occur but are not births in the approximating branching process since they do not infect new households.

For $m = 1, 2, ...,$ let $M^{(m)}$ be the smallest $k \ge 2$ such that $\chi_k^{(m)}$ belongs to the same household as $\chi_l^{(m)}$ for some $l = 1, 2, ..., k-1$, and let $\hat{M}^{(m)}$ be a random variable, taking values in $2, 3, ...,$ having survivor function

$$\mathbb{P}(\hat{M}^{(m)} > k) = \prod_{i=1}^{k-1} (1 - i n_{max}/N^{(m)}) \quad (k = 2, 3, ...).$$

Note that $M^{(m)}$ is stochastically greater than $\hat{M}^{(m)}$, since the maximum household size is n_{max}, and (cf. Aldous [1985], p.96) $m^{-1/2}\hat{M}^{(m)} \xrightarrow{\mathrm{D}} \hat{M}$ as $m \to \infty$, where $\xrightarrow{\mathrm{D}}$ denotes convergence in distribution and $\hat{M}$ has density $f(x) =$

$n_{max} x \mu_H^{-1} \exp\left(-n_{max}\mu_H^{-1}x^2/2\right)$ $(x > 0)$, with $\mu_H = \sum_{n=1}^{n_{max}} n\alpha_n$ being the mean household size. (Note that $m^{-1}N^{(m)} \to \mu_H$ as $m \to \infty$.)

By the Skorokhod representation theorem, the random variables $\hat{M}$, $M^{(m)}$ and $\hat{M}^{(m)}$ $(m = 1, 2, ...)$ may be defined on a common probability space so that $\mathbb{P}(M^{(m)} \geq \hat{M}^{(m)},\ (m = 1, 2, ...)) = 1$ and $m^{-1/2}\hat{M}^{(m)} \xrightarrow{\text{a.s.}} \hat{M}$ as $m \to \infty$. Further, that probability space may be augmented to carry random variables $\chi_k^{(m)}$ $(m = 1, 2, ...;\ k = 1, 2, ...)$ distributed as above and consistent with $M^{(m)}$ $(m = 1, 2, ...)$. Thus we may assume that the random variables $\hat{M}^{(m)}$ $(m = 1, 2, ...)$ and $\hat{M}$ are also defined on $(\Omega, \mathcal{F}, \text{P})$ and that there exists $B \in \mathcal{F}$ with $\mathbb{P}(B) = 1$, such that, for all $\omega \in B$,

$$M^{(m)}(\omega) \geq \hat{M}^{(m)}(\omega) \quad \text{and} \quad m^{-1/2}\hat{M}^{(m)}(\omega) \to \hat{M}(\omega) \quad \text{as } m \to \infty. \tag{4.6.2}$$

For $t \geq 0$, let $T(t)$ be the number of births in $E^{(\infty)}$ during $[0, t]$, including the ancestor. Then $T(t) = \sum_{(n,x,y)\in\mathcal{T}} Y_{n,x,y}(t)$ and it follows from (4.3.4) that $e^{-rt}T(t) \xrightarrow{\text{a.s.}} r^{-1}W$ as $t \to \infty$. Recall that $W = 0$ if and only if the branching process goes extinct. Thus there exists $C \in \mathcal{F}$, with $C \subseteq A^c$ and $\mathbb{P}(C) = \mathbb{P}(A^c)$, such that for all $\omega \in C$,

$$e^{-rt}T(t, \omega) \to r^{-1}W(\omega) \quad \text{as } t \to \infty. \tag{4.6.3}$$

Let $\omega \in B \cap C$ and $c \in (0, \frac{1}{2}r^{-1})$. Then it follows from (4.6.3) that $T(c\log m, \omega) < 2m^{rc}r^{-1}W(\omega)$ for all sufficiently large m. Also, (4.6.2) implies that $M^{(m)}(\omega) > \frac{1}{2}m^{1/2}\hat{M}(\omega)$ for all sufficiently large m. Hence, since $rc < 1/2$, for all sufficiently large m, every birth in $E^{(\infty)}(\omega)$ during $(0, c\log m]$ corresponds to a global contact with an uninfected household in $E^{(m)}(\omega)$ and (4.6.1) follows since $\mathbb{P}(B \cap C) = \mathbb{P}(A^c)$.

□

We turn now to estimation of $\lambda_L^{(n)}$ $(n = 2, 3, ..., n_{max})$ and γ. Suppose that the epidemic $E^{(m)}$ is observed at time $t^{(m)}$, where the sequence $(t^{(m)})$ satisfies (i) $t^{(m)} \to \infty$ as $m \to \infty$, (ii) $t^{(m)} \leq c\log m$ for all sufficiently large m, for some $c \leq (2r)^{-1}$. Suppose also that an estimator $\hat{r}^{(m)}$ of the growth rate r is available such that $\hat{r}^{(m)} \underset{A^c}{\xrightarrow{\text{a.s.}}} r$ as $m \to \infty$ where $\underset{A^c}{\xrightarrow{\text{a.s.}}}$ means convergence for P-almost all $\omega \in A^c$. It is easily verified that one such estimator is $\hat{r}^{(m)} = \log[(T^{(m)}(t^{(m)})/T^{(m)}(t^{(m)}/2))]/(t^{(m)}/2)$, where $T^{(m)}(t)$ is the total number of households that have been infected in $E^{(m)}$ by time t. Let $\hat{\boldsymbol{\theta}}^{(m)}_{full}$ denote the estimator obtained by maximising the function $L_{full}(\gamma, \lambda_L^{(2)}, \lambda_L^{(3)}, ..., \lambda_L^{(n_{max})} | \boldsymbol{a}, \hat{r}^{(m)})$ defined at (4.3.5) and extended to include γ. For ease of exposition, we assume that all infected households are observed, so, in our present notation, $a^{(m)}_{x,y} = Y^{(m)}_{n,x,y}(t^{(m)})$ for $(n, x, y) \in \mathcal{T}$. The following theorems are easily extended to the situation when only some infected households are observed; of course, the number of observed households must tend to infinity as $m \to \infty$ and the sampling mechanism must be independent of disease progression within households. In these theorems, it is convenient to denote the true value of $\boldsymbol{\theta}$ by $\bar{\boldsymbol{\theta}} = (\bar{\gamma}, \bar{\lambda}_L^{(2)}, \bar{\lambda}_L^{(3)}, ..., \bar{\lambda}_L^{(n_{max})})$.

Theorem 4.6.2. *Under the conditions of Theorem 4.6.1,*

$$\hat{\boldsymbol{\theta}}^{(m)}_{full} \xrightarrow[A^c]{a.s.} \bar{\boldsymbol{\theta}} \quad \textit{as } m \to \infty.$$

Proof. First note that from (4.3.5)

$$\hat{\boldsymbol{\theta}}^{(m)}_{full} = \text{argmax}\, \tilde{l}^{(m)}_{full}(\boldsymbol{\theta} | \boldsymbol{Y}^{(m)}, \hat{r}^{(m)}), \tag{4.6.4}$$

where

$$\tilde{l}^{(m)}_{full}(\boldsymbol{\theta} | \boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) = W^{-1} e^{-rt^{(m)}} \sum_{n=1}^{n_{max}} \sum_{(x,y) \in \mathcal{T}^{(n)}} Y^{(m)}_{n,x,y}(t^{(m)}) \log \tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)} | \boldsymbol{\theta}).$$

(Note that, unlike in Ball and Shaw [2015], we must sum across all household sizes since households of size 1 contribute information towards estimating $\bar{\gamma}$.)

Observe that, under the conditions satisfied by $(t^{(m)})$, Theorem 4.6.1 and (4.3.4) imply that, for all $(n, x, y) \in \mathcal{T}$,

$$W^{-1} e^{-rt^{(m)}} Y^{(m)}_{n,x,y}(t^{(m)}) \xrightarrow[A^c]{\text{a.s.}} \tilde{\alpha}_n \tilde{p}^{(n)}_{x,y}(r | \bar{\boldsymbol{\theta}}) \quad \text{as } m \to \infty. \tag{4.6.5}$$

Hence, since $\hat{r}^{(m)} \xrightarrow[A^c]{\text{a.s.}} r$ as $m \to \infty$, we have that for any $\boldsymbol{\theta} \in (0, \infty)^{n_{max}}$,

$$\tilde{l}^{(m)}_{full}(\boldsymbol{\theta} | \boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) \xrightarrow[A^c]{\text{a.s.}} \tilde{l}^{(\infty)}_{full}(\boldsymbol{\theta} | r) \quad \text{as } m \to \infty,$$

where

$$\tilde{l}^{(\infty)}_{full}(\boldsymbol{\theta} | r) = \sum_{n=1}^{n_{max}} \tilde{\alpha}_n \sum_{(x,y) \in \mathcal{T}^{(n)}} \tilde{p}^{(n)}_{x,y}(r | \bar{\boldsymbol{\theta}}) \log \tilde{p}^{(n)}_{x,y}(r | \boldsymbol{\theta}). \tag{4.6.6}$$

Standard arguments, (e.g. Silvey [1975], page 75) show that, for $n = 2, 3, ..., n_{max}$, the function $g_n(\boldsymbol{\theta}) = \sum_{(x,y)\in\mathcal{T}^{(n)}} \tilde{p}^{(n)}_{x,y}(r|\bar{\boldsymbol{\theta}}) \log \tilde{p}^{(n)}_{x,y}(r|\boldsymbol{\theta})$ has a unique global maximum at $\bar{\boldsymbol{\theta}}$. Hence, as a function of $\boldsymbol{\theta} \in (0,\infty)^{n_{max}}$, $\tilde{l}^{(\infty)}_{full}(\boldsymbol{\theta}|r)$ has a unique global maximum at $\bar{\boldsymbol{\theta}}$.

Fix K such that K is a compact subset of $(0,\infty)^{n_{max}}$ and $\bar{\boldsymbol{\theta}} \in K$. Then

$$\max_{\boldsymbol{\theta}\in K} |\tilde{l}^{(m)}_{full}(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) - \tilde{l}^{(\infty)}_{full}(\boldsymbol{\theta}|r)| \leq \sum_{n=1}^{n_{max}} \sum_{(x,y)\in\mathcal{T}^{(n)}} \max_{\boldsymbol{\theta}\in K} g^{(m)}_{n,x,y}(\boldsymbol{\theta}), \tag{4.6.7}$$

where

$$g^{(m)}_{n,x,y}(\boldsymbol{\theta}) = |W^{-1}e^{-rt^{(m)}} Y^{(m)}_{n,x,y}(t^{(m)}) \log \tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)}|\boldsymbol{\theta}) - \tilde{\alpha}_n \tilde{p}^{(n)}_{x,y}(r|\bar{\boldsymbol{\theta}}) \log \tilde{p}^{(n)}_{x,y}(r|\boldsymbol{\theta})|.$$

Now

$$g^{(m)}_{n,x,y}(\boldsymbol{\theta}) \leq \hat{g}^{(m)}_{n,x,y}(\boldsymbol{\theta}) + \check{g}^{(m)}_{n,x,y}(\boldsymbol{\theta}), \tag{4.6.8}$$

where

$$\hat{g}^{(m)}_{n,x,y}(\boldsymbol{\theta}) = W^{-1}e^{-rt^{(m)}} Y^{(m)}_{n,x,y}(t^{(m)}) |\log \tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)}|\boldsymbol{\theta}) - \log \tilde{p}^{(n)}_{x,y}(r|\boldsymbol{\theta})|$$

and

$$\check{g}^{(m)}_{n,x,y}(\boldsymbol{\theta}) = |\{W^{-1}e^{-rt^{(m)}} Y^{(m)}_{n,x,y}(t^{(m)}) - \tilde{\alpha}_n \tilde{p}^{(n)}_{x,y}(r|\bar{\boldsymbol{\theta}})\} \log \tilde{p}^{(n)}_{x,y}(r|\boldsymbol{\theta})|.$$

Using (4.6.5), for all $(n, x, y) \in \mathcal{T}$,

$$\max_{\boldsymbol{\theta}\in K} \check{g}^{(m)}_{n,x,y}(\boldsymbol{\theta}) \xrightarrow[A^c]{\text{a.s.}} 0 \quad \text{as } m \to \infty. \tag{4.6.9}$$

Further, for any $\boldsymbol{\theta} > \mathbf{0}$ (where $\mathbf{0}$ is a vector of zeros of length n_{max}) and $r, r' > 0$,

$$|\tilde{p}^{(n)}_{x,y}(r|\boldsymbol{\theta}) - \tilde{p}^{(n)}_{x,y}(r'|\boldsymbol{\theta})| \leq \int_0^\infty |e^{-rt} - e^{-r't}| \, \mathrm{d}t = |r - r'|/(rr'), \tag{4.6.10}$$

so, since $\log x$ is uniformly continuous on any bounded subinterval of $(0,\infty)$ and the estimator $\hat{r}^{(m)} \xrightarrow[A^c]{\text{a.s.}} r$ as $m \to \infty$, it follows using (4.6.5) that, for all $(n, x, y) \in \mathcal{T}$,

$$\max_{\boldsymbol{\theta}\in K} \hat{g}^{(m)}_{n,x,y}(\boldsymbol{\theta}) \xrightarrow[A^c]{\text{a.s.}} 0 \quad \text{as } m \to \infty. \tag{4.6.11}$$

Combining (4.6.4) - (4.6.9) yields

$$\max_{\boldsymbol{\theta}\in K} |\tilde{l}^{(m)}_{full}(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) - \tilde{l}^{(\infty)}_{full}(\boldsymbol{\theta}|r)| \xrightarrow[A^c]{\text{a.s.}} 0 \text{ as } m \to \infty, \tag{4.6.12}$$

whence, since $\tilde{l}^{(\infty)}_{full}(\boldsymbol{\theta}|r)$ has a unique global maximum at $\bar{\boldsymbol{\theta}}$,

$$\operatorname*{argmax}_{\boldsymbol{\theta}\in K} \tilde{l}^{(m)}_{full}(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)},\hat{r}^{(m)}) \xrightarrow[A^c]{a.s.} \bar{\boldsymbol{\theta}} \text{ as } m \to \infty. \tag{4.6.13}$$

To complete the proof we explore the behaviour of $l^{(m)}_{full}(\lambda_L|\boldsymbol{Y}^{(m)},\hat{r}^{(m)})$ as the $\lambda^{(n)}_L,\gamma \downarrow 0$ and the $\lambda^{(n)}_L,\gamma \uparrow \infty$. Loosely speaking, our aim is to show that there exist lower and upper bounds on each element of $\hat{\boldsymbol{\theta}}^{(m)}_{full}$ which must be satisfied for sufficiently large m, independently of other elements of $\hat{\boldsymbol{\theta}}^{(m)}_{full}$. Recall throughout this part of the proof that $\boldsymbol{\theta} = (\gamma,\lambda^{(2)}_L,\lambda^{(3)}_L,...,\lambda^{(n_{max})}_L)$ and $\hat{\boldsymbol{\theta}}^{(m)}_{full}$ and $\bar{\boldsymbol{\theta}}$ are vectors with elements denoted in the obvious manner.

We begin by considering $\lambda^{(n)}_L$ $(n = 2,3,...,n_{max})$. Let X denote the time of the first point in $(0,\infty)$ of a homogeneous Poisson process having rate $(n-1)\lambda^{(n)}_L$. Then $p^{(n)}_{n-2,2}(t|\boldsymbol{\theta}) \leq \mathbb{P}(X \leq t) = 1 - e^{-(n-1)\lambda^{(n)}_L t}$, so

$$\begin{aligned}\tilde{p}^{(n)}_{n-2,2}(r|\boldsymbol{\theta}) &\leq \int_0^\infty (1 - e^{-(n-1)\lambda^{(n)}_L t})e^{-rt}\,\mathrm{d}t\\ &= \frac{(n-1)\lambda^{(n)}_L}{r(r+(n-1)\lambda^{(n)}_L)}\\ &\leq (n-1)\lambda^{(n)}_L/r^2.\end{aligned} \tag{4.6.14}$$

For all n, we have that $\tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)}|\boldsymbol{\theta}) \leq 1/\hat{r}^{(m)}$ for all $(x,y) \in \mathcal{T}^{(n)}$, so

$$\log \tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)}|\boldsymbol{\theta}) + \log \hat{r}^{(m)} \leq 0. \tag{4.6.15}$$

Let

$$\begin{aligned}&l^{(m)}_*(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)},\hat{r}^{(m)})\\ &\quad = W^{-1}e^{-rt^{(m)}}\sum_{n=1}^{n_{max}}\sum_{(x,y)\in\mathcal{T}^{(n)}} Y^{(m)}_{n,x,y}(t^{(m)})(\log\tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)}|\boldsymbol{\theta}) + \log\hat{r}^{(m)})\\ &\quad = l^{(m)}_{full}(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)},\hat{r}^{(m)}) + W^{-1}e^{-rt^{(m)}}\sum_{n=1}^{n_{max}}\sum_{(x,y)\in\mathcal{T}^{(n)}} Y^{(m)}_{n,x,y}(t^{(m)})\log\hat{r}^{(m)},\end{aligned}$$

and, recalling (4.6.4), note that $\hat{\boldsymbol{\theta}}^{(m)}_{full} = \operatorname{argmax} l^{(m)}_*(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)},\hat{r}^{(m)})$.

Fix $\lambda_{n,0} > 0$. Then (4.6.14), (4.6.15) and, subsequently, (4.6.5) imply that, for all $\boldsymbol{\theta}$ such that $\lambda^{(n)}_L \in (0,\lambda_{n,0}]$,

$$\begin{aligned}l^{(m)}_*(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)},\hat{r}^{(m)}) &\leq W^{-1}e^{-rt^{(m)}}Y^{(m)}_{n,n-2,2}(t^{(m)})(\log(n-1) + \log\lambda_{n,0} - \log\hat{r}^{(m)})\\ &\xrightarrow[A^c]{\text{a.s.}} \tilde{\alpha}_n\tilde{p}^{(n)}_{n-2,2}(r|\bar{\boldsymbol{\theta}})[\log(n-1) + \log\lambda_{n,0} - \log r]\end{aligned} \tag{4.6.16}$$

as $m \to \infty$. Also, using (4.6.5) and (4.6.12),

$$l_*^{(m)}(\bar{\boldsymbol{\theta}}|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) \xrightarrow[A^c]{a.s.} l_{full}^{(\infty)}(\bar{\boldsymbol{\theta}}|r) + r^{-1}\log r \sum_{n=1}^{n_{max}} \tilde{\alpha}_n \quad \text{as } m \to \infty. \tag{4.6.17}$$

For n chosen such that $\tilde{\alpha}_n > 0$ (noting that if $\tilde{\alpha}_n = 0$ then estimation of $\lambda_L^{(n)}$ is unnecessary), choose $\lambda_{n,0} > 0$ and the right hand side of (4.6.16) is strictly less than the right hand side of (4.6.17). Let F_i be a function denoting the projection of a vector onto its i^{th} element. For example, if $\boldsymbol{x} = (x_1, x_2, ..., x_j)$, then, for $i \leq j$, $F_i(\boldsymbol{x}) = x_i$. Then, since $\hat{\boldsymbol{\theta}}_{full}^{(m)} = \operatorname{argmax} l_*^{(m)}(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)})$, it follows that for $n = 2, 3, ..., n_{max}$ and P-almost all $\omega \in A^c$, there exists $m_{n,0}(\omega)$ such that

$$F_n(\hat{\boldsymbol{\theta}}_{full}^{(m)}(\omega)) \notin (0, \lambda_{n,0}) \quad \text{for all } m \geq m_{n,0}(\omega). \tag{4.6.18}$$

Note that this behaviour is independent of the other elements of $\boldsymbol{\theta}$.

Let T_I denote the infectious period of the initial infective in a household of size n. Then $p_{n-1,1}^{(n)}(t|\boldsymbol{\theta}) = \mathbb{E}[e^{-(n-1)\lambda_L^{(n)}t}\mathbb{1}_{\{T_I>t\}}] \leq e^{-(n-1)\lambda_L^{(n)}t}$, whence $\tilde{p}_{n-1,1}^{(n)}(r|\boldsymbol{\theta}) \leq 1/((n-1)\lambda_L^{(n)} + r)$. Fixing $\lambda_{n,1}$ and arguing as in (4.6.16) implies that, for $\lambda_L^{(n)} \in [\lambda_{n,1}, \infty)$

$$\begin{aligned} l_*^{(m)}(\boldsymbol{\theta}|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) &\leq W^{-1}e^{-rt^{(m)}}Y_{n,n-1,1}^{(m)}(t^{(m)})[\log \hat{r}^{(m)} - \log(\hat{r}^{(m)} + (n-1)\lambda_{n,1})] \\ &\xrightarrow[A^c]{\text{a.s.}} \tilde{\alpha}_n \tilde{p}_{n-1,1}^{(n)}(r|\bar{\boldsymbol{\theta}})[\log \hat{r}^{(m)} - \log(\hat{r}^{(m)} + (n-1)\lambda_{n,1})] \end{aligned} \tag{4.6.19}$$

Choosing $\lambda_{n,1}$ large enough such that the right hand side of (4.6.19) is strictly less than the right hand side of (4.6.17) and arguing as before shows that there exists $\lambda_{n,1} < \infty$ such that, for P-almost all $\omega \in A^c$, there exists $m_{n,1}(\omega)$ such that

$$F_n(\hat{\boldsymbol{\theta}}_{full}^{(m)}(\omega)) \notin (\lambda_{n,1}, \infty) \quad \text{for all } m \geq m_{n,1}(\omega). \tag{4.6.20}$$

A similar argument holds for γ. Observe, by following the techniques above, that $\tilde{p}_{n-1,0}^{(n)}(r|\boldsymbol{\theta}) \leq \gamma/r^2$ and that $p_{n-1,1}^{(n)}(t|\boldsymbol{\theta}) \leq e^{-\gamma t}$. Thus the proof above can be easily adapted to show that there exist $\gamma_0, \gamma_1 > 0$ such that for P-almost all $\omega \in A^c$, there exists $m_0(\omega)$ and $m_1(\omega)$ such that

$$F_1(\hat{\boldsymbol{\theta}}_{full}^{(m)}(\omega)) \notin (0, \gamma_0) \quad \text{for all } m \geq m_0(\omega) \quad \text{and} \tag{4.6.21}$$

$$F_1(\hat{\boldsymbol{\theta}}_{full}^{(m)}(\omega)) \notin (\gamma_1, \infty) \quad \text{for all } m \geq 1(\omega). \tag{4.6.22}$$

Note again that this behavior is independent of all other parameters of $\boldsymbol{\theta}$. The theorem then follows from (4.6.13), (4.6.18), (4.6.20), (4.6.21) and (4.6.22). □

We now consider estimation of $\boldsymbol{\theta}$ based only on recoveries. For $m = 1, 2, ...,$ $n = 1, 2, ..., n_{max}$ and $t \geq 0$, let

$$Z_{n,j}^{(m)}(t) = \sum_{(x,y)\in A_j^{(n)}} Y_{n,x,y}^{(m)}(t) \quad (j = 1, 2, ..., n)$$

be the total number of size-n households in which j recoveries have been observed by time t in the epidemic $E^{(m)}$. Let $\hat{\boldsymbol{\theta}}_{rec}^{(m)}$ denote the estimator of $\boldsymbol{\theta}$ obtained by maximising the function $L_{rec}(\boldsymbol{\theta}|\boldsymbol{c}, \hat{r}^{(m)})$ described at (4.3.6) and extended to include γ, over any given set $K \in (0, \infty)^{n_{max}}$. (In our present notation $c_j^{(n)} = Z_{n,j}^{(m)}(t^{(m)})$.)

Theorem 4.6.3. *Under the conditions of Theorem 4.6.1,*

$$\hat{\boldsymbol{\theta}}_{rec}^{(m)} \xrightarrow[A^c]{a.s.} \bar{\boldsymbol{\theta}} \quad \textit{as } m \to \infty.$$

if K is any compact subset of $(0, \infty)^{n_{max}}$ *containing* $\bar{\boldsymbol{\theta}}$.

Proof. First note from (4.3.6) that $\hat{\boldsymbol{\theta}}_{rec}^{(m)} = \operatorname{argmax} \tilde{l}_{rec}^{(m)}(\boldsymbol{\theta}|\boldsymbol{Z}^{(m)}, \hat{r}^{(m)})$, where

$$\tilde{l}_{rec}^{(m)}(\lambda_L|\boldsymbol{Z}^{(m)}, \hat{r}^{(m)}) = W^{-1}e^{-rt^{(m)}} \sum_{n=1}^{n_{max}} \sum_{j=1}^{n} Z_{n,j}^{(m)}(t^{(m)}) \log \tilde{q}_j^{(n)}(\hat{r}^{(m)}|\boldsymbol{\theta}).$$

Using (4.6.5), for $n = 2, 3, ..., n_{max}$ and $j = 1, 2, ..., n$,

$$W^{-1}e^{-rt^{(m)}} Z_{n,j}^{(m)}(t^{(m)}) \xrightarrow[A^c]{\text{a.s.}} \tilde{\alpha}_n(r^{-1} - \tilde{q}_0^{(n)}(r|\bar{\boldsymbol{\theta}}))\tilde{q}_j^{(n)}(r|\bar{\boldsymbol{\theta}}) \quad \text{as } m \to \infty, \tag{4.6.23}$$

so, for any $\boldsymbol{\theta} \in (0, \infty)^{n_{max}}$,

$$\tilde{l}_{rec}^{(m)}(\boldsymbol{\theta}|\boldsymbol{Z}^{(m)}, \hat{r}^{(m)}) \xrightarrow[A^c]{\text{a.s.}} \tilde{l}_{rec}^{(\infty)}(\lambda_L|r) \text{ as } m \to \infty,$$

where

$$\tilde{l}_{rec}^{(\infty)}(\boldsymbol{\theta}|r) = \sum_{n=1}^{n_{max}} \tilde{\alpha}_n(r^{-1} - \tilde{q}_0^{(n)}(r|\bar{\boldsymbol{\theta}})) \sum_{j=1}^{n} \tilde{q}_j^{(n)}(r|\bar{\boldsymbol{\theta}}) \log \tilde{q}_j^{(n)}(r|\boldsymbol{\theta}). \tag{4.6.24}$$

Now

$$|\tilde{l}_{rec}^{(m)}(\boldsymbol{\theta}|\boldsymbol{Z}^{(m)}, \hat{r}^{(m)}) - \tilde{l}_{rec}^{(\infty)}(\boldsymbol{\theta}|r)| \leq \sum_{n=1}^{n_{max}} \sum_{j=1}^{n} (\hat{h}_{n,j}^{(m)}(\boldsymbol{\theta}) + \check{h}_{n,j}^{(m)}(\boldsymbol{\theta})), \tag{4.6.25}$$

where

$$\hat{h}_{n,j}^{(m)}(\boldsymbol{\theta}) = W^{-1}e^{-rt^{(m)}} Z_{n,j}^{(m)}(t^{(m)})| \log \tilde{q}_j^{(n)}(\hat{r}^{(m)}|\boldsymbol{\theta}) - \log \tilde{q}_j^{(n)}(r|\boldsymbol{\theta})|$$

and

$$\check{h}_{n,j}^{(m)}(\boldsymbol{\theta}) = |\{W^{-1}e^{-rt^{(m)}}Z_{n,j}^{(m)}(t^{(m)}) - \tilde{\alpha}_n(r^{-1} - \tilde{q}_0^{(n)}(r|\bar{\boldsymbol{\theta}}))\tilde{q}_j^{(n)}(r|\bar{\boldsymbol{\theta}})\} \log \tilde{q}_j^{(n)}(r|\boldsymbol{\theta})|.$$

For $n = 2, 3, ..., n_{max}$ and $j = 1, 2, ..., n$,

$$\tilde{q}_j^{(n)}(r|\boldsymbol{\theta}) = \tilde{a}_j^{(n)}(r|\boldsymbol{\theta}) / \tilde{a}_0^{(n)}(r|\boldsymbol{\theta})$$

where, for $j = 1, 2, ..., n$,

$$\tilde{a}_j^{(n)}(r|\boldsymbol{\theta}) = \sum_{(x,y)\in\mathcal{A}_j^{(n)}} \tilde{p}_{x,y}^{(n)}(r|\boldsymbol{\theta}) \quad \text{and}$$

$$\tilde{a}_0^{(n)}(r|\boldsymbol{\theta}) = r^{-1} - \sum_{y=1}^{n} \tilde{p}_{n-y,y}^{(n)}(r|\boldsymbol{\theta}).$$

Note that $|\mathcal{A}_j^{(n)}| = n + 1 - j \quad (j = 1, 2, ..., n)$. It follows from (4.6.10) that, for $n = 2, 3, ..., n_{max}$ and $j = 1, ..., n$,

$$|\tilde{a}_j^{(n)}(r|\boldsymbol{\theta}) - \tilde{a}_j^{(n)}(r'|\boldsymbol{\theta})| \leq (n + 1 - j)|r - r'|/(rr'). \tag{4.6.26}$$

Fix K such that K is a compact subset of $(0, \infty)^{n_{max}}$ and $\bar{\boldsymbol{\theta}} \in K$. It then follows from (4.6.23) and the continuity of $\tilde{a}_j^{(n)}(r|\boldsymbol{\theta})$ that for $n = 2, 3, ..., n_{max}$ and $j = 1, 2, ..., n$,

$$\max_{\boldsymbol{\theta}\in K} \check{h}_{n,j}^{(m)}(\boldsymbol{\theta}) \xrightarrow[A^c]{\text{a.s.}} 0 \quad \text{as } m \to \infty, \tag{4.6.27}$$

Further, (4.6.26) and the uniform continuity of $\log x$ imply that, for $n = 2, 3, ...,$ n_{max} and $j = 1, 2, ..., n$,

$$\max_{\boldsymbol{\theta}\in K} \hat{h}_{n,j}^{(m)}(\boldsymbol{\theta}) \xrightarrow[A^c]{\text{a.s.}} 0 \quad \text{as } m \to \infty, \tag{4.6.28}$$

since $\hat{r}^{(m)} \xrightarrow[A^c]{\text{a.s.}} r$ as $m \to \infty$. Similar to before, (4.6.24) implies that $\tilde{l}_{rec}^{(\infty)}(\boldsymbol{\theta}|r)$ has a unique global maximum at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. It follows using (4.6.25), (4.6.27) and (4.6.28), that, for any $a \in (0, \boldsymbol{\theta})$,

$$\operatorname*{argmax}_{\boldsymbol{\theta}\in K} \tilde{l}_{rec}^{(m)}(\boldsymbol{\theta}|\boldsymbol{Z}^{(m)}, \hat{r}^{(m)}) \xrightarrow[A^c]{\text{a.s.}} \bar{\boldsymbol{\theta}} \quad \text{as } m \to \infty. \tag{4.6.29}$$

The theorem follows.

□

Extending the proof of Theorem 4.6.3 to $K = (0, \infty)^{n_{max}}$ is more complicated than in the one-dimensional setting of Ball and Shaw [2015] and is not considered here. Similar results to the above also hold for SEIR and Reed-Frost models but are omitted here.

4.7 Discussion

We have demonstrated that, for an emerging epidemic, basing inference on the final size distribution of single-household epidemics usually leads to a biased estimate of local contact rates. A new estimator has been developed using the theory of Crump-Mode-Jagers branching processes which properly accounts for the dynamics of emerging epidemics in a population of households. The method has also been adapted to develop a similar estimator for discrete-time Reed-Frost epidemics and simulations have been used to show that these estimators have the potential to perform well in practice. This method assumes that an estimate of the exponential growth rate, r of a given epidemic is available. How best to estimate r is a challenge which remains open. It is also assumed that estimation is performed whilst an epidemic is in its exponentially growing phase and it should be checked that this assumption is reasonable.

Extending the proof of Theorem 4.6.3 to $K = (0, \infty)^{n_{max}}$ is perhaps the most obvious place to consider further research into the theory presented in this chapter. Other ideas to progress this work may include developing approximations to the Laplace transforms $\tilde{p}^{(n)}_{x,y}(r|\lambda^{(n)}_L)$ $(n, x, y) \in \mathcal{T}$ in order to relax the assumption of exponentially distributed latent periods and/or recovery rates that are needed to make the estimation described in Section 4.3 computationally feasible. This may be possible by adopting approaches similar to those given in Fraser [2007] or Pellis et al. [2011] for calculating r in the non-Markovian case.

It would also be useful to approximate standard errors of estimators using the method developed in this chapter, either using a parametric bootstrap or by determining the asymptotic distribution of the estimator. The latter would require central limit analogues of the results of Nerman [1981] that were exploited in Section 4.3. Standard cluster bootstrapping would not be appropriate for households epidemic data since it relies on the clusters within the data (in this case the outcomes in household of different sizes) behaving independently of each other. However, it may be possible to develop another version of the block bootstrap which accounts for the dependence between outcomes in different households that exists under our households epidemic model. Alternatively, Bayesian methods such as MCMC or ABC may be used to create credible intervals for estimators (see, for example, Cauchemez et al. [2004]) and then the properties of these intervals could be investigated from a frequentist perspective.

A far simpler progression to implement would be the extension of the method to the multitype case (see the model of Ball and Lyne [2001]), using the generalisations of Nerman [1981], in order to accommodate age or gender specific susceptibilities. The method can in principle also be extended to situations where information on the temporal progression of disease within households is available. This is discussed in greater detail in the concluding comments of Ball and Shaw [2015].

CHAPTER 5

Epidemics in vaccinated populations

This chapter is concerned with estimating the vaccination coverage required to prevent an epidemic from taking place using final size and emerging epidemic data and, as such, may be viewed as a practical application of the previous two chapters. In general, we consider a scenario in which some data are available for a given disease which we use to estimate its infectious contact parameters. We then wish to use the parameter estimates to adopt a vaccination strategy which will prevent a global outbreak of the same disease, either among the same population in future or in a nearby population. We also use this chapter to make further comparisons between the basic and Cauchemez models for local contact rate that were introduced in Chapter 3.

The structure of this chapter is as follows. We outline the post-vaccination epidemic model, vaccine action models and vaccination strategies, all obtained from the literature, in Section 5.1. Section 5.2 considers the notion of an optimal vaccination strategy in greater detail. In Section 5.3 we outline a procedure for estimating critical vaccination coverage and investigate the impact of an incorrect model choice when performing these estimations. The findings of the chapter are discussed briefly in Section 5.4.

5.1 Vaccination models and strategies

We introduce a post-vaccination threshold parameter, models for the effects of vaccination and vaccination strategies. This section follows a similar structure to Section 3 of Ball and Lyne [2006], which may be consulted for further details.

5.1.1 Post-vaccination threshold parameter

Consider the threshold parameter R_* introduced in Chapter 2. In the asymptotic case, as the number of households $m \to \infty$, the probability that an epidemic infects a strictly positive proportion of the population tends to zero if the threshold parameter $R_* \leq 1$. Therefore, if m is large, we can eradicate the possibility of a severe outbreak, by reducing the threshold parameter to 1 through vaccination.

First, let x_{nv} denote the proportion of households of size n containing v vaccinated individuals $(n = 1, 2, ..., n_{max};\ v = 0, 1, ..., n)$ and note that the probability of a global contact in the initial stages of an epidemic being with an individual in such a household is $\tilde{\alpha}_n x_{nv}$. In a vaccinated population, we consider a single-household epidemic to begin whenever any of the individuals in a fully susceptible household is contacted globally by an infective. This definition makes it possible for a single-household epidemic to have size 0 if the globally contacted individual has been vaccinated and is able to resist infection as a consequence. (Note that we now distinguish between the phrases "infectious contact" and "contact" since some contacts between an infected and a susceptible individual do not lead to the susceptible becoming infected.) Let $\mu_{n,a,v}(\lambda_L^{(n)})$ be the mean number of global contacts coming out of a single-household epidemic, under this definition, in a household of size n with v vaccinated individuals and a individuals contacted globally by infectives outside of the household. Then,

$$R_v = \sum_{n=1}^{n_{max}} \tilde{\alpha}_n \sum_{v=0}^{n} x_{nv} \mu_{n,1,v}(\lambda_L^{(n)}). \tag{5.1.1}$$

Note that (5.1.1) provides a generic post-vaccination parameter which can be used for any vaccine action model. To obtain a more specific post-vaccination threshold parameter for a given vaccine action model, we must consider the exact form of $\mu_{n,1,v}(\lambda_L^{(n)})$ under that model.

5.1.2 Vaccine action models

We consider two vaccine action models in this chapter. The first of these is the *all-or-nothing vaccine* (see, for example, Halloran et al. [1992] and Becker and Starczak [1998]). This vaccine renders its subjects completely immune to the disease, independently, with probability ϵ but otherwise has no effect.

Under the all-or-nothing vaccine, successfully vaccinated individuals are not at all susceptible, whereas unsuccessfully vaccinated individuals are fully susceptible. This implies that the number of susceptibles in a household of size n is k, where $n-k$ denotes the number of successful vaccinations that have taken place. The distribution of the number of successful vaccinations in a household with v vaccinated individuals follows a binomial distribution with v trials and probability ϵ of success. Hence, for $k = n-v, n-v+1, ..., n$, the probability that a household of size n with v vaccinated members contains exactly k susceptibles is $\binom{v}{n-k}\epsilon^{n-k}(1-\epsilon)^{v-n+k}$. Note that the probability that global contact with a household of size n with only k susceptibles starting a single-household epidemic is k/n. Thus, under the all-or-nothing vaccine,

$$\mu_{n,1,v}(\lambda_L^{(n)}) = \lambda_G \mathbb{E}[T_I] \sum_{k=n-v}^{n} \binom{v}{n-k} \epsilon^{n-k}(1-\epsilon)^{v-n+k} \frac{k}{n} \mu_{k,1}(\lambda_L^{(n)}), \qquad (5.1.2)$$

where $\mu_{k,1}(\lambda_L^{(n)})$ is as defined in Chapter 2. Note that although vaccination reduces the number of susceptibles in a household, the local contact rate is still dependent on the total household size, n. This is the key distinction of (5.1.2) from the post-vaccination threshold parameter for the all-or-nothing vaccine given by, for example, Ball and Lyne [2002b, 2006], which follow the basic local mixing model (see Chapter 3).

Our second vaccine model is the *non-random response vaccine*, which has a predetermined response and is a specific version of the vaccine model of Becker and Starczak [1998]. Following that paper, we consider a non-random vaccine to have an effect $(A, B) \in [0,1]^2$ on a given individual, where A and B denote the relative susceptibility and infectivity respectively of the individual in comparison to their unvaccinated state. This is to say that all infectious contact rates towards vaccinated individuals are multiplied by A and all infectious contact rates from a vaccinated individual are multiplied by B, should they become infected.

We now derive the $\mu_{n,1,v}(\lambda_L^{(n)})$ under the non-random response vaccine model, First, note that we effectively have a multitype epidemic under this model with two types of individuals. Following Ball and Lyne [2006], let unvaccinated individuals be referred to as type-1 and vaccinated individuals be referred to as type-2 and let $\mathbf{\Lambda}_L^{(n)} = [\lambda_L^{(n)}]_{ij}$ be the local infection rate matrix for a household of size n ($n = 2, 3, ..., n_{max}$). For a vaccine with response (a, b), we have

$$\boldsymbol{\Lambda}_L^{(n)} = \begin{pmatrix} \lambda_L^{(n)} & a\lambda_L^{(n)} \\ b\lambda_L^{(n)} & ab\lambda_L^{(n)} \end{pmatrix}.$$

In a size-n household with v vaccinees there are $n - v$ type-1 individuals and v type-2 individuals. Define the expected number of type-j individuals infected in a single household epidemic with one initial infective of type-i to be $\mu_{(n,v),i,j}(\lambda_L^{(n)})$. (Note that we now assume that the initial individual is indeed infected rather than just contacted.) If global contact is made with a type-2 individual in a fully susceptible household, a single household epidemic ensues with probability a. Similarly, the expected number of global contacts made by a type-2 individual is scaled by b. Hence

$$\begin{aligned}\mu_{n,1,v}(\lambda_L^{(n)}) = \Bigg(& \frac{n-v}{n}\left[\mu_{(n-v,v),1,1}(\lambda_L^{(n)}) + b\mu_{(n-v,v),1,2}(\lambda_L^{(n)})\right] \\ & + \frac{v}{n}a\left[\mu_{(n-v,v),2,1}(\lambda_L^{(n)}) + b\mu_{(n-v,v),2,2}(\lambda_L^{(n)})\right]\Bigg)\lambda_G\mathbb{E}[T_I].\end{aligned}$$

The $\mu_{(n,v),i,j}(\lambda_L^{(n)})$ may be calculated using the following method of Ball [1986] and notation of Ball and Lyne [2006]. For $i = 1, 2$ and $l_1, l_2 = 0, 1, 2, ...$, let $h_i(l_1, l_2) = \phi(l_1[\lambda_L^{(n)}]_{i1} + l_2[\lambda_L^{(n)}]_{i2})$, where $\phi(t) = \mathbb{E}[e^{-tT_I}]$, as given in Chapter 2. Then, for $i = 1, 2$ and $n_1, n_2 = 0, 1, 2, ...$, let $\beta_{n_1,n_2}^{(n)}$ be defined recursively by

$$\sum_{l_1=0}^{n_1}\sum_{l_2=0}^{n_2}\binom{n_1}{l_1}\binom{n_2}{l_2}\beta_{n_1,n_2}^{(n)}[h_i(l_1,l_2)]^{n_1-l_1}[h_i(l_1,l_2)]^{n_2-l_2} = n_i.$$

Then, for $n_1 = 1, 2, ..., n_2 = 0, 1, 2, ...$ and $i = 1, 2,,$

$$\begin{aligned}\mu_{(n_1,n_2),1,i}(\lambda_L^{(n)}) = n_i & \\ & - \sum_{l_1=0}^{n_1-1}\sum_{l_2=0}^{n_2}\binom{n_1-1}{l_1}\binom{n_2}{l_2}\beta_{n_1,n_2}^{(n)}[h_i(l_1,l_2)]^{n_1-l_1}[h_i(l_1,l_2)]^{n_2-l_2}\end{aligned}$$

and, for $n_1 = 0, 1, 2, ..., n_2 = 1, 2, ...$ and $i = 1, 2,,$

$$\begin{aligned}\mu_{(n_1,n_2),1,i}(\lambda_L^{(n)}) = n_i & \\ & - \sum_{l_1=0}^{n_1}\sum_{l_2=0}^{n_2-1}\binom{n_1}{l_1}\binom{n_2-1}{l_2}\beta_{n_1,n_2}^{(n)}[h_i(l_1,l_2)]^{n_1-l_1}[h_i(l_1,l_2)]^{n_2-l_2}.\end{aligned}$$

The generalised version of the non-random vaccine response model is that of a *discrete vaccine response* Becker and Starczak [1998]. Under this model, the vaccine response (A, B) is given by a pair of random variables with a distribution supported on finitely many points in $\mathbb{R}^2$

$$\mathbb{P}(A = a_i, B = b_i) = p_i \quad (i = 1, 2, ..., k),$$

for some finite k such that $\sum_{i=1}^{k} p_i = 1$. We do not consider this vaccine model in general, however, it should be noted that the all-or-nothing vaccine is a form of this model in which $k = 2$, $a_1 = 0$, $a_2 = b_2 = 1$, $p_1 = \epsilon$, $p_2 = 1 - \epsilon$ and b_1 is arbitrary (since successfully vaccinated individuals never become infective). The post-vaccination threshold parameter, R_v, may be derived for this model by extending the methodology used for the non-random vaccine response in the manner described in Section 3.2.3 of Ball and Lyne [2006].

The value $1 - \mathbb{E}[AB]$ gives a measure of how efficient a vaccine is and is referred to as vaccine *efficacy*. (Note that this is not the only definition available of vaccine efficacy, see Becker et al. [2006], but is the traditional measure.) From the above statements it is clear that $1 - AB$ is the efficacy of the non-random response vaccine and that ϵ is the efficacy of the all-or-nothing vaccine as defined above. For the remainder of this chapter we shall compare vaccines with the same efficacy and denote that efficacy by ϵ. In particular an all-or-nothing vaccine with efficacy ϵ will be compared to non-random response vaccines with $A = B = \sqrt{1 - \epsilon}$ and $A = 1 - \epsilon$, $B = 1$. The latter of these, in which $B = 1$ and thus vaccination takes no affect once an individual has become infected, is known as the *leaky vaccine*.

5.1.3 Vaccination strategies

Let c denote the proportion of individuals in the population that are to be vaccinated. We consider three potential vaccination strategies. The first of these is the *random individuals* strategy in which the individuals chosen to be vaccinated are chosen uniformly at random from the entire population. Under this strategy we approximate that, for $n = 1, 2, ..., n_{max}$; $v = 1, 2, ..., n$, $x_{nv} = \binom{n}{v} c^v (1 - c)^{n-v}$. Under the *random households* strategy, entire households are vaccinated which are again chosen uniformly, at random from the population. Here we approximate that $x_{nn} = c$ and $x_{n0} = 1 - c$. These approximations for the random individuals and random households vaccination strategies become exact as $m \to \infty$.

The final strategy to consider is the *optimal* vaccination strategy. A strategy is considered optimal in this context if it reduces the post-vaccination threshold parameter such that $R_v \leq 1$ by vaccinating as few people as possible. Note however that such a strategy may not be unique or may not exist. Ball et al. [2004a] and Ball and Lyne [2006] show that by letting $h_{nv} = m_n x_{nv}$ be the number of households of size n with v vaccinated members, $M_{n,v} = n\mu_{n,1,v}(\lambda_L^{(n)})/N$, the post-vaccination threshold parameter becomes

$$R_v = \sum_{n=1}^{n_{max}} \sum_{v=0}^{n} h_{nv} M_{n,v}. \tag{5.1.3}$$

This shows that R_v can be determined by assigning every household a value $M_{n,v}$, as defined above, and summing across all household values in the population. More importantly, the reduction in R_v from vaccinating an individual in a household of size n with v currently vaccinated individuals is given by $G_{n,v} = M_{n,v} - M_{n,v+1}$. Note that $G_{n,v}$ is always non-negative since the expected size of a single-household epidemic, $\mu_{nv}(\lambda_L^{(n)})$, cannot be increased by vaccinating an extra individual in the household. Also observe that

$$c = \sum_{n=1}^{n_{max}} \sum_{v=0}^{n} v h_{nv}.$$

Under any given vaccination strategy, the usual aim is to achieve $R_v \leq 1$ since this prevents any epidemic from taking off in the manner described in Chapter 2. Let $c_v^{(ind)}$, $c_v^{(house)}$ and $c_v^{(opt)}$ denote the critical vaccination coverage (CVC) for the random individuals, random households and optimal vaccination strategies respectively, where the CVC refers to the minimum proportion of individuals in a population that need to be vaccinated in order to achieve $R_v \leq 1$.

5.2 Forms of the optimal vaccination strategy

In this section we focus on the effect that model choice can have on the optimal vaccination strategy. Under most circumstances, the optimal strategy is to find the maximal $G_{n,v}$ such that $x_{nv} > 0$ (or equivalently $h_{nv} > 0$) and to vaccinate an individual in the associated household state. This reduces the threshold parameter R_v by $G_{n,v}$ and the process should continue until $R_v \leq 1$. The exception to this rule is the case where $G_{n,v}$ does not decrease as v increases for fixed n. Here the strategy needs modifying to account for the possibility that it may be optimal to vaccinate two or more individuals in the same household before moving on. Such circumstances occur when $G_{n,v} < G_{n,v+1}$ and hence $2G_{n,v} < G_{n,v} + G_{n,v+1}$, implying that the gain from vaccinating two individuals in a household of size n with v vaccinees is greater than that of vaccinating individuals in different households of size n with v vaccinees. This phenomenon is explored in more detail in Ball and Lyne [2002b] and Ball et al. [2004a].

Another special case to consider is that in which, for all n, $\lambda_L^{(n)} = 0$. This gives a homogeneously mixing population in which there is clearly no difference in the effect of vaccinating one individual over another and thus all strategies are equal. Specifically, for all n and $v = 0, 1, ..., n-1$, $G_{n,v} = \lambda_G \mathbb{E}[T_I]\epsilon/N$ under any of the vaccine models outlined in Section 5.1.2 with efficacy ϵ, where N is the population size. Therefore, the optimal vaccination strategy is driven by the local dynamics of an epidemic.

We now consider epidemics with non-zero local contact rates and shall ignore the global contact rate for the remainder of this section since it merely scales the lower-triangular gain matrix $\boldsymbol{G}$ (formed by the $G_{n,v}$ values) which determines the optimal strategy. As such, the remainder of this section assumes that $\lambda_G = 1$ and T_I takes a negative exponential distribution with rate 1. This is largely for ease of illustration, although it should be noted that the underlying distribution of T_I does effect the $G_{n,v}$ beyond simply rescaling. Similar results to those shown below may be found for other distributions of T_I, such as gamma or constant. We also restrict ourselves to the case $n_{max} = 5$ for illustrative convenience.

Table 5.1 shows the gain matrices for an epidemic under the basic model with $\lambda_L = 0.6$ (see Chapter 3) using four vaccines: a *perfect* vaccine ($\epsilon = 1$) and three vaccines with efficacy $\epsilon = 0.5$. These are an all-or-nothing, vaccine, a leaky vaccine and a non-random response vaccine with $A = B = \sqrt{0.5}$. Note that the action model of the perfect vaccine is irrelevant since it renders all vaccinees non-infective, either through loss of susceptibility or infectivity. It is assumed here that $\mathbb{P}(A = 0) = 1$ when referring to a perfect vaccine and thus the vaccinee is left fully immune to infection. For illustrative purposes, the $G_{n,v}$ are multiplied by the unspecified population size N and bracketed superscripts are used to rank the values from highest to lowest for ease of reading off the optimal strategy.

Ball et al. [1997] proposed that the optimal vaccination strategy under the basic model with a perfect vaccine takes a form known as the equalising strategy. Under this strategy, households with the largest number of unvaccinated individuals are targeted for further vaccination. Under an imperfect vaccine, Ball and Lyne [2002a] outline a generalisation to a strategy which they refer to as the *conditional equalising strategy* and describe as targeting "households with the largest expected number of susceptibles" for further vaccination (although it is not obvious exactly how to interpret this explanation for a non-random response vaccine). As discussed above, optimal vaccination strategies work by curtailing the local dynamics of an epidemic and the idea under these strategy is to minimise the expected number of susceptibles in larger households so that the expected number of susceptibles in local groups is as equal as possible. The $\epsilon = 1$ portion of Table 5.1 provides a simple illustration of this strategy. Under a perfect vaccine, vaccinated individuals retain no susceptibility and thus the number of susceptibles in their household is effectively reduced by one post-vaccination. With this in mind, the table clearly shows that the optimal vaccination strategy at any given time is to vaccinate a single individual in any of the households containing the largest possible number of unvaccinated individuals. Note also that when $\epsilon = 1$, $G_{n,v}$ depends only on the value of $n - v$.

If the vaccine is not fully effective, as in the three vaccines with efficacy $\epsilon = 0.5$ shown in Table 5.1, the exact form of the optimal vaccination strategy is not as immediately obvious. However, on closer inspection, the idea of conditional equalisation still holds under the basic model. We can still see from the table that optimal strategy attempts to effectively reduce the number of susceptibles in the larger households, as per the explanation by Ball and Neal [2002]. The difference with a less effective all-or-nothing vaccine is that vaccinated individuals may still be susceptible and thus one must vaccinate a greater number of individuals in the larger households in order to effectively reduce their size. Observe also that the $G_{n,v}$ values are all smaller under the less effective vaccine. These values represent a relative reduction in R_v by vaccinating a given indi-

Table 5.1: Gain matrices for a basic model epidemic, with $\lambda_L = 0.6$ under a perfect vaccine and all-or-nothing, non-random $A = B$ and leaky vaccines with efficacy $\epsilon = 0.5$. The superscripts rank the gains from highest to lowest for ease of reading off the optimal vaccination strategy

n	v=0	v=1	v=2	v=3	v=4
	Perfect, $\epsilon = 1$				
1	$1.0000^{(11)}$				
2	$1.7500^{(7)}$	$1.0000^{(11)}$			
3	$2.8835^{(4)}$	$1.7500^{(7)}$	$1.0000^{(11)}$		
4	$4.4267^{(2)}$	$2.8835^{(4)}$	$1.7500^{(7)}$	$1.0000^{(11)}$	
5	$6.3334^{(1)}$	$4.4267^{(2)}$	$2.8835^{(4)}$	$1.7500^{(7)}$	$1.0000^{(11)}$
	All-or-nothing, $\epsilon = 0.5$				
1	$0.5000^{(15)}$				
2	$0.8750^{(13)}$	$0.6875^{(14)}$			
3	$1.4418^{(9)}$	$1.1584^{(11)}$	$0.9229^{(12)}$		
4	$2.2134^{(4)}$	$1.8276^{(6)}$	$1.4930^{(8)}$	$1.2080^{(10)}$	
5	$3.1667^{(1)}$	$2.6900^{(2)}$	$2.2588^{(3)}$	$1.8759^{(5)}$	$1.5419^{(7)}$
	Non-random, $A = B = \sqrt{0.5}$				
1	$0.5000^{(15)}$				
2	$0.8287^{(13)}$	$0.6905^{(14)}$			
3	$1.2429^{(10)}$	$1.1179^{(11)}$	$0.9805^{(12)}$		
4	$1.7458^{(6)}$	$1.6262^{(7)}$	$1.5046^{(8)}$	$1.3747^{(9)}$	
5	$2.3221^{(1)}$	$2.2035^{(2)}$	$2.0915^{(3)}$	$1.9820^{(4)}$	$1.8690^{(5)}$
	Leaky, $A = 0.5$				
1	$0.5000^{(15)}$				
2	$0.8317^{(13)}$	$0.6875^{(14)}$			
3	$1.2969^{(9)}$	$1.1096^{(11)}$	$0.9348^{(12)}$		
4	$1.8829^{(4)}$	$1.6659^{(6)}$	$1.4534^{(8)}$	$1.2492^{(10)}$	
5	$2.5512^{(1)}$	$2.3255^{(2)}$	$2.0948^{(3)}$	$1.8629^{(5)}$	$1.6338^{(7)}$

vidual and thus it is clear that from a mathematical viewpoint that a greater number of individuals need to be vaccinated to achieve the CVC under less effective vaccines. This observation is trivial but nonetheless reassuring.

The non-random vaccines shown in Table 5.1 illustrate these points even more clearly. The non-random response $A = B = \sqrt{0.5}$ vaccine displays an optimal strategy in which it is always better to vaccinate individuals in larger households if possible. This even occurs if all but one individuals in a household are vaccinated and there are other households in the population with just one fewer individual residing in them and no vaccinees. Other than in households of size 1, the $G_{n,v}$ are all smaller, suggesting that the non-random response $A = B$ vaccine is actually less effective than an all-or-nothing vaccine of the same efficacy when a vaccinated individual mixes locally as well as globally. Note, however, that the optimal strategy still takes a conditional equalising form under this vaccine model, since larger households are still targeted first. They simply require greater vaccination coverage to reduce their overall susceptibility than under the all-or-nothing vaccine.

The leaky vaccine displays $G_{n,v}$ values which are generally smaller than for the all-or-nothing vaccine, as proven formally in Ball et al. [2004a], but larger than the non-random $A = B$ vaccine of the same efficacy. It shares the same optimal strategy as the all-or-nothing vaccine for these particular parameter values, although differences in some relative gains such as $G_{4,1} > G_{5,4}$ are less pronounced than for the all-or-nothing vaccine, suggesting that the leaky vaccine may display an optimal strategy more similar to that of the non-random $A = B$ model under different parameter choices.

The optimal vaccination strategy has the potential to deviate from conditional equalisation under the model presented in this thesis in which the local contact rate varies with n. In particular, one would expect the form of the optimal strategy to vary under the Cauchemez model (again, see Chapter 3) if η is sufficiently large, since the local contact rate is more significant in smaller households under this model. Table 5.2 shows gain matrices from Cauchemez model epidemics with $\lambda_L = 0.2$ and $\lambda_L = 4$, both with $\eta = 1$, so for all n, $\lambda_L^{(n)} = \lambda_L / n$ in each epidemic. An all-or-nothing vaccine with $\epsilon = 0.75$ and a non-random response vaccine with $A = B = 0.5$ are considered in both cases.

Deviation from the conditional equalising strategy can be seen immediately in

Table 5.2: Gain matrices and optimal strategies for Cauchemez model epidemics, with $\eta = 1$. We consider epidemics with high and low local infectious rates, an all-or-nothing vaccine with $\epsilon = 0.75$ and a non-random response vaccine with $A = B = 0.5$

n	v=0	v=1	v=2	v=3	v=4
			All-or-nothing, $\lambda_L = 0.2$		
1	$0.7500^{(15)}$				
2	$0.8864^{(7)}$	$0.7841^{(14)}$			
3	$0.9530^{(4)}$	$0.8711^{(8)}$	$0.7978^{(13)}$		
4	$0.9935^{(2)}$	$0.9250^{(5)}$	$0.8624^{(9)}$	$0.8053^{(12)}$	
5	$1.0210^{(1)}$	$0.9620^{(3)}$	$0.9074^{(6)}$	$0.8569^{(10)}$	$0.8101^{(11)}$
			All-or-nothing, $\lambda_L = 4$		
1	$0.7500^{(15)}$				
2	$1.7500^{(10)}$	$1.000^{(14)}$			
3	$2.8650^{(6)}$	$1.9216^{(9)}$	$1.2036^{(13)}$		
4	$4.0625^{(3)}$	$2.9844^{(5)}$	$2.0820^{(8)}$	$1.3818^{(12)}$	
5	$5.3212^{(1)}$	$4.1476^{(2)}$	$3.1054^{(4)}$	$2.2322^{(7)}$	$1.5434^{(11)}$
			Non-random, $\lambda_L = 0.2$		
1	$0.7500^{(15)}$				
2	$0.8842^{(8)}$	$0.7854^{(14)}$			
3	$0.9361^{(4)}$	$0.8778^{(10)}$	$0.8068^{(13)}$		
4	$0.9571^{(2)}$	$0.9238^{(5)}$	$0.8797^{(9)}$	$0.8242^{(12)}$	
5	$0.9626^{(1)}$	$0.9466^{(3)}$	$0.9210^{(6)}$	$0.8856^{(7)}$	$0.8400^{(11)}$
			Non-random, $\lambda_L = 4$		
1	$0.7500^{(15)}$				
2	$1.5833^{(12)}$	$1.0833^{(14)}$			
3	$2.4392^{(7)}$	$1.9283^{(10)}$	$1.4141^{(13)}$		
4	$3.3192^{(3)}$	$2.7920^{(5)}$	$2.2757^{(8)}$	$1.7592^{(11)}$	
5	$4.2202^{(1)}$	$3.6763^{(2)}$	$3.1488^{(4)}$	$2.6350^{(6)}$	$2.1235^{(9)}$

the $\lambda_L = 0.2$ gain matrices. Under an imperfect vaccine, conditional equalising suggests that it is better to vaccinate an individual in a size-n household with v vaccinated individuals than an individual in a household of size $n - v + 1$ with no vaccinated individuals. Other than when $n = v + 1$, the opposite is shown to be true for our $\lambda_L = 0.2$ epidemics under the all-or-nothing vaccine and is generally the case for the non-random response $A = B$ vaccine. The exception under the non-random response vaccine is among households in which two unvaccinated individuals remain.

For our epidemics with higher local contact rates, in which $\lambda_L = 4$, we see that the conditional equalising still holds as the optimal vaccination strategy. However, by recalling that a perfect vaccine effectively reduces household size by 1 when used, it is clear that an optimal strategy similar to that for the all-or-nothing vaccine $\lambda_L = 0.2$ epidemic would have been seen had we considered a vaccine with $\epsilon = 1$. Under an imperfect vaccine, vaccinated individuals may still become infected locally and the potential additional local infectious contact that they bring to their household can be more significant to the spread of an epidemic in a household than the increased local contact rate in a smaller household. Thus the optimal strategy may still continue to conform to conditional equalising, despite the change in how local contact rates are modelled. When λ_L is sufficiently large, the model essentially becomes the highly-infectious model of Becker and Dietz [1995] (see also Becker and Starczak [1997]) for which conditional equalising is the optimal vaccination strategy (see Ball and Lyne [2002b]).

The optimal vaccination strategy under the non-random response vaccine in the $\lambda_L = 0.2$ epidemic also deviates from that under the all-or-nothing vaccine in the same epidemic and that which we could expect from a perfect vaccine under the Cauchemez model in the case where two unvaccinated individuals remain in all households. Here, the optimal strategy under the non-random response vaccine advocates vaccinating individuals in households of size 5 before those in households of size $2, 3$ or 4. The explanation for this is similar to that as to why conditional equalisation still holds under the Cauchemez model for large enough λ_L. We may recall from Table 5.1 that vaccinating people in larger households is of greater importance under the non-random response vaccine than the all-or-nothing vaccine, particularly in the case where $A = B$. Note however that all four values in the gain matrix for households with two unvaccinated individuals are extremely close. Therefore, any deviation from the optimal vaccination strategy may not be too problematic in practice in terms of achieving CVC that is close to the optimal value.

We observe that deviation in the optimal vaccination strategy from conditional equalisation under the Cauchemez model is most likely to occur under larger η and ϵ, smaller λ_L (cf. Keeling and Ross [2015]) and, if vaccine efficacy is known, assuming that the available vaccine has a non-random response with $A = B$. However, given that smaller λ_L gives an epidemic closer to the homogeneously mixing case in which there is no optimal vaccination strategy this suggests that conditional equalisation may still be effective in practice, even under a Cauchemez model. This is investigated in greater detail in Section 5.3.

5.3 Estimating critical vaccination coverage

5.3.1 General approach

We attempt to estimate the vaccination coverage required to prevent a major outbreak of a disease by first estimating the parameters associated with the epidemic using the maximum pseudolikelihood methods outlined in Section 3.2 for final size data and in Section 4.3 for emerging epidemic data. In the emerging case we assume that information on infectives are available and thus use the full-pseudolikelihood method. It is assumed that all the necessary conditions set out in these sections and in Chapter 2 are satisfied. In particular, we assume that there is no latent period when dealing with emerging epidemic data (recall that this assumption is not required for final size data). We also assume that T_I follows a negative exponential distribution. In the case of emerging data, this distribution has an unknown rate γ (see Section 4.3.2). Recall from Chapter 2.1 that we may assume, without loss of generality, that $\mathbb{E}[T_I] = 1$ when dealing with final size data. We also assume that the vaccine response is known (all-or-nothing or non-random and the associated parameters). There is a wealth of information available on estimating vaccine efficacy, Longini and Halloran [1996] and Longini et al. [1998], for example, give estimation methods for a generalised form of the all-or-nothing vaccine outlined in this chapter. Becker et al. [2006] offer procedures for estimating the efficacy of a discrete response vaccine which, as we have already noted in Section 5.1.2, is a generalisation of both of our vaccine action models.

Once estimates are made of all the unknown parameters of an epidemic, it is possible to compute the pre-vaccination threshold parameter R_*. If $R_* \leq 1$ then no further action is needed since the epidemic is already sub-critical. If $R_* > 1$ then it is necessary to vaccinate members of the population and values are needed for the parameters x_{nv} $(n = 1, 2, ..., n_{max};\ v = 0, 1, ..., n)$ in order to compute the post-vaccination threshold parameter R_v, as outlined in Section 5.1. The $x_{n,v}$ are determined by the vaccination coverage c for the random households and random individuals strategies. For the optimal strategy, the $x_{n,v}$ are determined by the estimates of the $\lambda_L^{(n)}$ as well as c. (Note that λ_G and $\mathbb{E}[T_I]$ merely scale the $G_{n,v}$ values that determine the optimal strategy). Since ϵ and the x_{nv} are considered to be either estimated or known, we need only find estimates for λ_G and the $\lambda_L^{(n)}$ from final size data in order to estimate the critical vaccination coverage for an epidemic using (5.1.3). When dealing with emerging epidemic data, we begin by estimating the real-time growth rate r by using the method outlined in Section 4.5.1 in which we fit a straight line to the logarithm of the number of recoveries. The $\lambda_L^{(n)}$ and γ may then be estimated by using the maximum pseudolikelihood approach of Section 4.3. Finally, an estimate of λ_G can be obtained by using (4.3.3).

5.3.2 Simulation study

We use the methods outlined in Section 5.3.1 to illustrate the effects of estimating epidemics governed by the Cauchemez model with the simpler and more widely used basic model. In particular, we wish to assess the accuracy of using the basic model to estimate critical vaccination coverage for the three vaccination strategies outlined in Section 5.1.3 for epidemics with dynamics governed by the Cauchemez model. To achieve this, epidemic simulations are performed to generate emerging epidemic data and final size data.

Let $\boldsymbol{\theta} = (\lambda_G, \lambda_L, \eta, \gamma)$ be a vector denoting the unknown parameters of a given epidemic under the Cauchemez model. The data from the simulated epidemics that take off are used to determine one of four possible estimators for $\boldsymbol{\theta}$. These are denoted by $\hat{\boldsymbol{\theta}}_{EME}^{(Cauch)}$, $\hat{\boldsymbol{\theta}}_{FIN}^{(Cauch)}$, $\hat{\boldsymbol{\theta}}_{EME}^{(basic)}$ and $\hat{\boldsymbol{\theta}}_{FIN}^{(basic)}$. The subscripts EME and

FIN refer to estimators made from emerging and final size data respectively (and so no estimate of γ is made for the latter case) and the superscripts $(Cauch)$ and $(basic)$ refer to the Cauchemez and basic models respectively. For the basic model estimators, $\hat{\eta}$ is fixed at 0. An estimator for the CVC may be calculated from each of these four estimators using the methods of Section 5.1.3.

The parameter choices for the epidemic used in the simulation study are as follows. As outlined in Section 5.3.1, the infectious period, T_I, is set to have an exponential distribution with rate $\gamma = 1$ and hence $\mathbb{E}[T_I] = 1$. Two population distributions are considered to reflect different household structures in different parts of the world. The first, $\boldsymbol{\alpha}_{UK} = [0.36, 0.30, 0.15, 0.10, 0.05, 0.04]$, represents a typical city in Western Europe and is based on data taken from the 2011 UK census for the urban area of Nottingham (Office for National Statstics [2011]). Due to the lack of households of size greater than 6, all households meeting this criterion have been truncated to be of size 6 for the sake of convenience.

The second structure, $\boldsymbol{\alpha}_{Ghana} = [0.20, 0.15, 0.15, 0.14, 0.12, 0.09, 0.05, 0.10]$, represents a typical city in West Africa and is taken from the combined urban data of the 2010 Ghanaian census (Ghana Statistical Service [2012]). All households of size 8 or above are truncated to size 8, again for the sake of convenience. The parameters governing disease transmission are set to be $\lambda_G = 0.8$, $\lambda_L = 2$ and $\eta = 1$. These values are chosen such that approximately 50% of an unvaccinated population becomes infected if the epidemic takes off under the $\boldsymbol{\alpha}_{UK}$ population structure (see Ball et al. [2010a], Ferguson et al. [2005]). For comparison, the pre-vaccination threshold parameter $R_* = 1.57$ under the $\boldsymbol{\alpha}_{UK}$ population structure and $R_* = 2.20$ under $\boldsymbol{\alpha}_{Ghana}$ for epidemics with parameters as outlined above. The value of $\eta = 1$ follows the suggestion of Cauchemez et al. [2004] and Chapter 3.

Consider the following vaccines with efficacy $\epsilon = 0.84$: an all-or-nothing vaccine, a leaky vaccine ($A = 0.16$) and a non-random response vaccine with $A = B = 0.4$. For the epidemic outlined above, Table 5.3 gives the true CVC for the epidemic outlined above under each vaccine action model, each of our three vaccination strategies and both the UK and Ghanaian population structures. Note from Table 5.3 that the CVC values are generally larger under the Ghanaian population structure which contains larger households and thus has potentially larger local outbreaks. The CVC is also smallest under the optimal

Table 5.3: Critical vaccination coverage (CVC) for the epidemic outlined in this simulation study under different population structures, vaccine action models (all with efficacy $\epsilon = 0.84$) and both the UK and Ghanaian population structures, as outlined in the main text

Population structure	Vaccine action model	$c_v^{(ind)}$	$c_v^{(house)}$	$c_v^{(opt)}$
UK	All-or-nothing, $\epsilon = 0.84$	0.2988	0.3976	0.2113
	Leaky, $A = 0.16$	0.3100	0.3998	0.2225
	Non-random, $A = B = 0.4$	0.3211	0.3998	0.2323
Ghana	All-or-nothing, $\epsilon = 0.84$	0.4269	0.5861	0.3492
	Leaky, $A = 0.16$	0.4430	0.5889	0.3643
	Non-random, $A = B = 0.4$	0.4684	0.5889	0.3903

strategy (as one would hope), followed by the random individuals strategy and finally the random households strategy is least effective. We also see that the all-or-nothing vaccine performs better than the leaky vaccine which in turn generally outperforms the non-random response $A = B$ vaccine, even though all three vaccines have the same efficacy. These observations are reassuring given previous results in the literature (see, for example, Ball et al. [2004a] and Ball and Lyne [2006]) and those seen in Section 5.2.

However, the two non-random response vaccines (Leaky and $A = B$) perform equally well under the random households strategy. In a household of size n in which every individual has been vaccinated with a non-random response vaccine which has effect (A, B), infectious contacts between a susceptible and infective occur at rate $AB\lambda_L^{(n)}$. Thus, any two non-random response vaccines with the same efficacy will perform equally well under a random households strategy, since we have already established in Section 5.2 that only the vaccine efficacy ϵ affects the gain of vaccinating a given individual in terms of curtailing global infectious contacts.

Before looking at simulation studies, we should take account of the fact that, in practice, implementing the optimal vaccination strategy relies on having knowledge of λ_L and η. Specifically, the discussion in Section 5.2 shows that an estimate of η that is far enough away from its true value can lead to incorrect guess as to the form of the optimal vaccination strategy. As such, we now introduce another vaccination strategy which will be referred to as the *fitted optimal* strategy. This scheme uses the estimated parameters of an epidemic rather than the unknown true parameters to determine an "optimal" vaccination strategy. Thus, the fitted optimal strategy is not necessarily optimal.

Let $c_v^{(\overline{opt})}$ denote the CVC for an epidemic under the fitted optimal vaccination strategy and $\hat{c}_v^{(\overline{opt})}$ be its estimator. Also, let $\hat{c}_v^{(ind)}$ and $\hat{c}_v^{(house)}$ be estimators of the CVC under the random individuals and random households strategies respectively. Note that Table 5.3 does not contain true values for the fitted optimal strategy since it is dependent on observed data and thus is a random variable.

We simulate final size data for a population of $m = 500$ households for the UK population structure and $m = 300$ households for the Ghanaian population structure so that the total population size is similar for both populations. ($N = 1150$ for the UK structured population whilst $N = 1020$ for the Ghanian structured population.) In both cases 1000 epidemics were simulated, with the data used to give estimates $\hat{\boldsymbol{\theta}}_{FIN}^{(Cauch)}$ and $\hat{\boldsymbol{\theta}}_{FIN}^{(basic)}$, which were then used to give estimates of $c_v^{(ind)}$, $c_v^{(house)}$ and $c_v^{(\overline{opt})}$ for each of the vaccine action models considered in Table 5.3.

As explained in Chapter 4, estimating epidemic parameters from emerging data is only reliable if the population is large enough such that there is a point in the epidemic at which the proliferation of infected households still resembles the branching process outlined in Chapter 2. However, we must also ensure that there are enough infected households to give a reliable estimate of θ_{EME}. As such, we consider epidemics in a population of $m = 10000$ households ($N = 23000$) for the UK population structure and $m = 6000$ households ($N = 20400$) for the Ghanaian population structure, both of which are used to provide emerging epidemic data after 500 recoveries have been observed, (see Section 4.5.3). An estimate $\hat{r}$ of the real-time growth rate is made as described in Section 5.3.1 by using the polyfit function in MATLAB and ignoring the first 20 recoveries (also see Section 4.5.1). Again 1000 epidemics are simulated and the data used give CVC estimates for each strategy under the basic and Cauchemez models.

For this simulation study, we focus on the value $\hat{c}_v - c_v$, i.e. the difference between estimated CVC and its true value. If this value is positive, then the population would be over-vaccinated if the estimated CVC is used, potentially

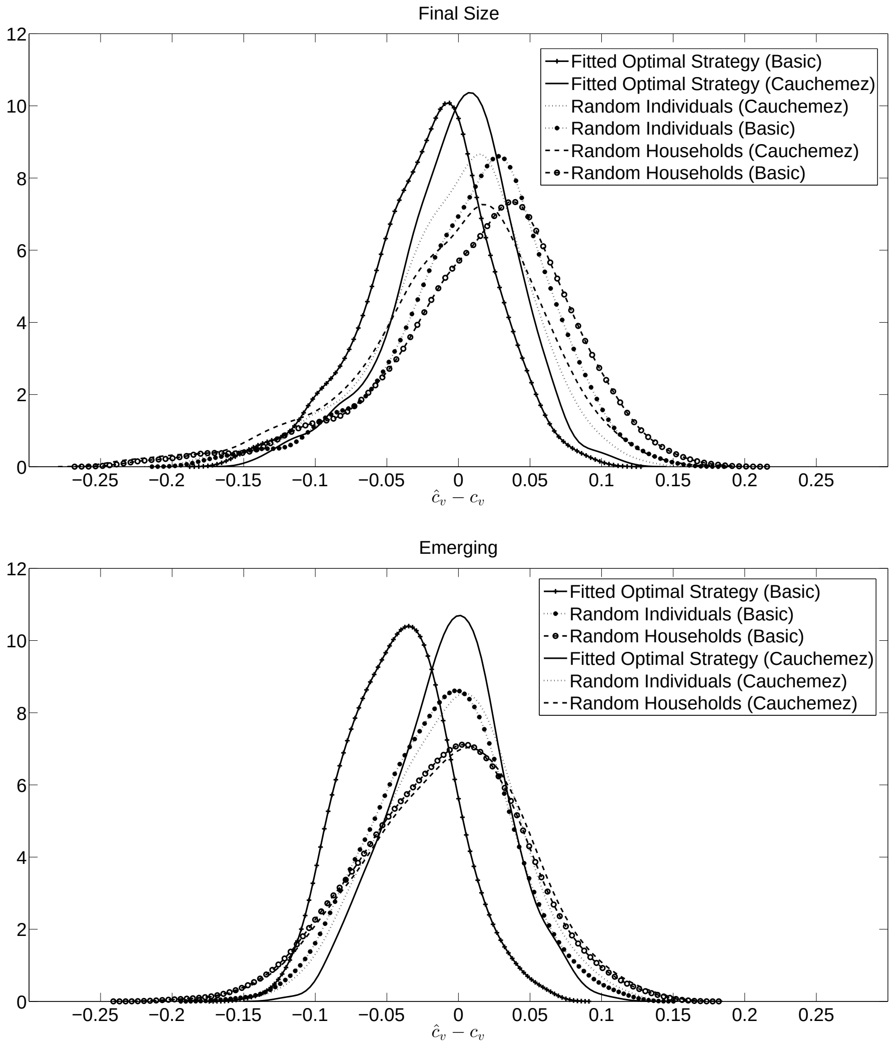

Figure 5.1: Kernel density estimates of $\hat{c}_v - c_v$ for each of our vaccination strategies. These plots are based on 1000 simulations of the epidemic outlined in this section for the UK population structure and use an all-or-nothing vaccine with efficacy $\epsilon = 0.84$. We consider both final size data and emerging epidemic data and estimate the CVC assuming both a basic and Cauchemez model for the epidemics. The true CVC values were $c_v^{(ind)} = 0.2988$, $c_v^{(house)} = 0.3976$ and $c_v^{(\overline{opt})}$ took values in the range $[0.2113, 0.2160]$

wasting valuable resources. If it is negative then the population will be under-vaccinated, leaving it vulnerable to a global outbreak. In Figure 5.1 we illustrate how choice of vaccination strategy under our all-or-nothing vaccine affects the potential to under/over-vaccinate a population for our epidemic, using both our simulated final size and emerging epidemic data from the population structure $\boldsymbol{\alpha}_{UK}$, by giving kernel density estimates of $\hat{c}_v - c_v$ from our 1000 epidemic simulations.

As one would expect given that we simulated epidemics from the Cauchemez model, the distribution of $\hat{c}_v - c_v$ is centred around zero for the fitted optimal, random individuals and random households strategies if the Cauchemez model is used to estimate the unknown parameters of the epidemic and this is observed from both the final size and emerging epidemic simulations. If a basic model is used for parameter estimation, both our final size and emerging plots suggest that the fitted optimal strategy is more likely to under-vaccinate the population. The distribution of CVC estimates for random individuals and random households strategies are both still loosely centred around the true CVC value when the basic model is used for parameter estimation. The plots suggest that over-vaccination may be more likely under these strategies if CVC estimates are made using the basic model from final size data.

We also observe that the variation of the estimators $\hat{\boldsymbol{\theta}}^{(basic)}_{FIN}$, $\hat{\boldsymbol{\theta}}^{(Cauch)}_{FIN}$, $\hat{\boldsymbol{\theta}}^{(basic)}_{EME}$ and $\hat{\theta}^{(Cauch)}_{EME}$ can all lead to rather large errors in CVC estimators. These errors are greater under the random individuals and random households strategies although this may be attributed to the fact that these strategies generally require more individuals to be vaccinated than the fitted optimal strategy so this is not entirely unexpected. (Generally speaking, variance of $\hat{c}_v$ is reduced when c is closer to 0 or 1.) It is also worth noting that the variance of the estimators of $\boldsymbol{\theta}$ reduces as the number of households m increases and so this is a less problematic issue if large enough data are available (cf. Chapters 3 and 4). We discuss this further in Section 5.3.3.

Figure 5.1 only shows kernel density estimate of the probability density function of the random variable $\hat{c}_v - c_v$ under the all-or-nothing vaccine. We now turn our attention to comparing our three vaccine action models, as given in Table 5.3. Kernel density estimates of the distribution of $\hat{c}_v^{(\overline{opt})} - c_v^{(\overline{opt})}$ for our epidemic under the $\boldsymbol{\alpha}_{UK}$ population structure are given in Figure 5.2 for each

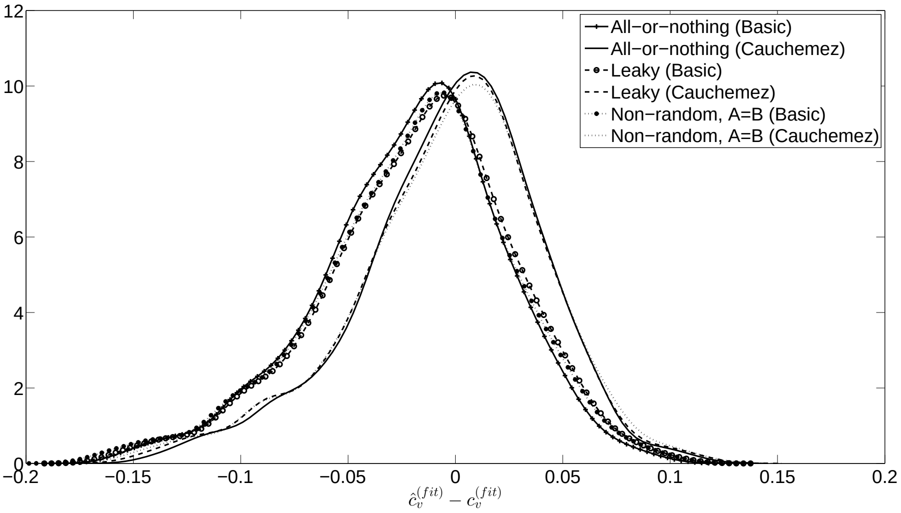

Figure 5.2: Kernel density estimates $\hat{c}_v^{(\overline{opt})} - c_v^{(\overline{opt})}$ for the 1000 final size data simulations of the epidemic outlined in this section under the UK population structure. Plots are shown for estimates based on assumption of a basic model and a Cauchemez model for local epidemic dynamics and three vaccine action models are considered, each with efficacy $\epsilon = 0.84$. These are the all-or-nothing vaccine, a leaky vaccine and a non-random response vaccine with $A = B = 0.4$. The true CVC under the fitted optimal strategy, $c_v^{(\overline{opt})}$, took values in the range $[0.2113, 0.2160]$ for the all-or-nothing vaccine, $[0.2225, 0.2275]$ for the leaky vaccine and $[0.2323, 0.2379]$ under the non-random response vaccine with $A = B$

of our vaccine action models, using final size data.

Despite the differences between the three vaccine action models that we have observed in Section 5.2 and Table 5.3, we observe that the three vaccine models exhibit very similar behaviour with respect to the distribution of $\hat{c}_v^{(\overline{opt})} - c_v^{(\overline{opt})}$ under each vaccine model. Figure 5.2 shows this to be the case whether the correct Cauchemez model or incorrect basic model are used to estimate $\boldsymbol{\theta}$. Similar plots for emerging epidemic data and alternative vaccination strategies are omitted but show similar results. Thus we conclude that whilst the vaccine action model affects the value of the CVC under any given vaccination strategy/population structure/data type, it does not appear to have much bearing on the probability of one under/over-estimating the CVC, if the methods outlined in this chapter are used.

Finally, we use our simulation study to consider the effects of population structure when estimating the CVC for an epidemic. Kernel density estimates of the distribution of $\hat{c}_v^{(\overline{opt})} - c_v^{(\overline{opt})}$ for our epidemic under the all-or-nothing vaccine are given in Figure 5.3 for both the α_{UK} and α_{Ghana} population structures. In Section 5.3.3 we discuss how population structure affects whether one would expect to over-estimate or under-estimate the CVC if an incorrect model choice for local contact rates is assumed. However, from this figure, we see that the variance of $\hat{c}_v^{(\overline{opt})} - c_v^{(\overline{opt})}$ appears to be greater under the complex Ghanaian population structure than the simpler UK structure, especially when final size data are used. Similar plots under other vaccination models and strategies reveal a similar trend and thus are omitted.

The illustrations from this simulation study show that highly inaccurate CVC estimates are plausible for realistic sizes of data, even if the correct model is chosen. This is particularly likely if there is significant variety in the household sizes within a population and if the more practical random households or random individuals vaccination strategies are selected, rather than attempting to find an optimal vaccination strategy. In the following section, we investigate exactly when parameterising a Cauchemez model epidemic using the basic model would be expected to lead to the most severe cases of under/over-estimation of the CVC and also illustrate the effects of under-vaccination in terms of the expected final outcome if an epidemic in an under-vaccinated population takes off.

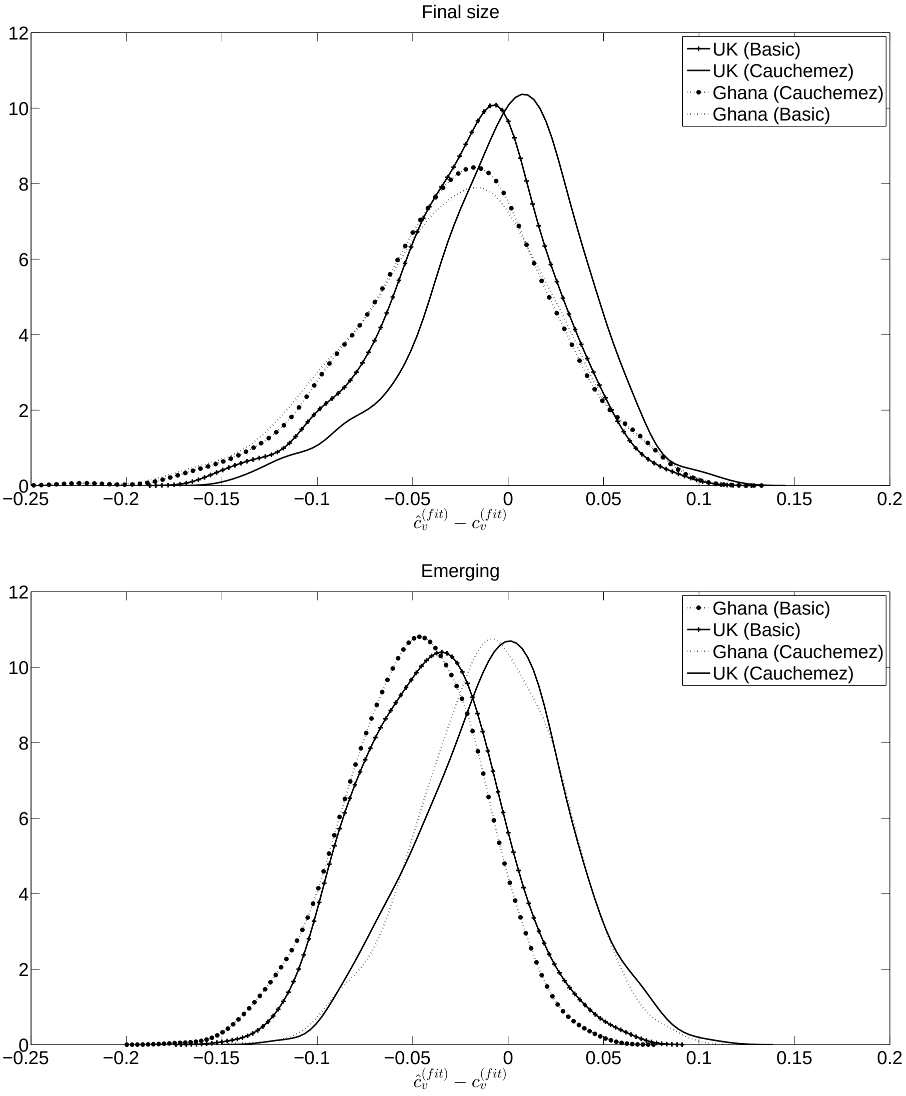

Figure 5.3: Kernel density estimates of $\hat{c}_v^{(\overline{opt})} - c_v^{(\overline{opt})}$ based on 1000 simulations of the outlined epidemic under the α_{UK} and α_{Ghana} population structures, using both the basic and Cauchemez models to estimate the epidemic parameters. Estimates are based on an all-or-nothing vaccine with efficacy, $\epsilon = 0.84$. The true CVC under the fitted optimal strategy, $c_v^{(\overline{opt})}$, took values in the range $[0.2113, 0.2160]$ under α_{UK} and $[0.3492, 0.3669]$ under α_{Ghana}

5.3.3 Perfect data analysis

The variances of the estimators $\hat{\theta}_{FIN}$ and $\hat{\theta}_{EME}$ decrease as the population size increases. Whilst the simulation study presented in Section 5.3.2 suggests that there is little difference between using the basic and Cauchemez models when estimating CVC, this may not be the case if data from a larger population are available. Here we consider data from an infinite population, as described in Section 3.1 for final size data and in Section 4.5.2 for emerging epidemic data. We wish to ascertain the circumstances under which CVC estimates using the basic model are at their least accurate, ignoring the variance of $\hat{\theta}_{FIN}$ and $\hat{\theta}_{EME}$. Perhaps more importantly, we also consider the expected final outcome of epidemics in which the population has been under-vaccinated as a result of using the basic model.

As in the simulation study, we assume the specific form of the Cauchemez model in which $\eta = 1$, unless stated otherwise. For the sake of convenience, we only consider perfect vaccines in this section ($\epsilon = 1$), apart from in Figure 5.7 in which we also consider imperfect all-or-nothing vaccines.

Figure 5.4 shows the effect of changing the global contact rate for the epidemic outlined in Section 5.3.2. The plot is given in terms of changing R_* (a value which has a one-to-one correspondence with the value of λ_G, as discussed at the end of Chapter 2). This offers an illustrative advantage in that the lower bound of $R_* = 1$, which must be exceeded for any epidemic to take off with non-zero probability, is consistent for epidemics with any parameters, making it easier to compare epidemics with different population structures. It also allows us to consider epidemics of similar severity under different population structures. All three vaccination strategies exhibit a similar pattern for both the $\boldsymbol{\alpha}_{UK}$ and $\boldsymbol{\alpha}_{Ghana}$ population structures and both types of observation in that the estimated CVC under the simple model is close to the true coverage required for very small R_* and tends back towards the true coverage required as R_* gets very large. This can be explained by the local contact rate (in which the basic model differs from the true Cauchemez model), becoming a less important factor in the epidemic as λ_G increases and the epidemic in question becomes globally driven.

Beyond this however, one can observe subtle differences between how the pop-

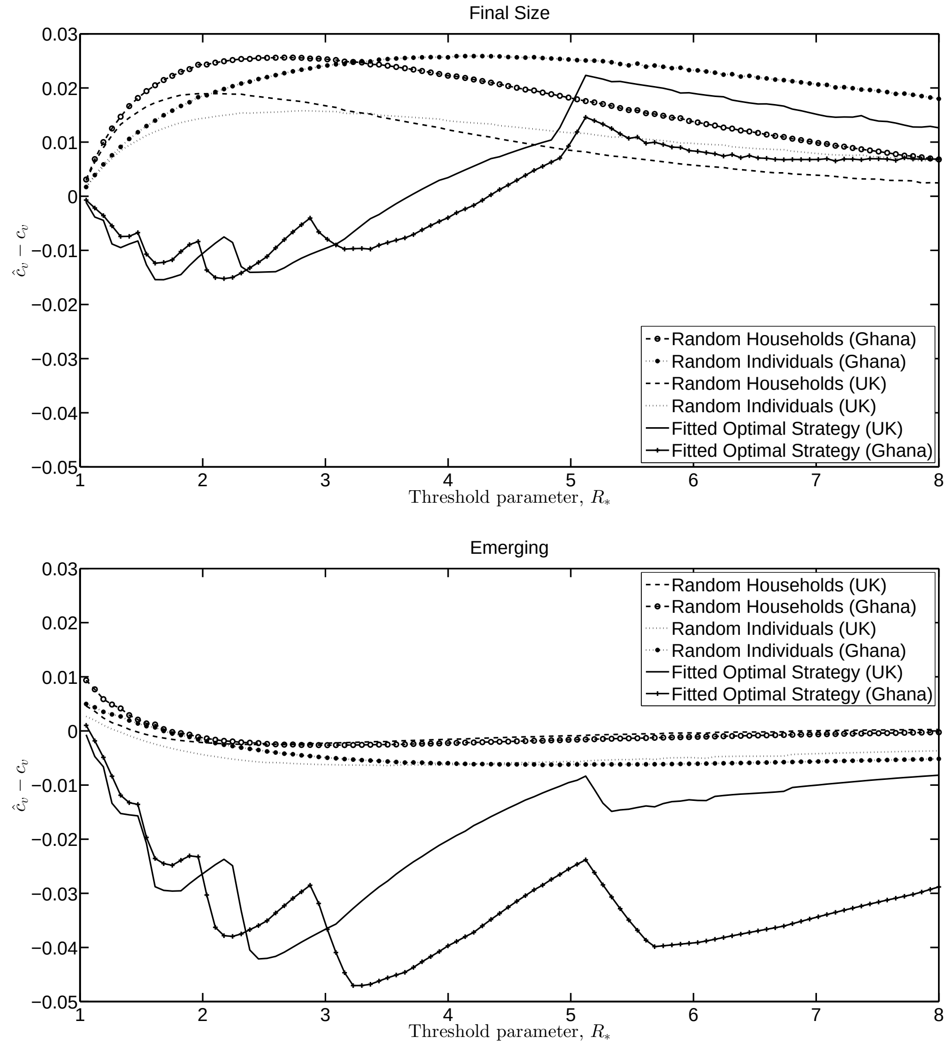

Figure 5.4: Plots showing how $\hat{c}_v - c_v$ changes as the threshold parameter R_* of an epidemic increases. Perfect data is assumed and the basic model (assuming $\eta = 0$) is used to find $\hat{c}_v$ for each of the possible vaccination strategies. The true values of the local contact rates for households of size n are fixed at $\lambda_L^{(n)} = 3/n$. We assume a perfect vaccine (i.e. $\epsilon = 1$). The true value of c_v is approximately 0 when R_* is close to 1. When $R_* = 8$, $c_v = 0.7215$, 0.7724 and 0.8750 under the UK population structure for the true optimal, random individuals and random households strategies respectively. Under the Ghanaian structure, these values are $c_v = 0.6427$, 0.7155 and 0.8750

ulation structure and type of data available affects the accuracy of critical vaccination coverage estimates under the simple model. For the emerging epidemic data, the random households and random individuals critical vaccination coverage estimates from the simple model appear to perform at a similar level regardless of the population distribution but under final size data, estimates of critical vaccination coverage from the $\boldsymbol{\alpha}_{Ghana}$ population model are seen to be less accurate than estimates from the $\boldsymbol{\alpha}_{UK}$ model, in which the majority of the population reside in smaller households. The most distinctive plots in Figure 5.4 however are those relating to the fitted optimal strategy. In particular, there are intervals on all four fitted optimal strategy plots in during which the estimated critical vaccination coverage diverges sharply from the true value. These intervals occur when the increased severity of the epidemic is combatted by vaccinating individuals in the largest households, since the simple model predicts a far greater gain from vaccinating such individuals than is achieved under the true model (c.f. Section 5.2). As individuals in smaller households are vaccinated to combat the increase in λ_G (or R_*), the CVC estimates recover towards the true value. The exception to this rule is severe epidemics under final size data, for which Figure 5.4 shows that the CVC is generally overestimated. Here the sharp divergences from the true critical value occurs when a high proportion of individuals in smaller households (size-2) are vaccinated and the simple model underestimates the gain from vaccinating these individuals.

It is important to note the range of values that c_v takes for each population structure and vaccination strategy shown in Figure 5.4 as R_* increases. As one would expect, the range is wide since we consider epidemics which barely take off and those which would be expected to infect an extremely high proportion of the population without intervention. The true value of c_v is approximately 0 when R_* is close to 1. When $R_* = 8$, $c_v = 0.7215, 0.7724$ and 0.8750 under the UK population structure for the true optimal, random individuals and random households strategies respectively. Under the Ghanaian structure, these values are $c_v = 0.6427, 0.7155$ and 0.8750. Thus the scale of the errors shown in CVC estimates shown in Figure 5.4 vary drastically and this should be borne in mind when observing the figure.

However, we also observe that $c_v^{(house)}$ is the same for both population structures when $R_* = 8$. Specifically, under a perfect vaccine, $c_v^{(house)} = 1 - 1/R_*$ regardless of the population structure (mirroring the CVC under a perfect vaccine of $1 - 1/R_0$ for a homogeneously mixing population). Although not explicitly stated, this is illustrated within Figure 2 of Ball and Lyne [2006].

Similar plots to Figure 5.4 in which λ_L is adjusted under the Cauchemez model and the threshold parameter $R_* = 2$ is fixed are given in Figure 5.5. As one may expect, the CVC estimate under all strategies and both population structures is extremely accurate when λ_L is small and the epidemic is almost exclusively globally driven. As local contact rates increase, the difference between the true and estimated models take effect and the critical vaccination coverage estimates diverge from the true value. As the local contact rate gets particularly large however, the CVC estimate becomes more accurate again. Consider, for example, the final epidemic considered in the plot for which $\lambda_L = 10$ and hence $\lambda_L^{(8)} = 1.25$. The expected size of a single-household epidemic in an unvaccinated household of size 8 is given by $\mu_8(1.25) = 7.15$. This suggests that almost every individual is expected to be infected in a single-household epidemic and hence even if the basic model gives a considerable overestimate of λ_L, it cannot drastically overestimate any of the $\mu_{nv}(\lambda_L^{(n)})$ values which determine the CVC.

Note again that in Figure 5.5, as in Figure 5.4, the true CVC values vary across the x-axis and thus the scale of the errors shown in the plot also varies. When $\lambda_L = 0$, all three vaccination strategies are equivalent and $c_v = 0.5$ under both population structures. When $\lambda_L = 10$, $c_v = 0.2231$ and 0.3414 under the UK population structure for the true optimal and random individuals strategies respectively. Under the Ghanaian structure, these values are $c_v = 0.2246$ and 0.3109. Once again, this should be borne in mind when considering the scale of the errors of the CVC estimates given in this figure.

When estimating vaccination coverage, we have stated that our intention is to achieve $R_v \leq 1$, since this eradicates the possibility of an epidemic taking off in a large enough population. Suppose we vaccinate some members of the population but too few to reach critical coverage. It is of interest to know the expected proportion of individuals in the population, z_v, that will become infected by the epidemic under these circumstances, if the epidemic takes off. We now look to assess the impact of using the basic model to estimate CVC for epidemics with different values of η under the fitted optimal vaccination strategy, which we

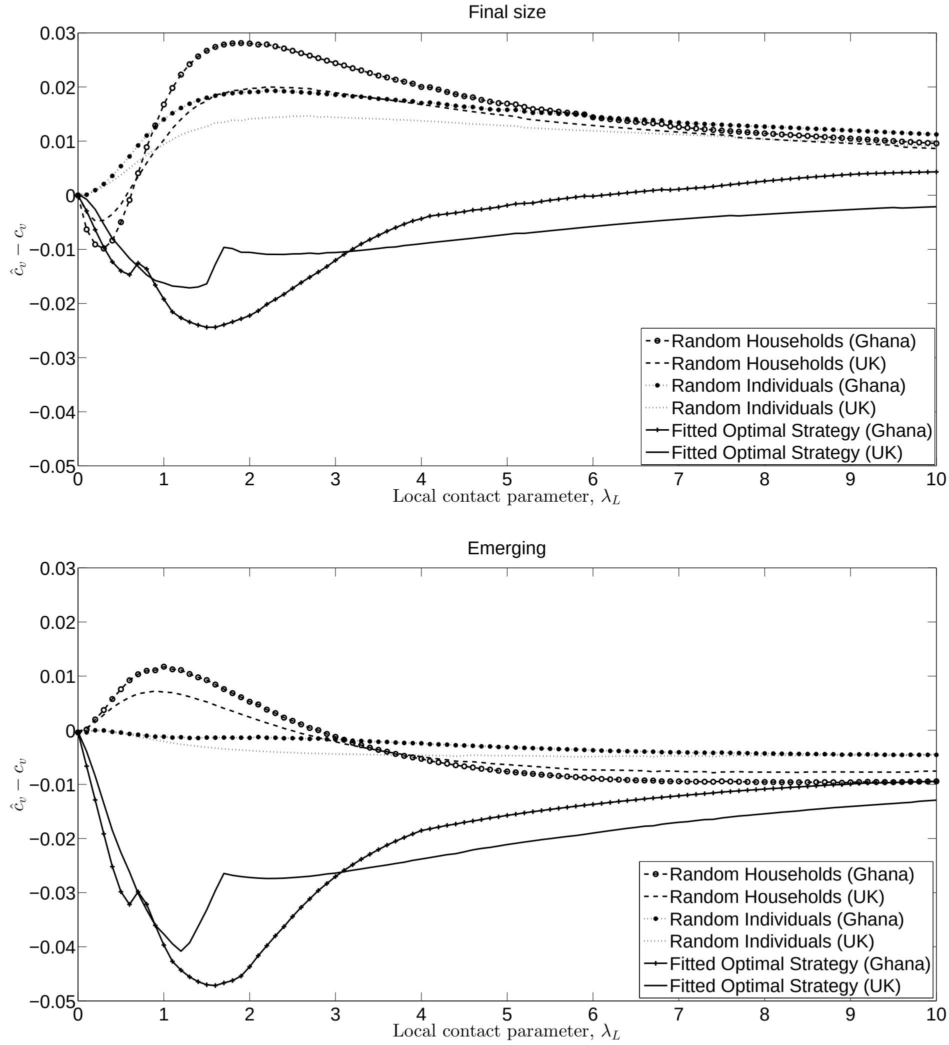

Figure 5.5: Plots showing how $\hat{c}_v - c_v$ changes as the local contact parameter λ_L of an epidemic increases if a perfect vaccine is used. The threshold parameter is fixed at $R_* = 2$ and thus $c_v^{(house)} = 0.5$ regardless of the population structure or value of λ_L used. When $\lambda_L = 0$, all three vaccination strategies are equivalent and $c_v = 0.5$ under both population structures. When $\lambda_L = 10$, $c_v = 0.2231$ and 0.3414 under the UK population structure for the true optimal and random individuals strategies respectively. Under the Ghanaian structure, these values are $c_v = 0.2246$ and 0.3109

have seen generally underestimates c_v.

For ease of illustration, we only consider the all-or-nothing vaccine for the remainder of this section. Recall from Chapter 2 that (2.3.3) and (2.3.4) give an implicit equation for z in an unvaccinated population. We now look to adapt these equations to the case when some members of the population have been vaccinated using an all-or-nothing vaccine. Let π_v be the probability that a given individual avoids global contact from any infective over the course of an epidemic, including contacts that do not result in infection as a result of our given individual having been successfully vaccinated. Then, following the same logic used to determine (2.3.3) in Section 2.3,

$$\pi_v = [\exp(-\lambda_G \mathbb{E}[T_I]/N)]^{Nz_v} = \exp(-\lambda_G z \mathbb{E}[T_I]). \tag{5.3.1}$$

Note that, under an all-or-nothing vaccine, the expected number of people that ultimately become infected by a single household epidemic in a household of size n, with v vaccinated individuals and a individuals contacted globally by infectives outside of the household, is given by

$$\mu_{n,a,v}(\lambda_L^{(n)})/\lambda_G \mathbb{E}[T_I]. \tag{5.3.2}$$

Thus, by once again considering the arguments used in Section 2.3, we obtain

$$z_v = (\lambda_G \mathbb{E}[T_I])^{-1} \sum_{n=1}^{n_{max}} n^{-1} \tilde{\alpha}_n \sum_{v=0}^{n} \sum_{a=1}^{n} x_{nv} \binom{n}{a} (1-\pi)^a \pi^{n-a} \mu_{n,a,v}(\lambda_L^{(n)}), \tag{5.3.3}$$

Equations (5.3.1) and (5.3.3) now give an implicit solution for z_v under the all-or-nothing vaccine. As with the pre-vaccination version, we are interested in the second solution, for which $z_v \in (0,1)$ and which only exists when $R_v > 1$, as this determines the expected proportion of individuals that become infected if the epidemic takes off. Note that these methods cannot be extended to other vaccine action models since (5.3.2) does not necessarily hold. However, calculation of z_v under the non-random response vaccine, or indeed its discrete response generalisation, is possible using the multitype epidemic model methods of Ball and Lyne [2001].

As the value of η increases, the less accurate the CVC estimator provided by the basic model should become. Hence, if we are estimating $c_v^{(opt)}$, we expect to underestimate $c_v^{(opt)}$ more severely as η increases and thus we expect z_v to increase with η. We have also observed, in Figures 5.4 and 5.5, that regardless of whether final size or emerging epidemic data are available, the basic model generally provides its least accurate estimators when the epidemic severity and local contact rates are large but not to the extent that global/local outbreaks are almost certain to occur following an initial infectious presence. If R_* is very large, the epidemic becomes globally driven and thus the local dynamics we consider here play a less important role in the spread of the outbreak. If λ_L is very large then single household epidemics are likely to infect everyone in the household who is not fully immune to the disease, thus rendering the differences between adopting a basic or Cauchemez model negligible. The exception to this rule is if vaccination coverage is also very large (MMR, for example, has both large local contact rates and a high vaccination coverage), in which case the difference between the models may become non-negligible.

Plots depicting the expected proportion of individuals infected in populations vaccinated according to estimates of the CVC using the fitted optimal strategy from the basic model for different values of η are depicted in Figure 5.6. For each value of η used, the value of λ_L is determined by setting $\lambda_G = 1$ and $R_* = 1.95$ for the UK population structure and $R_* = 2.75$ under the Ghanaian structure. This sets $\lambda_L = 2$ when $\eta = 1$ which Figure 5.5 shows to be approximately the level at which basic model estimation performs worst under the fitted optimal vaccination strategy.

Figure 5.6 shows an increase in z_v as η increases however, it is the values that z_v takes that are most interesting. Even at the $\eta = 1$ level suggested by Cauchemez et al. [2004], $z_v \approx 0.1$ if emerging epidemic data are used. This seems rather large for a population that has supposedly been vaccinated well enough to prevent an epidemic from taking off. The parameter values and vaccination strategy for Figure 5.6 were deliberately chosen as an extreme, but realistic, example of under-vaccination given perfect data. The random individuals and random households strategies are generally expected to over-vaccinate (in comparison to the CVC) for epidmeic models with $\eta > 0$ and thus their plots are not included here since $z_v = 0$ when over-vaccination occurs.

This point is illustrated further in Figure 5.7 in which we consider the specific case of $\eta = 1$ from the epidemics considered in Figure 5.6. Again, the plot shows the expected proportions of individuals that become that become in-

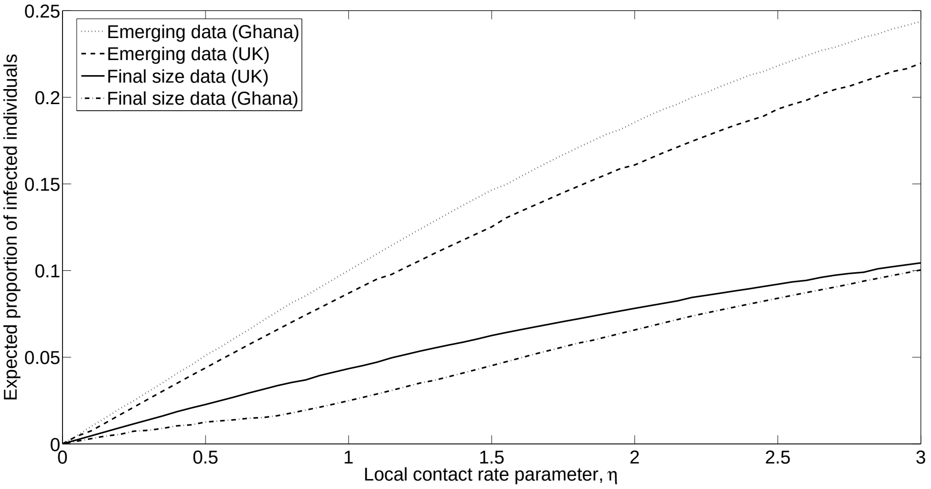

Figure 5.6: Plots depicting z_v in epidemics that have been vaccinated using the fitted optimal vaccination strategy for the basic model for increasing values of the local contact rate parameter η. The value $\lambda_G = 1$ is fixed along with $R_* = 1.95$ for plots based on the UK population structure and $R_* = 2.75$ for the Ghanaian structure. The vaccine is assumed to be perfect

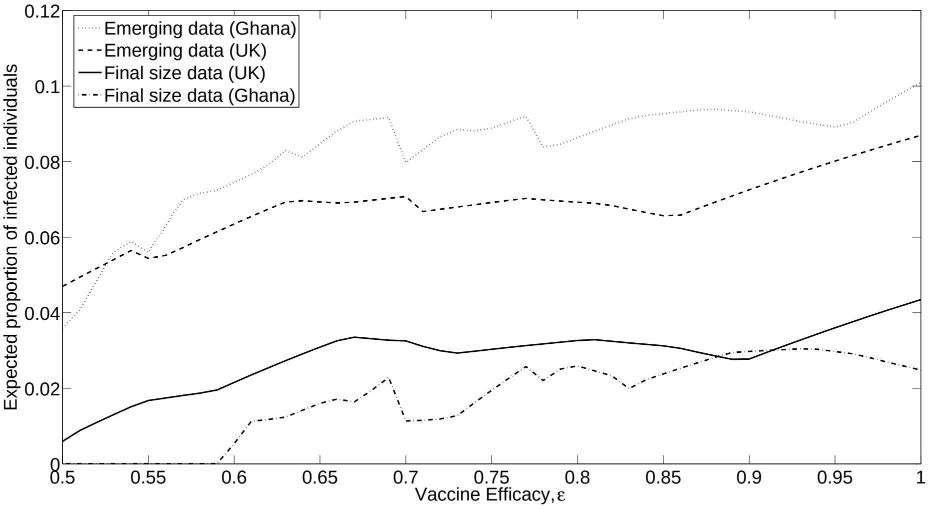

Figure 5.7: Plots depicting z_v in epidemics that have been vaccinated using the fitted optimal vaccination strategy for the basic model under all-or-nothing vaccines with different efficacies. The epidemic parameters are $\lambda_G = 1$, $\lambda_L = 2$ and $\eta = 1$

fected in a population vaccinated using the fitted optimal strategy from the basic model but by using all-or-nothing vaccines of varying effectiveness. In general, we observe that more effective vaccines can lead to more severe under vaccination in terms of the eventual final outcome of an epidemic if an incorrect model choice for the local dynamics of an epidemic is used.

It can be seen from both figures that emerging data leads to a more severe under vaccination in the circumstances set out here. From Figure 5.6 we observe that the extent of under/over-vaccination and its effects are far less predictable under the Ghanaian structure than the UK population structure. Small changes in vaccine efficacy are more likely to result in changes in the form of the fitted optimal strategy in population with a greater variety of household sizes and this affects the extent to which one under/over-vaccinates the population. This explains both the greater ranges and increased complexity of the curves in Figure 5.6 relating to the Ghanaian population structure compared to those representing the UK structure. Plots for non-random vaccines yield similar results to the above and thus are omitted.

Simulations from Section 5.3.2 show that under-vaccination is possible under any vaccination strategy and vaccine action model. Figures 5.6 and 5.7 should therefore serve as a warning that as accurate data as possible is needed to estimate parameters associated with the spread of an epidemic, since seemingly marginal under-vaccination can still lead to a reasonably severe epidemic taking place. Note also that whilst we observe the fitted optimal strategy being more prone to under vaccinate if one mistakenly assumes a basic model over a Cauchemez model, it is intuitive to note that the random individuals and, in particular, the random households strategies are more likely to under vaccinate if the converse holds and η is assumed to be strictly positive. (Recall from the Figures 5.5 and 5.4 that the random households strategy is particularly prone to over-vaccination in the current setting.)

5.4 Discussion

We have investigated how model selection when analysing epidemic data affects estimates of the vaccination coverage that would be required to prevent a future outbreak of the same epidemic from taking off. In particular we found that if practical vaccination strategies (random individuals and random households) are used, then estimating the parameters of an epidemic under the more widely used basic model generally provides a good approximation for critical vaccination coverage required should the actual epidemic have more closely resembled a Cauchemez model with $\eta = 1$, (see Chapter 3 and Cauchemez et al. [2004]). If anything, we observe that overestimation of critical vaccination coverage may be more likely in these circumstances. Model selection becomes a greater issue if one plans on developing an optimal vaccination strategy and we show that, even if a large amount of data are available from a previous outbreak, critical vaccination coverage estimates made using the basic model could under-vaccinate a population to the extent that an epidemic may still take off and a sizeable proportion of the population become infected.

We also show that the true form of the optimal vaccination strategy can differ significantly under the Cauchemez model from that of the conditional equalisation, which has been hypothesised to be the optimal strategy under the basic model (see Ball and Lyne [2002b]) for the widely used all-or-nothing and non-random response vaccine action models outlined in this thesis. Deviation from this strategy has been shown to be possible under both of these vaccine models however we show that it is particularly likely to occur under the Cauchemez model if the vaccine in question is highly effective and that this deviation occurs more readily under an all-or-nothing vaccine than a non-random response vaccine of the same efficacy.

Epidemic models in which the optimal vaccination strategy differs from conditional equalisation have also been considered by Keeling and Shattock [2012], who note that in non-interacting communities, small vaccine stockpiles which are not great enough to achieve critical vaccination coverage in any of the communities should be focussed on the smallest populations first. Also, Keeling and Ross [2015] consider a households model similar to that presented in this thesis, in which the within household transmission rate is dependent upon household size. They find that large maximal household sizes and small within household transmission rates are most likely to break the assumption of conditional equalisation for the optimal vaccination strategy.

From the simulation studies, it is clear that the variability of critical vaccination coverage estimates is far more likely to be the cause of under/over-vaccination using the random households or random individuals vaccination strategies. As such it would be useful to attach standard errors to our estimators of critical vaccination coverage. This would be attainable by attaching errors to parameter estimates using the methods of Chapter 3 and considering CVC as a function of $\boldsymbol{\theta}$. This is considered by Ball et al. [2004b] under the setting of Ball and Lyne [2001] and thus may be adapted to our model.

By attaching standard errors to CVC estimators, one may then consider the effects of vaccinating at the upper bound of some confidence interval for the CVC. A similar study to Section 5.3 could then be carried out to assess both how often vaccination set with some margin for error may leave an epidemic subcritical and what the effects of this may be if an incorrect model was chosen estimating CVC. The methods presented in this chapter may also be extended to the discrete vaccine response model or any other vaccine action models and thus this may also be an area for potential future research.

CHAPTER 6

Concluding comments

Results relating to inference from a stochastic SIR epidemic among a populations of households using both emerging and final size data from outbreaks have been developed. Here we give a brief summary of the results obtained in the thesis, ideas for extending this work and offer some general comments on the future of the field.

We have investigated the use of three different ways of modelling the within-household contact rate in a population:

1. **The basic model** in which the local contact rate is independent of household size.

2. **The Cauchemez model** in which the local contact rate varies with household size with respect to a parameter η (see Cauchemez et al. [2004]).

3. **The unrestricted model** in which local contact rates in households of different sizes are independent of each other.

Since these models are nested, we have presented all of our theory in terms of the unrestricted model which was set out in detail in Chapter 2.

In Chapter 3 we used the central limit theorem of Ball and Lyne [2001] to present theory for performing hypothesis tests based on maximum pseudolikelihood estimates of epidemic parameters from final size data. The tests were given in a general setting however we placed a specific focus on tests which could be used to select an appropriate epidemic model from those listed above based on final size households data obtained from a given outbreak. In particular, we showed that performing hypothesis tests to select one of these models does not require knowledge of the proportion of households in the population that have been sampled in the data. We also provided details on calculating the covariance matrices required to perform these tests.

This theory was illustrated using real life final size data obtained from influenza outbreaks. Our numerical studies in Section 3.6 suggested that the extra simplicity afforded by the basic or Cauchemez models may often provide a better fit to observed data than the full model. However, this study was only used to illustrate our theory since the data were relatively small and only gave information for influenza. As such it would seem to be ill-advised to dismiss any of the models outlined above without performing suitable hypothesis testing on any given households epidemic data that became available.

The use of hypothesis testing as a method for model selection was justified in Section 3.7. We stated that other popular model selection tools such as AIC, BIC and cross-validation all rely on data points in a sample being independent and that their asymptotic properties are unknown for data such as households epidemic data where dependence is weak but nonetheless exists. Further investigation into the properties of these methods in this scenario would prove beneficial for any future research focussing on attaching a “best-fitting” model to a stochastic epidemic with more than one level of mixing.

In Chapter 4 it was demonstrated that using the final size distribution of a single household epidemic generally results in obtaining a biased estimator for the within household infection rate of an emerging epidemic. We used branching process theory to develop an estimator which correctly accounts for the true nature of an emerging epidemic and showed that this estimator is strongly consistent. Using similar theory, we also provided an estimator for the local contact probability for data obtained from emerging Reed-Frost epidemics among a population of households. The estimator was also shown to be applicable whether infective and recovered or only recovered individuals in an emerging epidemic.

Simulation studies were carried out to illustrate that the derived estimator has the potential to perform well when applied to a real life data set and these were followed by a series of numerical illustrations depicting the bias of estimators obtained using the final size distribution of a single household epidemic. In order to be of practical use, the estimator relies on there being no latent period for the epidemic,individuals in the population having an exponentially distributed infectious period T_I and having an estimator of the exponential growth rate r of an epidemic available. We have shown that the assumption of no latent period and an exponentially distributed recovery rate may be relaxed using the phase method and assuming that one or both of these may be considered to have a J-stage Erlang distribution which is made up of J independently distributed exponentially distributed durations. However, this could become difficult to implement computationally if J is large.

The problems of how best to estimate r and approximating the Laplace transforms $\tilde{p}_{x,y}^{(n)}(r|\lambda_L^{(n)})$ $(n, x, y) \in \mathcal{T}$ (see Section 4.3) for non-phase-exponentially distributed latent periods/recovery rates are also open, although we have indicated in Section 4.7 the latter may be possible by adopting approaches similar to those given in Fraser [2007] or Pellis et al. [2011] for calculating r in the non-Markovian case. Approximating the standard error of the estimator derived in Section 4.3 is another potentially key area of future research. In Section 4.7 we have suggested that computationally intensive Bayesian methods such as ABC or MCMC may be used to calculate credible intervals for the standard error and this may be the most realistic method. Alternatively, bootstrapping may be possible but an alternative to the cluster bootstrap would have to be developed since this method relies on data from households of different sizes being independent. A final possibility would be to determine the asymptotic distribution of the estimator which would require central limit analogues of the results of Nerman [1981] that were exploited in Section 4.3.

In Chapter 5 we discussed the estimation of critical vaccination coverage for epidemics among households using emerging and final size data. In particular, we investigated how the form of the optimal vaccination strategy can vary under the three specific models outlined above and how incorrect model selection can lead to expected over/under-vaccination a population.

Simulation studies in this chapter showed that the variance of an estimator for critical vaccination coverage is the most likely cause of under/over-vaccination of a population if one uses a random individuals or random households strategy to vaccinate the population rather than the optimal vaccination strategy. As such, research should be made into attaching standard errors to estimators of critical vaccination coverage, which is attainable using the methods of Chapter 3 by considering the critical vaccination coverage as a function of the parameters of an epidemic. These could then be used to create confidence intervals for critical vaccination coverage under a given strategy. A further extension would be to extend the methods of Chapter 5 to the discrete vaccine response model which is a generalisation of the non-random and all-or-nothing vaccine action models used in that chapter.

In a wider context, it would also be fruitful to conduct further investigation using real data as to which of the households epidemic models outlined above best encapsulates the dynamics of the spreading of various diseases in which there appears to be increased levels of mixing at household level. It should also be possible to incorporate the methods used throughout this thesis into more general/complex models, such as the network epidemic model, the households-workplace model (Ball and Neal [2002]) or a model in which global infectious pressure does not increase linearly with the number of infectives (O'Neill and Wen [2012]).

The general future of the epidemiology field was discussed in great detail in the *Challenges in modelling infectious disease dynamics* edition of the Epidemics journal in March 2015. The households model presented in this thesis falls under the banner of the metapopulation models discussed by Ball et al. [2015] in their contribution to this journal. Their suggestions for the future development of the model include improving the theory for endemic diseases under the household structure (e.g. the SIS, Susceptible → Infective → Susceptible model), generalising theory to more complex population structures (such as the households-workplace model), developing inferential methods for emerging epidemics and improving the efficiency of computational methods used to calculate growth rates and threshold parameters. It is hoped that the work in this thesis has provided a contribution towards improving the understanding of the theory and the tools available for inference in epidemics among households, particularly in the emerging epidemic setting.

References

H. Abbey. An examination of the Reed-Frost theory of epidemics. *Human Biology*, 24(3):201–233, 1952.

C. L. Addy, I. M. Longini, and M. Haber. A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, 47(3):961–974, 1991.

B. Ainseba and M. Iannelli. Optimal screening in structured SIR epidemics. *Mathematical Modelling of Natural Phenomena*, 7(3):12–27, 2012.

D. J. Aldous. Exchangeability and related topics. *Springer Lecture Notes in Mathematics*, 1117:1:198, 1985.

H Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis*. Springer Science and Business Media, 2000.

J. Arino and S. Portet. Epidemiological implications of mobility between a large urban centre and smaller satellite cities. *Journal of Mathematical Biology*, 71(5): 1243–1265, 2015.

S. Asmussen. *Applied Probability and Queues*. Wiley, New York, 1987.

K. B. Athreya and P. E. Ney. *Branching Processes*. Springer-Verlag, Berlin, 1972.

N. T. J. Bailey. The total size of a general stochastic epidemic. *Biometrika*, 40 (1-2):177–185, 1953.

N. T. J. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. Mathematics in Medicine Series. Griffin, 1975.

F. G. Ball. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Advances in Applied Probability*, 18(2):289–310, 1986.

F. G. Ball and O. D. Lyne. Parameter estimation for SIR epidemics in households. *Bulletin of the International Statistical Institution. 52nd Session Contributed Papers, Tome LVIII, Book*, 2:251–252, 1999.

F. G. Ball and O. D. Lyne. Stochastic multitype SIR epidemics among a population partitioned into households. *Advances in Applied Probability*, 33(1):99–123, 2001.

F. G. Ball and O. D. Lyne. Epidemics among a population of households. In *Mathematical Approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory*, volume 126, pages 115–142. Springer, 2002a.

F. G. Ball and O. D. Lyne. Optimal vaccination policies for stochastic epidemics among a population of households. *Mathematical Biosciences*, 177:333–354, 2002b.

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

F. G. Ball and O. D. Lyne. Statistical inference for epidmeics among a population of households. *(In Preparation)*, 2016.

F. G. Ball and P. Neal. A general model for stochastic SIR epidemics with two levels of mixing. *Mathematical Biosciences*, 180(12):73–102, 2002.

F. G. Ball and L. M. Shaw. Estimating the within-household infection rate in emerging SIR epidemics among a community of households. *Journal of Mathematical Biology*, 71(6-7):1705–1735, 2015.

F. G. Ball and L. M. Shaw. Inference for emerging epidemics among a community of households. In *Lecture Notes in Statistics Proceedings III Workshop on Branching Processes and Their Applications*, volume 219, pages 269–284. Springer, 2016.

F. G. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *The Annals of Applied Probability*, 7(1):46–89, 1997.

F. G. Ball, T. Britton, and O. D. Lyne. Stochastic multitype epidemics in a community of households: estimation and form of optimal vaccination schemes. *Mathematical Biosciences*, 191(1):19–40, 2004a.

F. G. Ball, T. Britton, and O. D. Lyne. Stochastic multitype epidemics in a community of households: estimation of threshold parameter R and secure vaccination coverage. *Biometrika*, 91(2):345–362, 2004b.

F. G. Ball, T. Britton, and D. Sirl. Household epidemic models with varying infection response. *Journal of Mathematical Biology*, 63(2):309–337, 2010a.

F. G. Ball, D. Sirl, and P. Trapman. Analysis of a stochastic SIR epidemic on a random network incorporating household structure. *Mathematical Biosciences*, 224(2):53–73, 2010b.

F. G. Ball, T. Britton, T. House, V. Isham, D. Mollison, L. Pellis, and G. Scalia-Tomba. Seven challenges for metapopulation models of epidemics, including households models. *Epidemics*, 10:63–67, 2015.

F. G. Ball, L. Pellis, and P. Trapman. Reproduction numbers for epidemic models with households and other social structures II: Comparisons and implications for vaccination. *Mathematical Biosciences*, 274:108–139, 2016.

A. D. Barbour and S. Utev. Approximating the ReedFrost epidemic process. *Stochastic Processes and their Applications*, 113(2):173–197, 2004.

M. S. Bartlett. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, page 109, 1956.

R. Bartoszyński. Branching processes and the theory of epidemics. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 259–269, 1967.

R. Bartoszyński. On a certain model of an epidemic. *Applicationes Mathematicae*, 2(13):139–151, 1972.

N. G. Becker. Vaccination programmes for rare infectious diseases. *Biometrika*, 59(2):443–453, 1972.

N. G. Becker. The use of mathematical models in determining vaccination policies. *Bulletin of the International Statistics Institute*, 46:478–490, 1975.

N. G. Becker and K. Dietz. The effect of household distribution on transmission and control of highly infectious diseases. *Mathematical Biosciences*, 127(2):207–219, 1995.

N. G. Becker and D. N. Starczak. Optimal vaccination strategies for a community of households. *Mathematical Biosciences*, 139(2):117–132, 1997.

N. G. Becker and D. N. Starczak. The effect of random vaccine response on the vaccination coverage required to prevent epidemics. *Mathematical Biosciences*, 154(2):117–135, 1998.

N. G. Becker, T. Britton, and P. D. O'Neill. Estimating vaccine effects from studies of outbreaks in household pairs. *Statistics in Medicine*, 25(6):1079–1093, 2006.

T. Britton. Stochastic epidemic models: a survey. *Mathematical Biosciences*, 225 (1):24–35, 2010.

T. Britton and F. Giardina. Introduction to statistical inference for infectious diseases. *arXiv:1411.3138 [q-bio, stat]*, 2014.

S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boëlle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23(22):3469–3487, 2004.

D. Clancy and P. D. O'Neill. Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scandinavian Journal of Statistics*, 34(2):259–274, 2007.

N. Demiris and P. D. O'Neill. Bayesian inference for epidemics with two levels of mixing. *Scandinavian Journal of Statistics*, 32(2):265–280, 2005.

O. Diekmann and J. A. P. Heesterbeek. *Mathematical Epidemiology of Infectious diseases: Model Building, Analysis and Interpretation*. John Wiley & Sons, 2000.

K. Dietz. Transmission and control of arbovirus diseases. In D Ludwig and K. L. Cooke, editors, *Epidemiology*, pages 104–121. SIAM, Philadelphia, 1975.

K. Dietz and J. A. P. Heesterbeek. Daniel Bernoulli's epidemiological model revisited. *Mathematical Biosciences*, 180(1):1–21, 2002.

L. I. Dublin and A. J. Lotka. On the true rate of natural increase: As exemplified by the population of the United States, 1920. *Journal of the American Statistical Association*, 20(151):305–339, 1925.

P. D. En'ko. The epidemic course of some infectious diseases. *Vrac*, 10:1008–1010, 1889.

P. D. En'ko. On the course of epidemics of some infectious diseases. *International Journal of Epidemiology*, 18(4):749–755, 1989.

N. M. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214, 2005.

J. P. Fox and C. E. Hall. *Viruses in Families: Surveillance of Families as a Key to Epidemiology of Virus Infections.* PSG Publishing Company Inc., 1980.

C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS One*, 2(8):e758, 2007.

M. K. Gamado, G. Streftaris, and S. Zachary. Modelling under-reporting in epidemics. *Journal of Mathematical Biology*, 69(3):737–765, 2014.

Ghana Statistical Service. 2010 Population and housing census post enumeration survey report, 2012. URL `www.statsghana.gov.gh/`.

E. Goldstein, K. Paur, C. Fraser, E. Kenah, J. Wallinga, and M. Lipsitch. Reproductive numbers, epidemic spread and control in a community of households. *Mathematical Biosciences*, 221(1):11–25, 2009.

V. Gontcharoff. *Détermination des Fonctions Entières par Interpolation*. Hermann, Paris, 1937.

G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2nd edition, 2001.

M. E. Halloran, M. Haber, and I. M. Longini. Interpretation and estimation of vaccine efficacy under heterogeneity. *American Journal of Epidemiology*, 136(3): 328–343, 1992.

J. A. P. Heesterbeek. A brief history of R 0 and a recipe for its calculation. *Acta Biotheoretica*, 50(3):189–204, 2002.

H. W. Hethcote. Mathematical models for the spread of infectious diseases. In D Ludwig and K. L. Cooke, editors, *Epidemiology*, pages 122–131. SIAM, Philadelphia, 1975.

H. W. Hethcote. An immunization model for a heterogeneous population. *Theoretical Population Biology*, 14(3):338–349, 1978.

H. W. Hethcote and P. Waltman. Optimal vaccination schedules in a deterministic epidemic model. *Mathematical Biosciences*, 18(3):365–381, 1973.

R. E. Hope Simpson. Infectiousness of communicable diseases in the household:(measles, chickenpox, and mumps). *The Lancet*, 260(6734):549–554, 1952.

T. House, N. Inglis, J. V. Ross, F. Wilson, S. Suleman, O. Edeghere, G. Smith, B. Olowokure, and M. J. Keeling. Estimation of outbreak severity and transmissibility: Influenza A(H1N1)pdm09 in households: literature review (additional file 1). *BMC Medicine*, 10:117, 2012.

T. House, J. V. Ross, and D. Sirl. How big is an outbreak likely to be? Methods for epidemic final-size calculation. 469(2150):20120436, 2013.

C. Y. Hsu, A. M. F. Yen, L. S. Chen, and H. H. Chen. Analysis of household data on influenza epidemic with Bayesian hierarchical model. *Mathematical Biosciences*, 261:13–26, 2015.

P. Jagers. *Branching Processes with Biological Applications*. Wiley, London, 1975.

P. Jagers. Stabilities and instabilities in population dynamics. *Journal of Applied Probability*, pages 770–780, 1992.

N. L. Johnson and S. Kotz. *Continuous Univariate Distributions: Distributions in Statistics*. Hougton Mifflin, 1970.

M. J. Keeling and K. T. D. Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2011.

M. J. Keeling and J. V. Ross. Optimal prophylactic vaccination in segregated populations: When can we improve on the equalising strategy? *Epidemics*, 11:7–13, 2015.

M. J. Keeling and A. Shattock. Optimal but unequitable prophylactic distribution of vaccine. *Epidemics*, 4(2):78–85, 2012.

D. G. Kendall. Deterministic and stochastic epidemics in closed populations. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 149–165, 1956.

W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 115, pages 700–721. The Royal Society, 1927.

H. Kesten and B. P. Stigum. A Limit Theorem for Multidimensional Galton-Watson Processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966.

A. A. King, M. Domenech de Celles, F. M. G. Magpantay, and P. Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B*, 282(1806):20150347, 2015.

E. S. Knock and T. Kypraios. Bayesian non-parametric inference for infectious disease data. *arXiv:1411.2624 [stat]*, 2014.

L. L. H. Lau, H. Nishiura, H. Kelly, D. K. M. Ip, G. M. Leung, and B. J. Cowling. Household transmission of 2009 pandemic influenza A(H1N1): a systematic review and meta-analysis. *Epidemiology*, 23(4):531–542, 2012.

I. M. Longini and M. E. Halloran. A frailty mixture model for estimating vaccine efficacy. *Applied Statistics*, 45(2):165–173, 1996.

I. M. Longini and J. S. Koopman. Household and community transmission parameters from final distributions of infections in households. *Biometrics*, 38 (1):115–126, 1982.

I. M. Longini, K. Sagatelian, W. N. Rida, and M. E. Halloran. Optimal vaccine trial design when estimating vaccine efficacy for susceptibility and infectiousness from multiple populations. *Statistics in Medicine*, 17(10):1121–1136, 1998.

M. López-Garcia. Stochastic descriptors in an SIR epidemic model for heterogeneous individuals in small networks. *Mathematical Biosciences*, 271:42–61, 2016.

D. Ludwig. Final size distribution for epidemics. *Mathematical Biosciences*, 23 (1-2):33–46, 1975.

J. Ma, P. Van der Driessche, and F. H. Willeboordse. Effective degree household network disease model. *Journal of Mathematical Biology*, 66(1-2):75–94, 2012.

J. Ma, J. Dushoff, B. M. Bolker, and D. J. D. Earn. Estimating initial epidemic growth rates. *Bulletin of Mathematical Biology*, 76(1):245–260, 2014.

G. Macdonald. The analysis of equilibrium in malaria. *Tropical Diseases Bulletin*, 49(9):813–829, 1952.

G. Macdonald. The measurement of malaria transmission. *Proceedings of the Royal Society of Medicine*, 48(4):295, 1955.

A. Martin-Löf. Symmetric sampling procedures, general epidemic processes and their threshold limit theorems. *Journal of Applied Probability*, 23:265–282, 1986.

R. M. May and R. M. Anderson. Spatial heterogeneity and the design of immunization programs. *Mathematical Biosciences*, 72(1):83–111, 1984.

A. G. McKendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1925.

A. S. Monto, J. S. Koopman, and I. M. Longini. Tecumseh study of illness. XIII. Influenza infection and disease, 19761981. *American Journal of Epidemiology*, 121(6):811–822, 1985.

P. Neal. Efficient likelihood-free Bayesian computation for household epidemics. *Statistics and Computing*, 22(6):1239–1256, 2012.

P. Neal. A household SIR epidemic model incorporating time of day effects. *Journal of Applied Probability*, (to appear), 2016.

P. Neal and T. Kypraios. Exact Bayesian inference via data augmentation. *Statistics and Computing*, 25(2):333–347, 2015.

O. Nerman. On the convergence of supercritical general (CMJ) branching processes. *Probability Theory and Related Fields*, 57(3):365–395, 1981.

M. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002.

J. Neyman and E. L. Scott. A stochastic model of epidemics. *Stochastic Models in Medicine and Biology*, 45:83, 1964.

Office for National Statstics. 2011 Census: aggregate data (England and Wales), 2011.

P. D. O'Neill and C. H. Wen. Modelling and inference for epidemic models featuring non-linear infection pressure. *Mathematical Biosciences*, 238(1):38–48, 2012.

T. Ouboter, R. Meester, and P. Trapman. Stochastic SIR epidemics in a population with households and schools. *Journal of Mathematical Biology*, 72(5): 1177–1193, 2015.

L. Pellis, N. M. Ferguson, and C. Fraser. Epidemic growth rate and household reproduction number in communities of households, schools and workplaces. *Journal of Mathematical Biology*, 63(4):691–734, 2011.

L. Pellis, F. G. Ball, and P. Trapman. Reproduction numbers for epidemic models with households and other social structures. I. Definition and calculation of R0. *Mathematical Biosciences*, 235(1):85–97, 2012.

L. Pellis, S. E. F. Spencer, and T. House. Real-time growth rate for general stochastic SIR epidemics on unclustered networks. *Mathematical Biosciences*, 265:65–81, 2015.

P. Picard and C. Lefèvre. A unified analysis of the final size and severity distribution in collective Reed-Frost epidemic processes. *Advances in Applied Probability*, 22(2):269–294, 1990.

Registrar General for England and Wales. 1961 Census: aggregate data (Great Britain), 1961.

S. Riley, C. Fraser, C. A. Donnelly, A. C. Ghani, L. J. Abu-Raddad, A. J. Hedley, G. M. Leung, L-M. Ho, T-H. Lam, T. Q. Thach, P. Chau, K-P. Chan, S-V. Lo, P-Y. Leung, T. Tsang, W. Ho, K-H Lee, E. M. C. Lau, N. M. Ferguson, and R. M. Anderson. Transmission Dynamics of the Etiological Agent of SARS in Hong Kong: Impact of Public Health Interventions. *Science*, 300(5627):1961–1966, 2003.

R. Ross. *The Prevention of Malaria*. Dutton, 2nd edition, 1911.

R. Ross. An application of the theory of probabilities to the study of a priori pathometry. Part I. In *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, pages 204–230, 1916.

R. Ross and H. P. Hudson. An application of the theory of probabilities to the study of a priori pathometry. Part II. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 93, pages 212–225. The Royal Society, 1917a.

R. Ross and H. P. Hudson. An Application of the Theory of Probabilities to the Study of a priori Pathometry. Part III. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 93, pages 225–240, 1917b.

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

E. Santermans, N. Goeyvaerts, A. Melegaro, W. J. Edmunds, C. Faes, M. Aerts, P. Beutels, and N. Hens. The social contact hypothesis under the assumption of endemic equilibrium: Elucidating the transmission potential of VZV in Europe. *Epidemics*, 11:14–23, 2015.

G. Scalia-Tomba. Asymptotic final-size distribution for some chain-binomial processes. *Advances in Applied Probability*, 17(3):477–495, 1985.

G. Scalia-Tomba. On the asymptotic final size distribution of epidemics in heterogeneous populations. In *Stochastic Processes in Epidemic Theory*, pages 189–196. Springer, 1990.

G. Scalia-Tomba, A. Svensson, T. Asikainen, and J. Giesecke. Some model based considerations on observing generation times for communicable diseases. *Mathematical Biosciences*, 223(1):24–31, 2010.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980.

S. D. Silvey. *Statistical Inference*, volume 7 of *Monographs on Statistics and Applied Probability*. CRC Press, 1975.

H. M. Taylor. Some models in epidemic control. *Mathematical Biosciences*, 3: 383–398, 1968.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, 2009.

P. Trapman. On analytical approaches to epidemics on networks. *Theoretical Population Biology*, 71(2):160–173, 2007.

T. K. Tsang, L. L. H. Lau, S. Cauchemez, and B. J. Cowling. Household Transmission of Influenza Virus. *Trends in Microbiology*, 24(2):123–133, 2016.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, 1996.

J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1609):599–604, 2007.

R. K. Watson. On an epidemic in a stratified population. *Journal of Applied Probability*, 9(3):659–666, 1972.

P. Whittle. The outcome of a stochastic epidemica note on Bailey's paper. *Biometrika*, 42(1-2):116–122, 1955.

K. Wickwire. Mathematical models for the control of pests and infectious diseases: a survey. *Theoretical Population Biology*, 11(2):182–238, 1977.