Noname manuscript No.
(will be inserted by the editor)

Estimating the within-household infection rate in emerging SIR epidemics among a community of households

Frank Ball · **Laurence Shaw**

Received: date / Accepted: date

Abstract This paper is concerned with estimation of the within-household infection rate λ_L for a susceptible $\rightarrow$ infective $\rightarrow$ recovered epidemic among a population of households, from observation of the early, exponentially growing phase of an epidemic. Specifically, it is assumed that an estimate of the exponential growth rate is available from general data on an emerging epidemic and more-detailed, household-level data are available in a sample of households. Estimates of λ_L obtained using the final size distribution of single-household epidemics are usually biased owing to the emerging nature of the epidemic. A new method, which accounts correctly for the emerging nature of the epidemic, is developed by exploiting the asymptotic theory of supercritical branching processes and proved to yield a strongly consistent estimator of λ_L as the population and sampled households both tend to infinity in an appropriate fashion. The theory is illustrated by simulations which demonstrate that the new method is feasible for finite populations and numerical studies are used to explore how changes to the parameters governing the spread of an epidemic affect the bias of estimates based on single-household final size distributions.

Keywords Household epidemic model · SIR epidemic · Emerging epidemic · Parameter estimation · Branching process

Mathematics Subject Classification (2000) 92D30 · 62M05 · 60J85

F. Ball
School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK
Tel.: +44-(0)115-9514969
Fax: +44-(0)115-9514951
E-mail: frank.ball@nottingham.ac.uk

L. Shaw
School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK
E-mail: pmxlmsha@exmail.nottingham.ac.uk

1 Introduction

Mathematical models are being used increasingly to inform public health policy concerning control of emerging infections, see, e.g. Ferguson *et al.* [17] and Fraser *et al.* [19] for applications to avian influenza A(H5N1) and swine influenza A(H1N1), respectively. A key role for such models is to evaluate the effectiveness of possible strategies for containment of an emerging infection. In order to accomplish this, estimates are required of parameters used to define the model in question. This paper considers such estimation from data collected in the early phase of an emerging epidemic, using the model of Ball *et al.* [11] for the spread of an SIR (susceptible $\rightarrow$ infective $\rightarrow$ recovered) epidemic among a population partitioned into households.

The model of Ball *et al.* [11] assumes that an infectious individual makes two types of contacts, *local* contacts, i.e. with individuals chosen uniformly at random from the individual's household, and *global* contacts, i.e. with individuals chosen uniformly at random from the entire population. Although an oversimplification, this structure, which includes a departure from homogeneous mixing that is clearly present in human populations, yields a model that (i) is amenable to considerable mathematical analysis and (ii) leads to important insights into disease dynamics and control, such as the impact of household structure on the performance of vaccination strategies (Becker and Dietz [13], Becker and Starczak [14] and Ball and Lyne [9]). A household component is present in many complex simulation models (see, e.g. Ferguson *et al.* [17]). Moreover, data at a household level are often collected during emerging infections; see Cauchemez *et al.* [16] and House *et al.* [23] for analyses of such data for influenza A(H1N1) transmission in the United States and England, respectively.

For many stochastic models of epidemics with few initial infectives, if the disease does not die out quickly then, during the early stages of an epidemic, the number of infectives grows exponentially until saturation effects take over. Early exponential growth is also seen in many real-life epidemics and there has been a growing interest in quick inference methods during this stage of an epidemic. Assuming a homogeneously mixing population, Wallinga and Lipsitch [32] provided a simple estimate of the basic reproduction number R_0 (see, e.g. Heesterbeek and Dietz [21]) from an observed exponential growth rate r and knowledge of the generation interval for the disease. Fraser [18] extended this methodology to a community of households, using a closed-form approximate method for determining the exponential growth rate of the households epidemic model. Fraser gives two illustrative applications of his methodology, to pandemic influenza and measles, using historical data to obtain estimates of within-household transmission parameters. As Fraser notes, these transmission parameters could be quite different for future pandemics, so methods are required for estimating such parameters from data on an emerging infection.

The following scenario is considered in this paper. It is assumed that the household size distribution for the population is known (this is usually available from census data), an estimate of the exponential growth rate r is available from general data on an emerging epidemic and more-detailed, household-level data are available in a sample of households. The primary goal is to estimate the local (within-household) infection rate λ_L from this information, whilst the epidemic is still in its emerging

phase. For most of the paper it is assumed, primarily for ease of notation, that there is no latent period and that the infectious period distribution is known, though both of these assumptions may be relaxed. For inference based on final outcome data (e.g. Knock and O'Neill [25] and Ball and Lyne [10]), estimates of infection rates are (i) invariant to very general assumptions concerning a latent period and (ii) confounded with the scale of the infectious period distribution. Neither is true for inference in an emerging epidemic. The partial nature of the assumed available data renders full maximum likelihood estimation difficult, if indeed feasible; the amount of unobserved data is such that computationally intensive methods for incomplete data, such as the EM and data augmentation MCMC, may well be problematic. Thus an alternative estimation procedure is developed and shown to give a strongly consistent estimator of λ_L as the population and sampled households both tend to infinity in an appropriate fashion.

It is well known that the early stages of an SIR epidemic among a community of households may be approximated by a branching process in which individuals correspond to single-household epidemics. Thus if, for example, the available data consist of the total number of cases in completed sub-epidemics within households, it is tempting to estimate λ_L by fitting the usual final size distribution for a single-household epidemic (see, e.g. Ball [4]) to such data. However, as illustrated in Section 3.2, this leads to λ_L being underestimated because in an emerging epidemic the completed single-household epidemics are likely to be the smaller ones. An improved estimate may be obtained by including single-household epidemics that are still ongoing at the time when estimation is performed, using right-censoring for their size, but, as also demonstrated in Section 3.2, the resulting estimate is still biased. In order to obtain unbiased estimates, one needs to account correctly for the emerging nature of the epidemic which produced these data. (Similar issues arise in estimating the generation time of an infectious disease early in an epidemic [31].) The main purpose of this paper is to show that this can be achieved by using the theory of Nerman [28] concerning the asymptotic behaviour of counts of characteristics associated with supercritical general (i.e. Crump-Mode-Jagers) branching processes applied to the above-mentioned branching process which approximates the early stages of an epidemic in a community of households.

The paper is structured as follows. The households epidemic model of Ball *et al.* [11] is described in Section 2 and the early stages of epidemics in a large population is considered in Section 3. The threshold behaviour of the model is outlined in Section 3.1. Estimation of λ_L by fitting the usual final size distribution to single-household epidemics, both without and with censoring, is considered and shown to be inadequate in Section 3.2. The new method, which incorporates correctly the emerging nature of the epidemic is described in Section 4. The theory for the method is developed in Section 4.1 for the situtations when, at the time the inference is performed, (i) complete knowledge of the numbers of infective and recovered individuals in each household is available, and (ii) (sometimes the more realistic scenario) only the numbers of recovered individuals in each household are available. Some extensions of the theory and implementation issues are considered in Section 4.2. The theory as developed does not make any assumptions concerning the infectious period distribution, other than it possesses a moment-generating function, but it does need

to be specified. However, the method is easy to implement only if single-household epidemic dynamics are Markovian, i.e. if the infectious period follows an exponential distribution, though phase-type distributions can also be accommodated. Extensions to incorporate a latent period and allow for the rate of the exponential distribution used to model the infectious period to be unknown are discussed briefly, as is allowing λ_L to depend on household size. Similar theory is developed for in Section 5 for a households Reed-Frost type model, in which the latent period is constant and the infectious period is reduced to a single point in time, using multitype Galton-Watson branching process. Simulations depicting how the estimation methodologies developed in Sections 4 and 5 perform in practice are shown in Section 6, while other plots in this section illustrate how changes to the parameters governing the spread of an epidemic affect the bias of the estimates based on single-household final size distributions. Proofs that the estimators derived in Section 4 are strongly consistent under suitable conditions are given in Section 7. Finally, some concluding comments are given in Section 8.

2 Model

The model used is based on that of Ball *et al*. [11] for describing the spread of an SIR epidemic in a population that has been partitioned into households. For a population in which n_{max} is the size of the largest household, let m_n be the number of households of size n, for $n = 1,2,...,n_{max}$, so that $m = \sum_{n=1}^{n_{max}} m_n$ and $N = \sum_{n=1}^{n_{max}} nm_n$ are, respectively, the total numbers of households and individuals in the population. Also, for $n = 1,2,...,n_{max}$, 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 who reside in households of size n.

The epidemic is initiated by a small number of individuals becoming infected at time $t = 0$. Once infected, an individual remains in this state for the duration of its infectious period, which for each individual is independently and identically distributed according to a random variable T_I, having an arbitrary but specified distribution. Once its infectious period is over, an individual is recovered and it plays no further part in the epidemic. During its infectious period, a given infective makes global contacts with any other given individual in the population at points of a homogeneous Poisson process having rate λ_G/N and it makes additional local contacts with any given individual in the same household at points of a homogeneous Poisson process having rate λ_L. All the Poisson processes describing infectious contacts (whether or not either or both of the individuals involved are the same) and the random variables describing the infectious periods are mutually independent. Whenever an infective makes contact with a susceptible individual, the susceptible becomes infected and is immediately able to transmit infection. Thus there is no latent period, though this can be relaxed; see Section 4.2. The process continues until there is no infective remaining in the population, at which point the epidemic is deemed to have ceased.

3 Early stages of an epidemic

3.1 Threshold Parameter

When the number of households m is large, the probability of a global infectious contact in the early stages of an epidemic being with a susceptible in a previously infected household is small. Thus, the initial behaviour of an epidemic in a community of households can be approximated by a branching process of infected households, in which each global contact is assumed to be with an individual in a fully susceptible household. Let R_* be the mean number of global contacts that emanate from a typical household in this branching process. Then R_* is a threshold parameter for the households epidemic model, in that in the limit as $m \to \infty$, the epidemic takes off with non-zero probability if and only if $R_* > 1$; see Ball *et al.* [11], where calculation of R_* is described.

The remainder of this paper focuses exclusively on epidemics where this condition holds and is concerned with epidemics that do take off. It is assumed that $\mathbb{E}[T_I] = 1$ as this can be done without loss of generality by rescaling the time axis.

3.2 Basic approach to estimating λ_L

Suppose one wishes to estimate λ_L for an epidemic that is observed whilst it is still in its initial stages and is therefore still mimicking the infected households branching process outlined above. For $n = 1, 2, \ldots$ and $x = 0, 1, \ldots, n-1$, let $p^{(n)}_{basic}(x|\lambda_L)$ 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. By using Equation (2.5) of Ball [4], $p^{(n)}_{basic}(x|\lambda_L)$ $(x = 0, 1, \ldots, n-1)$ can be determined using the following triangular system of linear equations:

$$\sum_{i=1}^{j} \binom{n-i}{j-i} p^{(n)}_{basic}(n-i|\lambda_L)\phi(n-j)^i = \binom{n-1}{j-1}, \quad j = 1, 2, \ldots, n,$$

where $\phi(\theta) = \mathbb{E}[\exp(-\theta T_I)]$ $(\theta \geq 0)$ is the moment-generating function of T_I.

Let $a^{(n)}_{x,y}$ 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 λ_L by maximising the pseudolikelihood function

$$L_{basic}(\lambda_L|\boldsymbol{a}) = \prod_{n=1}^{n_{max}} \prod_{x=0}^{n-1} p^{(n)}_{basic}(x|\lambda_L)^{a^{(n)}_{x,0}}. \tag{3.1}$$

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 not used. 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 $n = 1, 2, \ldots n_{max}$ and $x = 0, 1, \ldots, n-1$,

let $q_{basic}^{(n)}(x|\lambda_L) = \sum_{i=0}^{x} p_{basic}^{(n)}(i|\lambda_L)$ 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 λ_L, with left-censoring for the number of survivors (i.e. right-censoring for the total size), by maximising

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

Figure 1 shows how well the basic and censored MPLE methods perform in practice. For these histograms, epidemics were simulated in a population containing 1 000 000 households, with estimates of λ_L 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 λ_L were made for the first 1000 epidemics to reach the 1000 recovery milestone. A large population was used to ensure that the simulated epidemics were still approximately mimicking a branching process at the time of estimation. The household distribution $\boldsymbol{\alpha}$ that was used was $[0.29, 0.34, 0.16, 0.14, 0.05, 0.02]$, i.e. $n_{max} = 6$ and $\alpha_1 = 0.29, \alpha_2 = 0.34, ..., \alpha_6 = 0.02$, as suggested by Fraser [18], and is based on UK census data from 2001 [34]. The infectious period was chosen to be exponentially distributed, the infectious parameters were $\lambda_G = 1$ and $\lambda_L = 1$, and all epidemics were initiated by a single individual, chosen uniformly at random from the population, becoming infected.

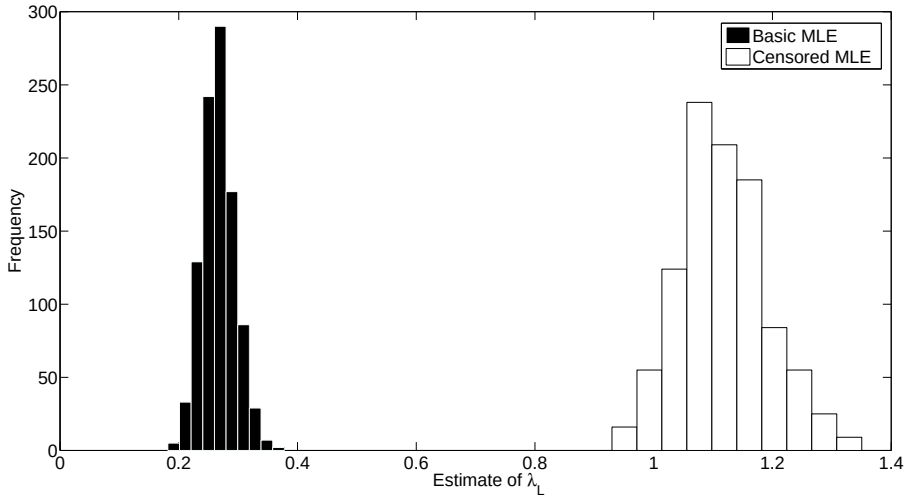

Fig. 1: Estimates of λ_L, with a true value of 1, from 1000 epidemic simulations using the basic and censored MPLE methods

It is clear from Figure 1 that the basic MPLE method severely underestimates λ_L. Households in which the epidemic spreads locally are more likely to still be infective at the time of observation than households infected at the same time but in which the initial infective does not infect any other individual locally. Consequently, the households that contain less severe local epidemics are more likely to be included in the basic MPLE estimate, causing the observed underestimate of λ_L. The censored MPLE approach appears to offer an improvement but repeated simulations with different parameters showed that this method generally overestimates λ_L, as is observed in Figure 1.

In order to obtain a more accurate estimate of λ_L 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 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 $\tilde{p}^{(n)}_{x,0}(r|\lambda_L)$ $(x = 0,1,...,n-1)$. Thus we assume that each observed household in the data has final size that comes from that distribution and estimate λ_L by maximising the pseudolikelihood obtained by replacing $p^{(n)}_{basic}(x|\lambda_L)$ by $\tilde{p}^{(n)}_{x,0}(\hat{r}|\lambda_L)$ in (3.1), where $\hat{r}$ is an estimate of the growth rate r; see (4.5) in the next section, where calculation of $\tilde{p}^{(n)}_{x,0}(r|\lambda_L)$ is explained.

4 A new method

4.1 A more accurate estimator

Consider the approximating branching process introduced in Section 3.1, 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 $E^{(n)}_H$ 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 $X^{(n)}_H(t)$ and $Y^{(n)}_H(t)$ be respectively the numbers of susceptibles and infectives in $E^{(n)}_H$ at time t. Let $\mathscr{T}^{(n)} = \{(x,y) : \ x = 0,1,...,n-1; \ y = 0,1,...,n-x\}$ and, for $(x,y) \in \mathscr{T}^{(n)}$, let $p^{(n)}_{x,y}(t|\lambda_L) = \mathbb{P}(X^{(n)}_H(t) = x, \ Y^{(n)}_H(t) = y)$ $(t \geq 0)$ and $\tilde{p}^{(n)}_{x,y}(r|\lambda_L) = \int_0^\infty e^{-rt} p^{(n)}_{x,y}(t|\lambda_L) \ \mathrm{d}t$ $(r \geq 0)$. Further, let $\xi^{(n)}_H$ be the point process describing times that global contacts emanate from $E^{(n)}_H$, so, for $t \geq 0$, $\xi^{(n)}_H([0,t])$ is the number of global contacts that emanate from $E^{(n)}_H$ during $[0,t]$. For $t \geq 0$ let $\mu^{(n)}(t) = \mathbb{E}[\xi^{(n)}_H([0,t])]$ and note that

$$\mu^{(n)}(\mathrm{d}t) = \lambda_G \sum_{(x,y)\in\mathscr{T}^{(n)}} y p^{(n)}_{x,y}(t|\lambda_L) \ \mathrm{d}t. \tag{4.1}$$

Let ξ_H be a mixture of $\xi^{(1)}_H, \xi^{(2)}_H, ..., \xi^{(n_{max})}_H$ with mixing probabilities $\tilde{\alpha}_1, \tilde{\alpha}_2, ..., \tilde{\alpha}_{n_{max}}$. Then ξ_H is a point process which describes the ages at which a typical individual reproduces in the approximating branching process. Note that this branching

process is a general (i.e. Crump-Mode-Jagers) branching process; e.g. Haccou *et al* [20], Section 3.3. For $t \geq 0$, let

$$\mu(t) = \mathbb{E}[\xi_H([0,t])] = \sum_{n=1}^{n_{max}} \tilde{\alpha}_n \mu^{(n)}(t). \tag{4.2}$$

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

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

Note that, from (4.1) and (4.2), r satisfies

$$\lambda_G \sum_{n=1}^{n_{max}} \tilde{\alpha}_n \sum_{(x,y)\in\mathscr{T}^{(n)}} y\tilde{p}_{x,y}^{(n)}(r|\lambda_L) = 1. \tag{4.3}$$

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. For $n = 1,2,...,n_{max}$ and $(x,y) \in \mathscr{T}^{(n)}$, 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 $\mathscr{T} = \{(n,x,y) : n = 1,2,...,n_{max} \text{ and } (x,y) \in \mathscr{T}^{(n)}\}$. For $t \geq 0$ and $(n,x,y) \in \mathscr{T}$, 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. Then it is easily verified that the conditions of Theorem 5.4 of Nerman [28] are satisfied and it follows from that theorem 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 $(n,x,y) \in \mathscr{T}$,

$$e^{-rt}Y_{n,x,y}(t) \xrightarrow{a.s.} \tilde{\alpha}_n \tilde{p}_{x,y}^{(n)}(r|\lambda_L)W \quad \text{as } t \to \infty. \tag{4.4}$$

Note that $\sum_{(x,y)\in\mathscr{T}^{(n)}} p_{x,y}^{(n)}(t|\lambda_L) = 1$, so $\sum_{(x,y)\in\mathscr{T}^{(n)}} \tilde{p}_{x,y}^{(n)}(r|\lambda_L) = 1/r \quad (n = 1,2,..., n_{max})$. 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 $\tilde{\alpha}_n r \tilde{p}_{x,y}^{(n)}(r|\lambda_L)$.

Return to the households epidemic model. Recall that for $(n,x,y) \in \mathscr{T}$, the number of households of size n that contain x susceptibles and y infectives when the epidemic is observed is denoted by $a_{x,y}^{(n)}$. 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, λ_L can be estimated by maximising the normalised pseudolikelihood function

$$L_{full}(\lambda_L|\boldsymbol{a},\hat{r}) = \prod_{n=2}^{n_{max}} \prod_{(x,y)\in\mathscr{T}^{(n)}} \tilde{p}_{x,y}^{(n)}(\hat{r}|\lambda_L)^{a_{x,y}^{(n)}}. \tag{4.5}$$

Note that households of size 1 provide no information about λ_L, so they do not contribute to L_{full}, and that L_{full}, L_{basic} and L_{cansor} are not true likelihood functions as

they assume independence between households. In Section 7 we prove that, under suitable conditions, the estimator $\hat{\lambda}_L = \text{argmax}\, L_{full}(\lambda_L|\boldsymbol{a},\hat{r})$ is strongly consistent as the number of households $m \to \infty$, i.e. that $\hat{\lambda}_L$ converges almost surely to the true value λ_L as $m \to \infty$.

Suppose that estimation is based only on completed single-household epidemics, as in the basic MPLE method. Then λ_L may be estimated by maximising

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

Observe that subject to mild conditions,

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

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 $n = 1,2,...,n_{max}$ and $j = 1,2,...,n$ let $c_j^{(n)}$ be the observed number of households of size n with j recoveries, let $\mathscr{A}_j^{(n)} = \{(x,y) \in \mathscr{T}^{(n)} : x+y = n-j\}$ and let

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

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

$$L_{rec}(\lambda_L|\boldsymbol{c},\hat{r}) = \prod_{n=2}^{n_{max}} \prod_{j=1}^{n} \tilde{q}_j^{(n)}(\hat{r}|\lambda_L)^{c_j^{(n)}}. \tag{4.6}$$

4.2 Practicalities and extensions

Estimates of λ_L based upon the L_{full} and L_{rec} pseudolikelihoods are both dependent on knowing $\tilde{p}_{x,y}^{(n)}(r|\lambda_L)$ for $(n,x,y) \in \mathscr{T}$, which is not practical in many circumstances. It is, however, 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.* [29] 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 2 shows the transition rates of $E_H^{(3)}$ as a CTMC and also assigns labels to each state $(x,y) \in \mathscr{T}^{(3)}$. 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 $\mathscr{T}^{(n)}$ of $E_H^{(n)}$ has size $s^{(n)} = |\mathscr{T}^{(n)}| = n(n+3)/2$. Let $Q^{(n)}(\lambda_L) = [q_{ij}^{(n)}(\lambda_L)]$

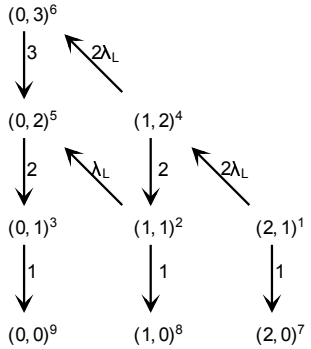

Fig. 2: Graphical representation of a single-household epidemic for households of size 3 as a CTMC, where (x,y) denotes the household state and state labels (shown as superfixes) for the CTMC are assigned as described. The values on the arrows represent transition rates between states in the single-household epidemic.

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

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

where $I_{s^{(n)}}$ is the $s^{(n)} \times s^{(n)}$ identity matrix.

The estimating procedure described in Section 4.1 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 $p_{x,y}^{(n)}(t|\lambda_L)$ replaced in an obvious fashion by $p_{x,y}^{(n)}(t|\lambda_L,\gamma)$ and (λ_L,γ) being estimated by maximising the appropriate normalised pseudolikelihood function. Note that for final outcome data it is impossible to estimate both λ_L 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 (p.71-78) [2]). 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 $\mathscr{T}^{(n)}$. One can also extend the model to an SEIR (susceptible $\rightarrow$ exposed $\rightarrow$ infectious $\rightarrow$ 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.

The methodology can be extended to allow the local contact rate to depend on household size. For $n = 1,2,...,n_{max}$, let $\lambda_L^{(n)}$ denote the local contact rate in a household of size n. The, provided there are enough households of each size in the sample, $(\lambda_L^{(2)},\lambda_L^{(3)},\ldots\lambda_L^{(n_{max})})$ can be estimated jointly, e.g. by replacing λ_L by $\lambda_L^{(n)}$ in (4.5). Alternatively, one can assume a specific form for $\lambda_L^{(n)}$, Cauchemez *et al.* [15] use $\lambda_L^{(n)} = \lambda_L/n$ for influenza, and estimate its unknown parameter (here λ_L) in the obvious fashion.

5 Application to the Reed-Frost model

5.1 The Reed-Frost model

Under the Reed-Frost model, the latent period is assumed to have a 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 $t = 0$ among a population having the same structure as that outlined in Section 2. 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 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. 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 $t+1$. The process continues until there is no infective left in the population.

Again, we consider the case of an emerging epidemic, 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 $N \rightarrow \infty$, 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 $\rho(p_L,\mu_G)$ of the approximating branching process. The parameter μ_G increases with $\rho(p_L,\mu_G)$ for fixed p_L, so for any estimate of p_L, an estimate for μ_G is pre-determined since it is assumed that $\rho(p_L,\mu_G)$ can be observed directly.

5.2 Estimating p_L

The local contact probability p_L can be estimated by approximating the early stages of a Reed-Frost epidemic with a discrete-time multitype branching process S. Define the type space of S as $\mathscr{T}_{RF} = \{(n,n-1,1) : 1 \le n \le n_{max}\} \cup \bigcup_{n=1}^{n_{max}} \{(n,x,y) : x \ge 0, y \ge 1, x+y < n\}$ and label the elements of $\mathscr{T}_{RF}$ as $1,2,...,k$ where $k = |\mathscr{T}_{RF}| = n_{max} + \sum_{n=2}^{n_{max}} \frac{n(n-1)}{2} = n_{max}(n_{max}^2+5)/6$. The type space includes all possible household states where infection is still present.

Let $\boldsymbol{M}$ be the mean matrix of S on $\mathscr{T}_{RF}$, 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 $\binom{x}{z}(1-(1-p_L)^y)^z(1-p_L)^{y(x-z)}$ 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 $\mathscr{T}_{RF}$ alive after t generations of S and let $\rho(p_L,\mu_G)$ be the maximal eigenvalue of $\boldsymbol{M}$. Assume that $\rho(p_L,\mu_G) > 1$, so the branching process is supercritical. Kesten and Stigum [24] show that if $\boldsymbol{u}(p_L,\mu_G)$ is the left-eigenvector associated with $\rho(p_L,\mu_G)$, normalised so that its components are non-negative and sum to one, then

$$\rho(p_L,\mu_G)^{-t}\boldsymbol{Y}_t \xrightarrow{a.s.} W\boldsymbol{u}(p_L,\mu_G) \quad \text{as } t \to \infty, \tag{5.1}$$

where W is a non-negative random variable such that $W = 0$ if and only if S becomes extinct. The eigenvector $\boldsymbol{u}(p_L,\mu_G)$ therefore gives the proportions of individuals of each type in S as $t \to \infty$, conditional upon S not going extinct. It follows from (5.1) that

$$\rho(p_L,\mu_G)^{-t}\sum_{t'=1}^{t}\boldsymbol{Y}_t' \xrightarrow{a.s.} \frac{\rho(p_L,\mu_G)}{\rho(p_L,\mu_G)-1}W\boldsymbol{u}(p_L,\mu_G) \quad \text{as } t \to \infty. \tag{5.2}$$

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 $i \in \mathscr{T}_{RF}$. A household in state (n,x,y) at time t' has probability $(1-p_L)^{xy}$ 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 (5.2) and the strong law of large numbers that, for $i = 1,2,...,k$,

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

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

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

where K is chosen such that

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

One can then estimate p_L by performing maximum pseudolikelihood estimation in exactly the same manner as described using L_{full} in Section 4.1. 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 4.1.

6 Numerical Illustrations

6.1 Methods of estimation

We illustrate applications of the preceding theory using simulation studies with parameter choices loosely based on Fraser's [18] analysis of varicella data. Simulations are performed on a population of $m = 10\ 000$ households with distribution $\boldsymbol{\alpha} = [0.13, 0.30, 0.23, 0.18, 0.09, 0.07]$. This distribution is based on the 1961 UK census data [34] and contains a higher proportion of larger households than the 2001 distribution used previously, meaning that local infectious contacts should have a greater effect 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 is 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 [22], 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)) 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, e.g. Ma *et al.* [27].

For illustrative purposes, estimates of λ_L in this subsection are given in terms of the secondary attack rate (SAR), as defined by Longini and Koopman [26]. The SAR is the probability that an infective infects 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 Sections 3 and 4 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 3 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 above formula) were made at regular intervals throughout the epidemic using the basic MPLE, censored MPLE and full-and-recovery-pseudolikelihood estimation methods (where the latter two use (4.5) and (4.6) respectively), and an additional estimate was made using the pseudolikelihood method of Ball and Lyne (2014) [10] (c.f. Section 5.1 of Ball *et al.* [11]) by considering the distribution of susceptible individuals in households of all sizes at the *end* of an epidemic. This is referred to as the *final-size* method of estimation. Note that for the basic MPLE method, it takes some time before the SAR is estimated to be any value other than 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 3 shows that once an epidemic has had sufficient time to establish itself, there is a window when the both the full and recovery CMJBP methods appear to give a good estimate of the SAR. Moreover, the length of this window 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 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 4.1 and those marked 'γ unknown' assume that γ is also estimated from the data, as described in Section 4.2. Estimations 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

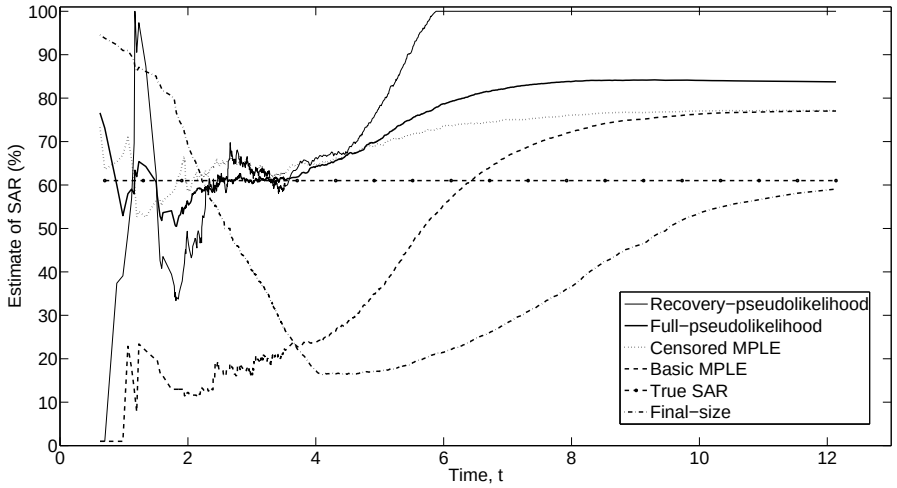

Fig. 3: Estimates of the SAR (true value 61%) through time for a single SIR households epidemic. The four estimation methods outlined earlier in the paper are shown along with estimates of the SAR using the final-size method.

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. The inset of Figure 4 shows a scatter plot of the estimates of (SAR, γ) using the full-pseudolikelihood CMJBP method, which indicates that the estimates of the SAR and γ are positively correlated.

Repeated simulations using different population sizes yielded very similar results to those seen in Figure 3, 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 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 [12] 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.

The above points are illustrated in Figure 5 which shows the mean squared error (MSE) of estimates of the SAR, using the full-pseudolikelihood method and assuming that $\gamma\ (=1)$ is known, throughout the emerging stages of 1000 simulated epidemics among populations with differing numbers of households but with the same population structure α, growth-rate r and SAR as given above. It is assumed that

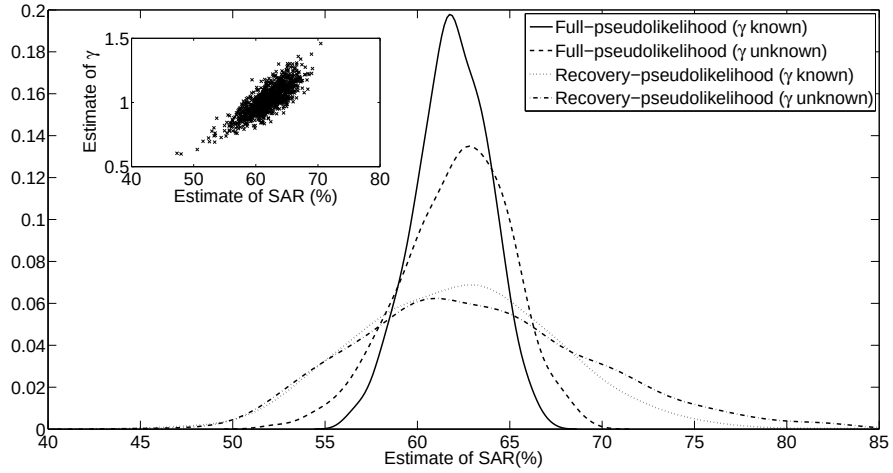

Fig. 4: Kernel density estimates of the distribution of the estimator the SAR (true value 61%) based on 1000 simulations of the outlined epidemic using the full and recovery CMJBP estimation methods, both with and without the recovery rate γ (true value 1.00) being also estimated. Inset: Scatter plot of estimates of (SAR, γ) for the full-pseudolikelihood (γ unknown) method.

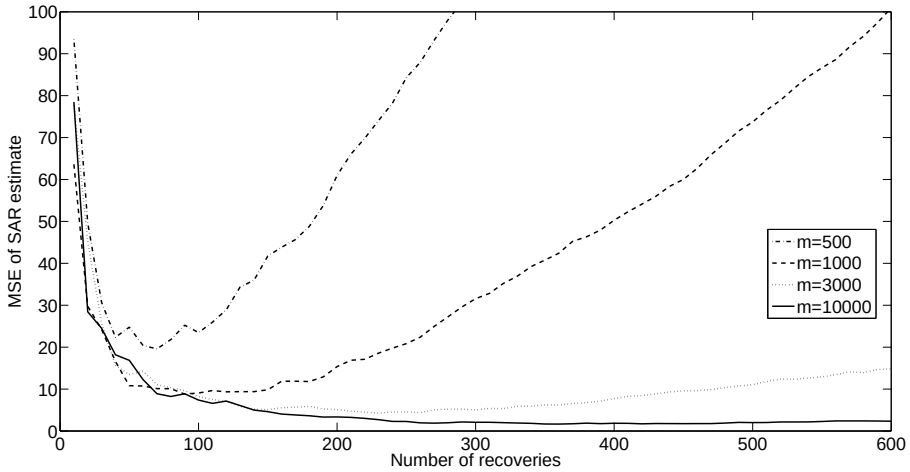

Fig. 5: MSE of estimates of the SAR using the full-pseudolikelihood method. See text for details.

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 MSE to settle to a reasonable 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.

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 6 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-pseudolikelihood 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 methods of estimation outlined in Section 4.1 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.

6.2 Relationship between parameters of the model and bias of the basic and censored MPLE methods

In Section 4 it is established that the new method of estimating λ_L in an emerging epidemic is unbiased, given an infinite population and assumptions regarding the time of estimation. It is also seen throughout this paper that the basic MPLE and censored MPLE provide inaccurate estimates of λ_L for various emerging epidemics. We now look to establish the extent of the bias of these two methods and how the bias is

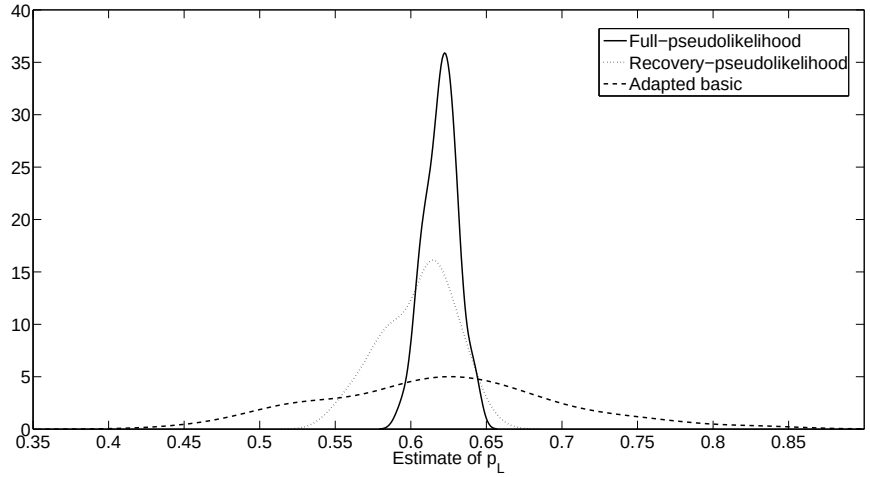

Fig. 6: Kernel density estimates of the distribution of the estimator of p_L (true value 0.61) based on 1000 simulations of Reed-Frost type epidemics; see text for details

affected by various parameters of an epidemic. This is achieved 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. We return to estimating λ_L rather then the *SAR* for this section as the *SAR* estimations provide no illustrative advantage in this asymptotic context when estimators have a variance of 0. Households data are considered to be perfect for an emerging epidemic in continuous-time with paramerters 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 $(n,x,y)\in\mathscr{T}$. (Note that with perfect data, $\hat{\lambda}_L = \text{argmax}\tilde{l}^{(\infty)}_{full}$, see equation (7.6) in Section 7.) Similarly, perfect data for an emerging Reed-Frost epidemic with parameters p_L and μ_G is achieved when the proportion of household in state (n,x,y) is exactly $p_{RF\,full}(n,x,y|p_L,\mu_G)$ for all $(n,x,y)\in\mathscr{T}_{RF}$. Note that in both cases, the distribution of household states representing perfect data is also dependent on the population structure $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,...,\alpha_{n_{max}})$.

6.2.1 Effect of local contact rate

Figure 7 illustrates the effect of λ_L 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 population distribution $\boldsymbol{\alpha} = [0.13, 0.30, 0.23, 0.18, 0.09, 0.07]$, as given in Section 6.1 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

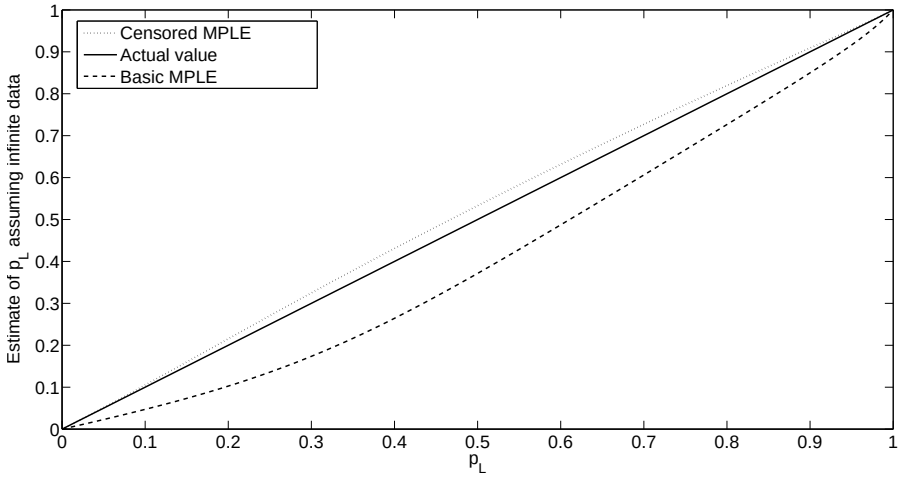

Fig. 7: Estimates of different values p_L assuming perfect data in emerging Reed-Frost type epidemics, $\rho = 2.248$, using the basic and censored MPLE methods

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.

6.2.2 Effect of household size

Figure 8 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 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 7).

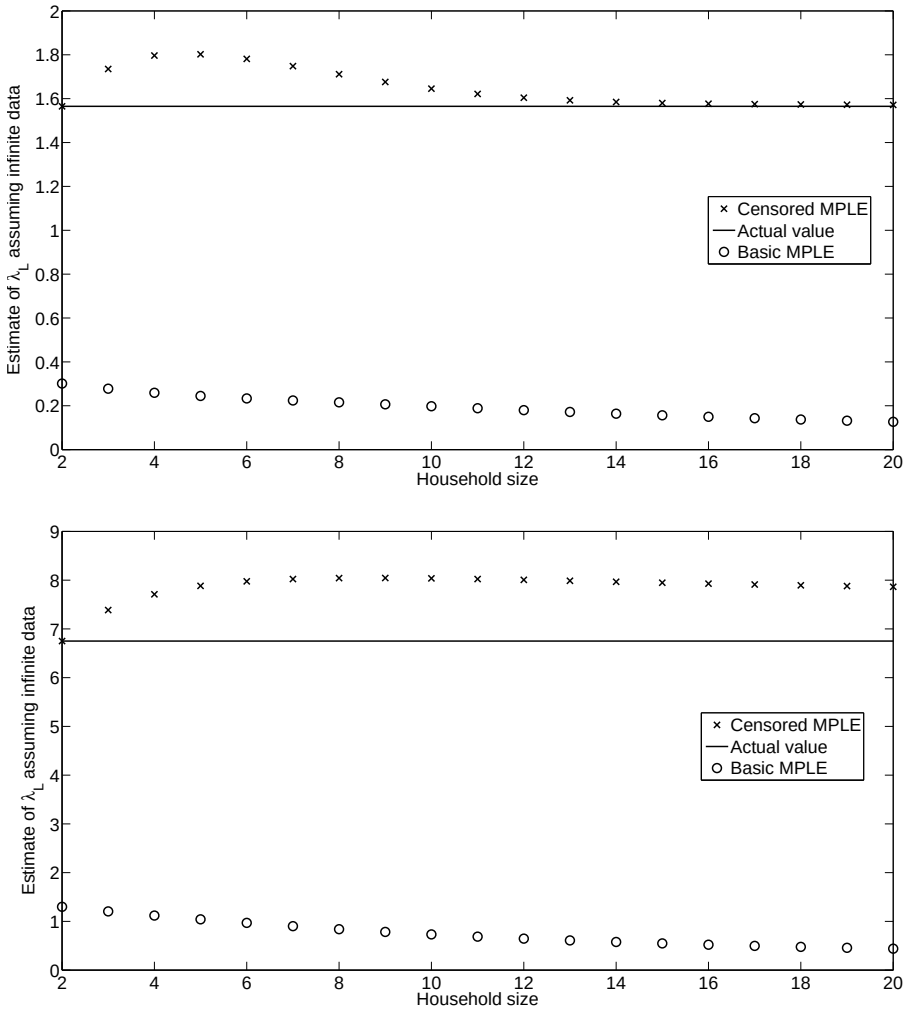

Fig. 8: Estimates of λ_L assuming perfect data for emerging epidemics, with $r = 1.762$, among populations with equal household sizes using the basic and censored MPLE methods. The upper plot takes $\lambda_L = 1.565$ for all household sizes. The lower plot adopts the model $\lambda_L^{(n)} = \lambda_L/n$, where n is household size and $\lambda_L = 6.75$

The lower plot of Figure 8 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 $\lambda_L = 6.75$ (see Section 4.2). 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 and censored MPLE approaches both become more biased as household size increases. 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.

6.2.3 Effect of growth rate

Figure 9 shows estimates of λ_L in emerging epidemics with λ_L and α as defined in Section 6.1. 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 4.1.

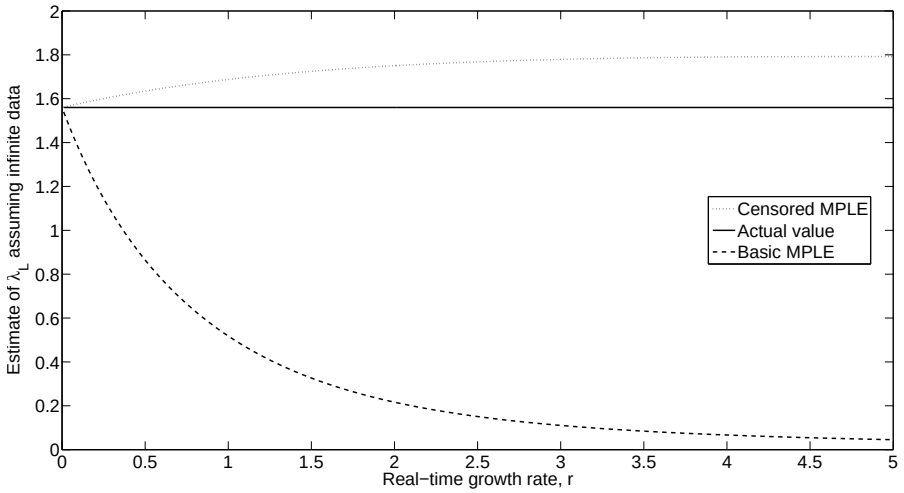

Fig. 9: Estimates of λ_L assuming perfect data in emerging epidemics with different real-time growth rates r using the basic and censored MPLE methods

7 Strong consistency of estimators

In this section we consider the asymptotic behaviour of the estimators of λ_L described in Section 4 as the number of households in the population tends to infinity. Specifically we show that, under suitable conditions, the estimators are strongly consistent, conditional upon the epidemic taking off.

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 Section 2 and has one initial infective, who is chosen uniformly at random from the population. The infective parameters (λ_L, λ_G) and the infectious period distribution

are all assumed to be independent of m, as is the maximum household size n_{max}. 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, introduced in Section 3 and analysed further in Section 4, which approximates the epidemic $E^{(m)}$ for suitably large m. Recall that for $(n,x,y) \in \mathscr{T}$, the number of individuals in $E^{(\infty)}$ having state (n,x,y) at time t is denoted by $Y_{n,x,y}(t)$. For $m = 1,2,...$, $(n,x,y) \in \mathscr{T}$ 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 $\mathscr{T}_L = \{(n,x,y) \in \mathscr{T} : y \geq 1\}$. For $t \geq 0$, let $Y(t) = \sum_{(n,x,y)\in\mathscr{T}_L} Y_{n,x,y}(t)$ denote the number of "live" individuals in $E^{(\infty)}$ at time t. Recall that r denotes the Malthusian parameter of $E^{(\infty)}$.

Theorem 7.1 *Suppose that $r > 0$. Then there is a probability space $(\Omega, \mathscr{F}, \mathrm{P})$ 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 $A = \{\omega \in \mathbb{R} : \lim_{t\to\infty} Y(t,\omega) = 0\}$ denote the set on which the branching process $E^{(\infty)}$ goes extinct. Then for* P*-almost all $\omega \in A^{\complement}$ and any $c \in (0, \frac{1}{2}r^{-1})$,*

$$\sup_{0\leq t\leq c\log m} \max_{(n,x,y)\in\mathscr{T}} |Y_{n,x,y}^{(m)}(t,\omega) - Y_{n,x,y}(t,\omega)| = 0 \tag{7.1}$$

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 $(\Omega, \mathscr{F}, \mathrm{P})$ 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, $\chi_1^{(m)}, \chi_1^{(m)}, ...$ 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 $(\Omega, \mathscr{F}, \mathrm{P})$ 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 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 contacts is 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 details are not required for the present proof.

For $m = 1,2,...$, let $M^{(m)}$ be the smallest $k \geq 2$ 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

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

Note that $M^{(m)}$ is stochastically greater than $\hat{M}^{(m)}$, since the maximum household size is n_{max}, and (c.f. Aldous [1], page 96) $m^{-1/2}\hat{M}^{(m)} \xrightarrow{D} \hat{M}$ as $m \to \infty$, where $\xrightarrow{D}$ denotes convergence in distribution and $\hat{M}$ has density $f(x) = n_{max} x \mu_H^{-1} \exp(-n_{max}\mu_H^{-1}x^2/2)$ $(x > 0)$, 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 $m^{-1/2}\hat{M}^{(m)} \xrightarrow{a.s.} \hat{M}$ 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 $(\Omega, \mathscr{F}, \mathrm{P})$ and that there exists $B \in \mathscr{F}$ with $\mathbb{P}(B) = 1$, such that, for all $\omega \in B$,

$$M^{(m)}(\omega) \geq \hat{M}^{(m)}(\omega) \quad \text{and} \quad m^{-1/2}\hat{M}^{(m)}(\omega) \to \hat{M}(\omega) \quad \text{as } m \to \infty. \tag{7.2}$$

For $t \geq 0$, let $T(t)$ be the number of births in $E^{(\infty)}$ during $[0,t]$, including the ancestor. Then $T(t) = \sum_{(n,x,y)\in\mathscr{T}} Y_{n,x,y}(t)$ and it follows from (4.4) that $e^{-rt}T(t) \xrightarrow{a.s.} r^{-1}W$ as $t \to \infty$. Recall that $W = 0$ if and only if the branching process goes extinct. Thus there exists $C \in \mathscr{F}$, with $C \subseteq A^c$ and $\mathbb{P}(C) = \mathbb{P}(A^c)$, such that for all $\omega \in C$,

$$e^{-rt}T(t,\omega) \to r^{-1}W(\omega) \quad \text{as } t \to \infty. \tag{7.3}$$

Let $\omega \in B \cap C$ and $c \in (0, \frac{1}{2}r^{-1})$. Then it follows from (7.3) that $T(c\log m, \omega) < 2m^{rc}r^{-1}W(\omega)$ for all sufficiently large m. Also, (7.2) 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 (7.1) follows since $\mathbb{P}(B \cap C) = \mathbb{P}(A^c)$. □

We turn now to estimation of λ_L. 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 $\hat{r}^{(m)} \xrightarrow[A^c]{a.s.} r$ as $m \to \infty$ where $\xrightarrow[A^c]{a.s.}$ 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 $\hat{\lambda}_{L,full}^{(m)}$ denote the estimator obtained by maximising the function $L_{full}(\lambda_L|\boldsymbol{a}, \hat{r}^{(m)})$ defined at (4.5). For ease of exposition, we assume that all infected households are observed, so, in our present notation, $a_{x,y}^{(m)} = Y_{n,x,y}^{(m)}(t^{(m)})$ for $(n,x,y) \in \mathscr{T}$. 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 λ_L by $\bar{\lambda}_L$.

Theorem 7.2 *Under the conditions of Theorem 7.1,*

$$\hat{\lambda}_{L,full}^{(m)} \xrightarrow[A^c]{a.s.} \bar{\lambda}_L \quad as\ m \to \infty.$$

Proof First note that from (4.5)

$$\hat{\lambda}_{L,full}^{(m)} = \text{argmax}\, \tilde{l}_{full}^{(m)}(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}), \tag{7.4}$$

where

$$\tilde{l}_{full}^{(m)}(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) = W^{-1}e^{-rt^{(m)}} \sum_{n=2}^{n_{max}} \sum_{(x,y)\in\mathscr{T}^{(n)}} Y_{n,x,y}^{(m)}(t^{(m)}) \log \tilde{p}_{x,y}^{(n)}(\hat{r}^{(m)}|\lambda_L).$$

Observe that,under the conditions satisfied by $(t^{(m)})$, Theorem 7.1 and (4.4) imply that, for all $(n,x,y) \in \mathscr{T}$,

$$W^{-1}e^{-rt^{(m)}} Y_{n,x,y}^{(m)}(t^{(m)}) \xrightarrow[A^c]{a.s.} \tilde{\alpha}_n \tilde{p}_{x,y}^{(n)}(r|\bar{\lambda}_L) \quad \text{as } m \to \infty. \tag{7.5}$$

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

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

where

$$\tilde{l}_{full}^{(\infty)}(\lambda_L|r) = \sum_{n=2}^{n_{max}} \tilde{\alpha}_n \sum_{(x,y)\in\mathscr{T}^{(n)}} \tilde{p}_{x,y}^{(n)}(r|\bar{\lambda}_L) \log \tilde{p}_{x,y}^{(n)}(r|\lambda_L). \tag{7.6}$$

Standard arguments, (e.g. Silvey [30], page 75) show that, for $n = 2,3,$ $\ldots, n_{max}$, the function $g_n(\lambda_L) = \sum_{(x,y)\in\mathscr{T}^{(n)}} \tilde{p}_{x,y}^{(n)}(r|\bar{\lambda}_L) \log \tilde{p}_{x,y}^{(n)}(r|\lambda_L)$ has a unique global maximum at $\bar{\lambda}_L$. Hence, as a function of $\lambda_L \in (0,\infty)$, $\tilde{l}_{full}^{(\infty)}(\lambda_L|r)$ has a unique global maximum at $\bar{\lambda}_L$.

Fix $0 < a < \bar{\lambda}_L < b < \infty$. Then

$$\max_{a\le\lambda_L\le b} |\tilde{l}_{full}^{(m)}(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) - \tilde{l}_{full}^{(\infty)}(\lambda_L|r)| \le \sum_{n=2}^{n_{max}} \sum_{(x,y)\in\mathscr{T}^{(n)}} \max_{a\le\lambda_L\le b} g_{n,x,y}^{(m)}(\lambda_L) \tag{7.7}$$

where

$$g_{n,x,y}^{(m)}(\lambda_L) = |W^{-1}e^{-rt^{(m)}} Y_{n,x,y}^{(m)}(t^{(m)}) \log \tilde{p}_{x,y}^{(n)}(\hat{r}^{(m)}|\lambda_L) - \tilde{\alpha}_n \tilde{p}_{x,y}^{(n)}(r|\bar{\lambda}_L) \log \tilde{p}_{x,y}^{(n)}(r|\lambda_L)|.$$

Now

$$g_{n,x,y}^{(m)}(\lambda_L) \le \hat{g}_{n,x,y}^{(m)}(\lambda_L) + \breve{g}_{n,x,y}^{(m)}(\lambda_L), \tag{7.8}$$

where

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

and

$$\breve{g}_{n,x,y}^{(m)}(\lambda_L) = |\{W^{-1}e^{-rt^{(m)}} Y_{n,x,y}^{(m)}(t^{(m)}) - \tilde{\alpha}_n \tilde{p}_{x,y}^{(n)}(r|\bar{\lambda}_L)\} \log \tilde{p}_{x,y}^{(n)}(r|\lambda_L)|.$$

Using (7.5), for all $(n,x,y) \in \mathscr{T}$,

$$\max_{a \le \lambda_L \le b} \check{g}^{(m)}_{n,x,y}(\lambda_L) \xrightarrow[A^c]{a.s.} 0 \quad \text{as } m \to \infty. \tag{7.9}$$

Further, for any $\lambda_L > 0$ and $r, r' > 0$,

$$|\tilde{p}^{(n)}_{x,y}(r|\lambda_L) - \tilde{p}^{(n)}_{x,y}(r'|\lambda_L)| \le \int_0^\infty |e^{-rt} - e^{-r't}| \, \mathrm{d}t = |r-r'|/(rr'), \tag{7.10}$$

so, since $\log x$ is uniformly continuous on any bounded subinterval of $(0,\infty)$ and $\hat{r}^{(m)} \xrightarrow[A^c]{a.s.} r$ as $m \to \infty$, it follows using (7.5) that, for all $(n,x,y) \in \mathscr{T}$,

$$\max_{a \le \lambda_L \le b} \hat{g}^{(m)}_{n,x,y}(\lambda_L) \xrightarrow[A^c]{a.s.} 0 \quad \text{as } m \to \infty. \tag{7.11}$$

Combining (7.4) - (7.9) yields

$$\max_{a \le \lambda_L \le b} |\tilde{l}^{(m)}_{full}(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) - \tilde{l}^{(\infty)}_{full}(\lambda_L|r)| \xrightarrow[A^c]{a.s.} 0 \text{ as } m \to \infty, \tag{7.12}$$

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

$$\operatorname*{argmax}_{a \le \lambda_L \le b} \tilde{l}^{(m)}_{full}(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) \xrightarrow[A^c]{a.s.} \bar{\lambda}_L \text{ as } m \to \infty. \tag{7.13}$$

To complete the proof we explore the behaviour of $l^{(m)}_{full}(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)})$ as $\lambda_L \downarrow 0$ and $\lambda_L \uparrow \infty$. Let X denote the time of the first point in $(0,\infty)$ of a homogeneous Poisson process having rate $(n-1)\lambda_L$. Then $p^{(n)}_{n-2,2}(t|\lambda_L) \le \mathbb{P}(X \le t) = 1 - e^{-(n-1)\lambda_L t}$, so

$$\tilde{p}^{(n)}_{n-2,2}(r|\lambda_L) \le \int_0^\infty (1 - e^{-(n-1)\lambda_L t}) e^{-rt} \, \mathrm{d}t \le (n-1)\lambda_L / r^2. \tag{7.14}$$

For all n, we have $\tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)}|\lambda_L) \le 1/\hat{r}^{(m)}$ for all $(x,y) \in \mathscr{T}^{(n)}$, so

$$\log \tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)}|\lambda_L) + \log \hat{r}^{(m)} \le 0. \tag{7.15}$$

Let

$$\begin{aligned} l^{(m)}_*(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) &= W^{-1} e^{-rt^{(m)}} \sum_{n=2}^{n_{max}} \sum_{(x,y) \in \mathscr{T}^{(n)}} Y^{(m)}_{n,x,y}(t^{(m)}) (\log \tilde{p}^{(n)}_{x,y}(\hat{r}^{(m)}|\lambda_L) + \log \hat{r}^{(m)}) \\ &= l^{(m)}_{full}(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) + W^{-1} e^{-rt^{(m)}} \sum_{n=2}^{n_{max}} \sum_{(x,y) \in \mathscr{T}^{(n)}} Y^{(m)}_{n,x,y}(t^{(m)}) \log \hat{r}^{(m)}, \end{aligned}$$

and, recalling (7.4), note that $\hat{\lambda}^{(m)}_{L,full} = \operatorname{argmax} l^{(m)}_*(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)})$.

Fix $\lambda_0 > 0$. Then (7.14) and (7.15) imply that, for all $\lambda_L \in (0, \lambda_0]$,

$$\begin{aligned} l^{(m)}_*(\lambda_L|\boldsymbol{Y}^{(m)}, \hat{r}^{(m)}) &\le W^{-1} e^{-rt^{(m)}} Y^{(m)}_{n,n-2,2}(t^{(m)}) (\log(n-1) + \log \lambda_0 - \log \hat{r}^{(m)}) \\ &\xrightarrow[A^c]{a.s.} \tilde{\alpha}_n \tilde{p}^{(n)}_{n-2,2}(r|\bar{\lambda}_L) [\log(n-1) + \log \lambda_0 - \log r] \end{aligned} \tag{7.16}$$

as $m \to \infty$. Also, using (7.5) and (7.12),

$$l_*^{(m)}(\bar{\lambda}_L|\boldsymbol{Y}^{(m)},\hat{r}^{(m)}) \xrightarrow[A^c]{a.s.} l_{full}^{(\infty)}(\bar{\lambda}_L|r) + r^{-1}\log r \sum_{n=2}^{n_{max}} \tilde{\alpha}_n \quad \text{as } m \to \infty. \tag{7.17}$$

Choose n such that $\alpha_n > 0$ and $\lambda_0 > 0$ such that the right hand side of (7.16) is strictly less than the right hand side of (7.17). Then, recalling since $\hat{\lambda}_{L,full}^{(m)} = \text{argmax}\, l_*^{(m)}(\lambda_L|\boldsymbol{Y}^{(m)},\hat{r}^{(m)})$, it follows that for P-almost all $\omega \in A^c$, there exists $m_0(\omega)$ such that

$$\hat{\lambda}_{L,full}^{(m)}(\omega) \notin (0,\lambda_0) \quad \text{for all } m \geq m_0(\omega). \tag{7.18}$$

Let T_I denote the infectious period of the initial infective in a household of size n. Then $p_{n-1,1}^{(n)}(t|\lambda_L) = \mathbb{E}[e^{-(n-1)\lambda_L t}\mathbb{1}_{\{T_I>t\}}] \leq e^{-(n-1)\lambda_L t}$, whence $\tilde{p}_{n-1,1}^{(n)}(r|\lambda_L) \leq 1/((n-1)\lambda_L + r)$. Arguing as before shows that there exists $\lambda_1 > 0$ such that, for P-almost all $\omega \in A^c$, there exists $m_1(\omega)$ such that

$$\hat{\lambda}_{L,full}^{(m)}(\omega) \notin (\lambda_1,\infty) \quad \text{for all } m \geq m_1(\omega). \tag{7.19}$$

The theorem then follows from (7.13), (7.18) and (7.19).

□

We now consider estimation of λ_L 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 $\hat{\lambda}_{L,rec}^{(m)}$ denote the estimator of λ_L obtained by maximising the function $L_{rec}(\lambda_L|\boldsymbol{c},\hat{r}^{(m)})$ described at (4.6). (In our present notation $c_j^{(n)} = Z_{n,j}^{(m)}(t^{(m)})$.)

Theorem 7.3 *Under the conditions of Theorem 7.1,*

$$\hat{\lambda}_{L,rec}^{(m)} \xrightarrow[A^c]{a.s.} \bar{\lambda}_L \quad \text{as } m \to \infty.$$

Proof First note from (4.6) that $\hat{\lambda}_{L,rec}^{(m)} = \text{argmax}\, \tilde{l}_{rec}^{(m)}(\lambda_L|\boldsymbol{Z}^{(m)},\hat{r}^{(m)})$, where

$$\tilde{l}_{rec}^{(m)}(\lambda_L|\boldsymbol{Z}^{(m)},\hat{r}^{(m)}) = W^{-1}e^{-rt^{(m)}} \sum_{n=2}^{n_{max}} \sum_{j=1}^{n} Z_{n,j}^{(m)}(t^{(m)}) \log \tilde{q}_j^{(n)}(\hat{r}^{(m)}|\lambda_L).$$

Using (7.5), for $n = 2,3,...,n_{max}$ and $j = 1,2,...,n$,

$$W^{-1}e^{-rt^{(m)}} Z_{n,j}^{(m)}(t^{(m)}) \xrightarrow[A^c]{a.s.} \tilde{\alpha}_n(r^{-1} - \tilde{q}_0^{(n)}(r|\bar{\lambda}_L))\tilde{q}_j^{(n)}(r|\bar{\lambda}_L) \quad \text{as } m \to \infty, \tag{7.20}$$

so, for any $\lambda_L \in (0,\infty)$, $\tilde{l}_{rec}^{(m)}(\lambda_L|\boldsymbol{Z}^{(m)},\hat{r}^{(m)}) \xrightarrow[A^c]{a.s.} \tilde{l}_{rec}^{(\infty)}(\lambda_L|r)$ as $m\to\infty$, where

$$\tilde{l}_{rec}^{(\infty)}(\lambda_L|r) = \sum_{n=2}^{n_{max}} \tilde{\alpha}_n(r^{-1}-\tilde{q}_0^{(n)}(r|\bar{\lambda}_L)) \sum_{j=1}^{n} \tilde{q}_j^{(n)}(r|\bar{\lambda}_L)\log\tilde{q}_j^{(n)}(r|\lambda_L). \tag{7.21}$$

Now

$$|\tilde{l}_{rec}^{(m)}(\lambda_L|\boldsymbol{Z}^{(m)},\hat{r}^{(m)}) - \tilde{l}_{rec}^{(\infty)}(\lambda_L|r)| \le \sum_{n=2}^{n_{max}}\sum_{j=1}^{n}(\hat{h}_{n,j}^{(m)}(\lambda_L)+\check{h}_{n,j}^{(m)}(\lambda_L)), \tag{7.22}$$

where $\hat{h}_{n,j}^{(m)}(\lambda_L) = W^{-1}e^{-rt^{(m)}}Z_{n,j}^{(m)}(t^{(m)})|\log\tilde{q}_j^{(n)}(\hat{r}^{(m)}|\lambda_L) - \log\tilde{q}_j^{(n)}(r|\lambda_L)|$ and $\check{h}_{n,j}^{(m)}(\lambda_L) = |\{W^{-1}e^{-rt^{(m)}}Z_{n,j}^{(m)}(t^{(m)}) - \tilde{\alpha}_n(r^{-1}-\tilde{q}_0^{(n)}(r|\bar{\lambda}_L))\tilde{q}_j^{(n)}(r|\bar{\lambda}_L)\}\log\tilde{q}_j^{(n)}(r|\lambda_L)|$. For $n=2,3,...,n_{max}$ and $j=1,2,...,n$,

$$\tilde{q}_j^{(n)}(r|\lambda_L) = \frac{\tilde{a}_j^{(n)}(r|\lambda_L)}{\tilde{a}_0^{(n)}(r|\lambda_L)},$$

where $\tilde{a}_j^{(n)}(r|\lambda_L) = \sum_{(x,y)\in\mathscr{A}_j^{(n)}}\tilde{p}_{x,y}^{(n)}(r|\lambda_L)$ $(j=1,2,...,n)$ and $\tilde{a}_0^{(n)}(r|\lambda_L) = r^{-1} - \sum_{y=1}^{n}\tilde{p}_{n-y,y}^{(n)}(r|\lambda_L)$. Note that $|\mathscr{A}_j^{(n)}| = n+1-j$ $(j=1,2,...,n)$. It follows from (7.10) that, for $n=2,3,...,n_{max}$ and $j=1,...,n$,

$$|\tilde{a}_j^{(n)}(r|\lambda_L) - \tilde{a}_j^{(n)}(r'|\lambda_L)| \le (n+1-j)|r-r'|/(rr'), \tag{7.23}$$

for all $\lambda_L > 0$.

Consider a household of size n. In the limit as $\lambda_L\to\infty$, as soon as one individual in the household is infected, the whole household becomes infected, so the number of removals in that household t time units after it was infected follows a binomial distribution with success probability $P^{(n)}(t) = \mathbb{P}(T_I \le t)$. It follows that, for $j=0,1,...,n$ and $r>0$, $\lim_{\lambda_L\to\infty}\tilde{a}_j^{(n)}(r|\lambda_L) \in (0,r^{-1}]$. Fix $a\in(0,\bar{\lambda}_L)$. It then follows from (7.20) and the continuity of $\tilde{a}_j^{(n)}(r|\lambda_L)$ that for $n=2,3,...,n_{max}$ and $j=1,2,...,n$,

$$\max_{a\le\lambda_L<\infty}\check{h}_{n,j}^{(m)}(\lambda_L) \xrightarrow[A^c]{a.s.} 0 \quad \text{as } m\to\infty, \tag{7.24}$$

Further, (7.23) and the uniform continuity of $\log x$ imply that, for $n=2,3,...,n_{max}$ and $j=1,2,...,n$,

$$\max_{a\le\lambda_L<\infty}\hat{h}_{n,j}^{(m)}(\lambda_L) \xrightarrow[A^c]{a.s.} 0 \quad \text{as } m\to\infty, \tag{7.25}$$

since $\hat{r}^{(m)} \xrightarrow[A^c]{a.s.} r$ as $m\to\infty$. Similar to before, (7.21) implies that $\tilde{l}_{rec}^{(\infty)}(\lambda_L|r)$ has a unique global maximum at $\lambda_L = \bar{\lambda}_L$. It follows using (7.22), (7.24) and (7.25), that, for any $a\in(0,\bar{\lambda}_L)$,

$$\operatorname*{argmax}_{a\le\lambda_L<\infty}\tilde{l}_{rec}^{(m)}(\lambda_L|\boldsymbol{Z}^{(m)},\hat{r}^{(m)}) \xrightarrow[A^c]{a.s.} \bar{\lambda}_L \quad \text{as } m\to\infty. \tag{7.26}$$

To complete the proof of the theorem, we obtain a uniform upper bound for $\tilde{l}_{rec}^{(m)}(\lambda_L|\boldsymbol{Z}^{(m)},\hat{r}^{(m)})$ for small λ_L. Two recoveries can occur in a household only if the initial infective has made at least one local infection, so, as at (7.14),

$$\hat{a}_2^{(n)}(r|\lambda_L) \leq \lambda_L(n-1)/r^2.$$

Also, there is at least one recovery in a household if the initial infective has recovered, so

$$\hat{a}_0^{(n)}(r|\lambda_L) \geq \int_0^\infty \mathbb{P}(T_I \leq t)e^{-rt}\,\mathrm{d}t = \phi(r)/r.$$

Hence, for $n = 2,3,\ldots,n_{max}$ and $\lambda_0 > 0$,

$$\tilde{q}_2^{(n)}(r|\lambda_L) \leq \lambda_0(n-1)/(r\phi(r)) \quad \text{for all } \lambda_L \in (0,\lambda_0].$$

Note that $\log\tilde{q}_j^{(n)}(r|\lambda_L) < 0$ for all n and j. We can now argue as in the derivation of (7.18) to show that λ_0 can be chosen so that, for P-almost all $\omega \in A^c$, there exists $m_2(\omega)$ such that

$$\hat{\lambda}_{L,rec}^{(m)}(\omega) \notin (0,\lambda_0) \quad \text{for all } m \geq m_0(\omega),$$

which, together with (7.26), completes the proof.

□

We omit the proofs but similar results to Theorems 7.1-7.3 hold for SEIR and Reed-Frost based models. Theorems 7.2 and 7.3 may also be extended to the case when the infectious period distribution has a parametric form with unknown parameters that need to be estimated. E.g. if the infectious period follows an exponential distribution with unknown rate γ, it is straightforward to show that, for any compact subset K of $(0,\infty)^2$, if (λ_L,γ) is estimated by maximising the relevant pseudolikelihood over K then the resulting estimator is strongly consistent. Extending this to $K = (0,\infty)^2$ is more complicated than in the one-dimensional setting of Theorems 7.2 and 7.3 and not considered here.

8 Concluding comments

In this paper we demonstrate that for an emerging SIR epidemic among a population partitioned into households, basing inference on the usual single-household final size distribution normally leads to a biased estimate of the within-household infection rate λ_L and use branching process theory to develop a new estimator which accounts correctly for the emerging nature of an epidemic. Although the model used is undoubtedly simpler than a real-life epidemic, the presence of households is a key departure from homogeneous mixing for human epidemics, and it seems likely that similar issues will arise in more complex settings when using data collected at a household level for inference during the exponentially growing phase of an outbreak. In particular, such data need to be modelled very carefully to ensure that the effects of a growing epidemic are incorporated correctly.

The new method is predicated upon the availability of an estimate of the exponential growth rate r. How best to estimate r for an emerging epidemic is an open

challenge (Ball *et al.* [6]) since, as illustrated by Figure 3, the exponentially growing phase occupies only a narrow time window and consequently care is required in choosing start and end time points for fitting it. Of course, the method assumes also that, at the time when estimation is performed, the epidemic is still in its exponentially growing phase and it should be checked that this is a reasonable assumption.

The new method has been shown to be computationally feasible under the assumption of no latent period and exponentially distributed infectious period. Extending its implementation to models with more realistic disease dynamics is an important area for research. One approach is via the phase method, see Section 4.2, though the matrices involved soon become large. Thus it would be worthwhile developing numerically amenable approximations to the key Laplace transforms $\tilde{p}^{(n)}_{x,y}(r|\lambda_L)$ $((n,x,y) \in \mathscr{T})$. Fraser [18] has developed a closed-form approximate method for calculating the growth rate r for quite general households models, which works well if both the maximum household size and the variance of the generation interval of the disease are not too large; it may be possible to apply related methods to approximate the aforementioned Laplace transforms.

It would be useful to attach standard errors to estimates obtained using the new method. One way of doing this is using a parametric bootstrap, along similar lines to Figure 4. Another approach is to determine the asymptotic distributions of the estimators, which would require central limit (or related) analogues of the almost sure results in Nerman [28].

The method can be extended to multitype SIR epidemics among a community of households, using the model of Ball and Lyne [8] together with multitype generalisations of Nerman [28]. This would accommodate age-stratified populations (e.g. children and adults), with age-specific susceptibilities, and also asymptomatic infections with different transmission parameters for symptomatic and asymptomatic cases. Note that the setting where all infectious episodes are governed by the same transmission parameters but infections are unobserved independently with a common parameter may be handled within the single-type framework, since the distribution of the number of observed cases in a households is obtained easily by conditioning on the total number of cases in that household and using binomial sampling.

The method can in principle also be extended to situations where information on the temporal progression of disease within households is available. In the Reed-Frost setting of Section 5, estimation can be generalised to the case when chains of infection within households are observed (rather than total number of cases) by extending the type-space of the approximating discrete-time multitype branching process to include such information. In the continuous-time setting of Section 4, suppose that inter-recovery times are observed. Consider the single-household epidemic $E^{(n)}_H$ described in Section 4.1, suppose that k recoveries occur in $(0,t]$, where $k = 1,2,\ldots,n$. Let t_1 denote the time of the first recovery and let $s_1,s_2,\ldots,s_k$ denote the k successive inter-recovery times, where s_k is the time elapsing between the kth recovery and t. Let $f^{(n)}_k(t_1,s_1,s_2,\ldots,s_{n-1}|\lambda_L)$ denote the joint-density of $s_1,s_2,\ldots,s_{k-1}$, including the information that no recovery occurs between the kth recovery and time t. Then using Theorem 5.4 of Nerman [28] shows that the contribution of such a household epidemic to the pseudolikelihood for λ_L is $\tilde{f}^{(n)}_k(\hat{r}|\lambda_L) =$

$\int_{t_A}^{\infty} e^{-\hat{r}t} f_k^{(n)}(t-t_A, s_1, s_2, \ldots, s_{n-1}|\lambda_L)\ \mathrm{d}t$, where $t_A = s_1 + s_2 + \cdots + s_k$, thus providing, at least in principle, a way of estimating λ_L.

Acknowledgements Laurence Shaw was supported by an EPSRC Doctoral Training grant.

References

1. D. J Aldous (1985) Exchangeability and related topics. Springer Lecture Notes in Math. 1117:1-198
2. S. Asmussen (1987) Applied Probability and Queues. Wiley, New York.
3. K. E. Athreya and P. E. Ney (1972) Branching Processes. Springer-Verlag, Berlin.
4. F. G. Ball (1986) A unified approch to the distribution of total size and total area under the trajectory of infectives in epidemic models. Adv Appl Prob, 18:289–310.
5. F. G. Ball (1996) Threshold behavior in stochastic epidemics among households. Athens Conference on Applied Probability and Time Series Analysis. Vol. I. Lecture Notes in Statist. (Springer, Berlin) 114:253–266.
6. F. G. Ball, T. Britton, T. House, V. Isham, D. Mollison, L. Pellis and G. Scalia Tomba (2014) Seven challenges for metapopulation models of epidemics, including households models. Epidemics, to appear.
7. F. G. Ball and P. J. Donnelly (1995) Strong approximations for epidemic models. Stoch Process Appl 55:1–21.
8. F. G. Ball and O. D. Lyne (2001) Stochastic multitype SIR epidemics among a population partitioned into households. Adv Appl Prob, 33:99–123.
9. F. G. Ball and O. D. Lyne (2002) Optimal vaccination schemes for stochastic epidemics among a population of households. Math Biosci 177&178:333–354.
10. F. G. Ball and O. D. Lyne (in preparation) Statistical inference for epidemics among a population of households.
11. F. G. Ball, D. Mollison and G. Scalia-Tomba (1997) Epidemics with two levels of mixing. Ann Appp Prob 7:46–89.
12. A. D. Barbour and S. Utev (2004) Approximating the Reed-Prost epidemic process. Stoch Process Appl 113:173–197.
13. N. G. Becker and K. Dietz (1995) The effect of the household distribution on transmission and control of highly infectious diseases. Math Biosci 127:207–219.
14. N. G. Becker and D. Starczak (1997) Optimal vaccination strategies for a community of households. Math Biosci 139:117–132.
15. S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron and P. Y. Boëlle (2004) A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statist Med 23:3469–3487.
16. S. Cauchemez, C. A. Donnelly, C. Reed, A. C. Ghani, C. Fraser, C. K. Kent, L. Finelli and N. M. Ferguson (2009) Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. N Engl J Med 361:2619–2627.
17. N. M. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Imasirithaworn and D. S. Burke (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature. 437:209–214.
18. C. Fraser (2007) Estimating individual and household reproduction numbers in an emerging epidemic. PLoS ONE 2(8):e758.
19. C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut, O. G. Pybus, H. Lopez-Gatell, C. M. Alpuche-Aranda, I. B. Chapela, E. P. Zavala, D. M. E. Guevara, F. Checchi, E. Garcia, S. Hugonnet and C. Roth (2009) Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324:1557–1561.
20. P. Haccou, P. Jagers and V. A. Vatutin (2005) Branching Processes: Variation, Growth, and Extinction of Populations. Cambridge University Press.
21. J. A. P. Heesterbeek and K. Dietz (1996) The concept of R_0 in epidemic theory. Statistica Neerlandica 50:89–110.
22. R. E. Hope-Simpson (1952) Infectiousness of communicable diseases in the household (measles, chickenpox and mumps). Lancet 260:549-554.

23. T. House, N. Inglis, J. V. Ross, F. Wilson, S. Suleman, O. Edeghere, G. Smith, B. Olowokure and M. J. Keeling (2012) Estimation of outbreak severity and transmissibility: Influenza (H1N1)pdm09 in households. BMC Medicine 10:117.
24. H. Kesten. and B. .P. Stigum (1966) A Limit Theorem for Multidimensional Galton-Watson Processes. The Ann Math Stat 37:1211–1223.
25. E. S. Knock and P. D. O'Neill (2014) Bayesian model choice for epidemic models with two levels of mixing. Biostatistics 15:46–59.
26. I. M. Longini. and J. S. Koopman (1982) Household and community transmission parameters from final distributions of infections in households. Biometrics 38:115–216.
27. J. Ma, J. Dushoff, B. M. Bolker and D. J. D Earn (2014) Estimating initial epidemic growth rates. Bull Math Biol 76:245–260.
28. O. Nerman (1981) On the convergence of supercritical gerneral (C-M-J) branching processes. Z. Warchscheinlichkeitstheorie, 57:365–395.
29. L. Pellis, N. M. Ferguson and C. Fraser (2011) Epidemic growth rate and household reproduction number in communities of households, schools and workplaces. J Math Biol 63:691–734.
30. S. D Silvey (1975) Statistical inference. Chapman and Hall, London.
31. G. Scalia Tomba, Å. Svensson, T. Asikainen and J. Giesecke (2010) Some model based considerations on observing generation times for communicable diseases. Math Biosci. 223:24–31.
32. J. Wallinga and M. Lipsitch (2007) How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc Lond B 274:599–604.
33. P. Whittle (1955) The outcome of a stochastic epidemic - a note on Bailey's paper. Biometrika 42:116–122.
34. *http://www.statistics.gov.uk/census/.*