Stat Biosci (2009) 1: 80–94
DOI 10.1007/s12561-009-9008-z

Estimating Temporal Transmission Parameters from Infectious Disease Household Data, with Application to Taiwan SARS Data

I-Shou Chang · Sian-Jhih Fu · Chung-Hsing Chen · Tsung-Hsi Wang · Chao A. Hsiung

Received: 27 February 2009 / Accepted: 3 June 2009 / Published online: 17 June 2009

Abstract Taking households having at least one infective as standard units and considering both a within-household infection rate and a global infection rate, we propose a Bayesian two level mixing S-I-R (susceptible-infective-removed) counting process model in which the transmission parameters may change over time and the parameters of interest are the within-household infection rate and the removal rate. Customized Markov chain Monte Carlo methods are developed for generating samples from the posterior distribution for inference purpose, based only on the removal times. The numerical performance of this method is examined in a simulation study. Applying this method to 2003 Taiwan SARS data, we find that the within-household infection rate decreases, the removal rate increases and their ratio is less than one and decreases significantly during the epidemic. This method allows the estimation of these parameters during the epidemic. For a rapidly transmitted disease, it provides a method to nearly real-time tracking of infection measures.

Keywords Counting process · MCMC methods · Metropolis-within-Gibbs algorithm · SARS · Transmission parameters

I-S. Chang and S.-J. Fu contribute equally and are joint first authors.

I-S. Chang
Institute of Cancer Research & Division of Biostatistics and Bioinformatics, National Health Research Institutes, Zhunan, Miaoli, Taiwan

S.-J. Fu
Center for Biomedical Databases, National Health Research Institutes, Zhunan, Miaoli, Taiwan

C.-H. Chen · C.A. Hsiung (✉)
Division of Biostatistics and Bioinformatics, National Health Research Institutes, 35, Keyan Road, Zhunan Town, Miaoli County 350, Taiwan
e-mail: hsiung@nhri.org.tw

T.-H. Wang
Centers for Disease Control, Taipei, Taiwan

1 Introduction

One of the focuses in recent studies of stochastic epidemic models is on models which feature a structured population. For example, when the disease transmission is from person to person and hence somewhat local in nature, cases will typically cluster according to some local community structure. The local community structure that deserves most attention has been the presence of small social groups such as households, schools and work places; the reason being not only that members of the same social group have a higher level of mixing, but also that household epidemic data are often easier to model and collect. Important contributions to the theory and applications using these models include [4, 7, 8, 10, 17, 20]. Works on outbreaks within households, in the presence of community infections, include [1, 4, 9, 24], among others.

A popular epidemic model for a community of households is the so-called two-level mixing model, discussed by Ball et al. [4]. This model assumes that, in a closed population partitioned into groups, an individual, during his infectious period, makes infectious contacts at population level at times given by the points of a Poisson process of rate λ_G, and hence those at individual level with a Poisson rate λ_G/N, where N is the population size; additionally, each infective individual makes an infectious contact with each individual in the same group with Poisson rate λ_L. We note that Becker and Hopper [9] considered a counting process model in which both between-households and within-household infection rates are explicitly described.

A mathematically simpler two-level mixing model was considered by Addy et al. [1]. Instead of using the aforementioned explicit global infection process for each infective, Addy et al. [1] utilize a single fixed probability that an individual avoids global infection.

In view of the population structure and the fact that control measures and their effects on transmission parameters may change over time, we propose in this paper a two-level mixing stochastic epidemic model in which the transmission parameters may change over time.

This work is motivated partly by the epidemiological study of SARS (severe acute respiratory syndrome) in Taiwan; see [29] and references therein. It was clear in the early period of the epidemic that it transmits from person to person through close contacts and that cases often cluster in households. As with other infectious diseases, it is of great interest to estimate the transmission parameters not only when the epidemic is over but also during the epidemic so as to assess the effects of various control measures and make suitable changes accordingly.

For an epidemic like SARS, cases are admitted to hospitals and are under quarantine or isolation there. For this reason, we assume that the removal times are available, but not the infection times.

We assume the population consists of disjoint households and we know the number of people in each household. The former assumption may be too restrictive, because some individuals cannot be uniquely assigned to a single household. To alleviate this problem, our analysis will be based on households having at least one infective. Although excluding households without any infective causes bias in the estimation of global infection rate, it is still useful in the study of within-household

infection and removal rate. Effects of control measures can then be studied in household setting.

Real-time tracking of control measures for emerging infections is gaining attention recently; see the commentary by Liopsitch and Bergstrom [22], for example. Wallinga and Teunis [28] proposed an elegant method for the estimation of instantaneous reproductive number R_t as it evolves over time in an epidemic. Later developments in this line include [12, 13]. Generally speaking, these methods require the knowledge of generation time, which is assumed to be invariant over calendar time, or need contact tracing data, in addition to daily case counts. Another method for the real-time estimation of the basic reproduction number R_0 was proposed by White and Paganno [30], which makes use of certain parametric model assumptions. Our method provides real-time estimation of the within-household infection rate and removal rate in a two-level mixing model; it seems to be useful in the real-time tracking of control measures within households.

In this paper, the dynamics of the transmission will be modeled by counting processes, as explained in [3, 6], for example. We construct a two-level mixing S-I-R (susceptible-infective-removed) counting process model for each household. There are two infectivity terms in the model: the global one is described by a deterministic intensity function in the spirit of Addy et al. [1], and the local one has not only a deterministic factor but also a random factor depending on the number of infectives in the same household. Both the global intensity and deterministic factor of the local term are allowed to change over time.

The inference is carried out within a Bayesian framework using carefully designed model-specific MCMC methods. In particular, concepts from both Gibbs sampler and Metropolis–Hastings reversible jump algorithms are used in designing the algorithm for sampling from the posterior distribution.

This paper is organized as follows. The model and the likelihoods are presented in Sect. 2. The details of the MCMC algorithm are in Sect. 3. Section 4 contains a simulation study to assess the numerical performance of our method. Section 5 applies our method to Taiwan SARS data, which shows that the within-household infectivity decreases, the removal rate increases and their ratio is less than 1 and decreases significantly throughout the period starting on March 18. Finally, Sect. 6 is a discussion on future investigations.

2 The Model and the Complete Data Likelihood

2.1 The Model

We consider a population consisting of households and suffering a transmissible disease. At any time point, each individual is assumed to be in one of the following three states: susceptible (S), infectious (I), or removed (R). Models using this assumption are called S-I-R models. A susceptible individual is healthy, may contract the disease in question and become an infective. An infectious individual or an infective is one who has become infected and can transmit the disease to others. A removed individual is one who plays no part in further disease spread; this could occur either

by actual immunity or by isolation following the appearance of symptoms. During the infectious period, an infective makes random contacts with others; if a contacted individual is susceptible, then he or she becomes infectious and is immediately able to infect other individuals.

We assume that individuals are homogeneous and mixed uniformly within each household, and no individual belongs to two different households. Suppose there are M households for which there is at least one infected member in each of these households. For $m = 1, \ldots, M$, let T_0^m denote the earliest time that an infective appears in the mth household, let $S_m(t)$, $I_m(t)$ and $R_m(t)$ respectively denote the number of susceptible, infectious and removed individuals at time $t \geq T_0^m$ in the mth household, where the time is calendar time. Let $N_m(t) = S_m(T_0^m) - S_m(t)$, denoting the number of individuals in the m-th household infected in the time interval $(T_0^m, t]$. For the sake of mathematical convenience, we assume $T_0^m \geq 0$ and $S_m(.)$, $I_m(.)$ and $R_m(.)$ all have right-continuous sample paths on $[0, \infty)$, which implies the total number of individuals in the m-th household is $S_m(T_0^m) + 1$. We assume $S_m(T_0^m)$ is known, and no two individuals are infected at the same time, as is usually assumed in counting process models.

Let $\mathcal{G}_t^m$ denote the σ-field generated by $\{S_m(u), I_m(u) \mid T_0^m \leq u \leq t\}$; it is equal to that generated by $\{N_m(u), R_m(u) \mid T_0^m \leq u \leq t\}$, since $S_m(t) + I_m(t) + R_m(t) = S_m(T_0^m) + 1$ for every $t \geq T_0^m$. Let $\mathcal{G}_t = \sigma\{\mathcal{G}_t^m \mid m = 1, \ldots, M\}$, the σ-field generated by all $\mathcal{G}_t^m$. We assume $\{N_m, R_m \mid m = 1, \ldots, M\}$ is a multivariate counting process and has the following intensities:

$$
\begin{aligned}
&Pr\big(N_m(t+h) - N_m(t) = 1, R_m(t+h) - R_m(t) = 0 \mid \mathcal{G}_t\big) \\
&\quad = h\alpha(t) S_m(t-) + h\beta(t) I_m(t-) \bar{S}_m(t-) + o(h), \qquad (1) \\
&Pr\big(N_m(t+h) - N_m(t) = 0, R_m(t+h) - R_m(t) = 1 \mid \mathcal{G}_t\big) \\
&\quad = h\gamma(t) I_m(t-) + o(h), \qquad (2) \\
&Pr\big(N_m(t+h) - N_m(t) = 0, R_m(t+h) - R_m(t) = 0 \mid \mathcal{G}_t\big) \\
&\quad = 1 - h\alpha(t) S_m(t-) - h\beta(t) I_m(t-) \bar{S}_m(t-) - h\gamma(t) I_m(t-) + o(h). \qquad (3)
\end{aligned}
$$

Here $\bar{S}_m(t) = \frac{S_m(t)}{S_m(T_0^m)}$, and

$$
\begin{aligned}
\alpha(t) &= \frac{\alpha_1}{1 + \alpha_2 \exp(-\alpha_3 t)}, \\
\beta(t) &= \frac{\beta_1}{1 + \beta_2 \exp(-\beta_3 t)}, \\
\gamma(t) &= \frac{\gamma_1}{1 + \gamma_2 \exp(-\gamma_3 t)},
\end{aligned}
$$

for some positive real numbers α_1, α_2, β_1, β_2, γ_1 and γ_2 and real numbers α_3, β_3 and γ_3. Concepts of multivariate counting processes and their intensities can be found, for example, in [2, 11]. It is for the sake of convenience that we use logistic functions α, β and γ in the above modeling; other choices are possible.

The assumptions (1), (2) and (3) provide an epidemic model that allows two levels of mixing. Specifically, the first term on the right-hand side of (1) describes the global infection exerted on a susceptible by a deterministic intensity $\alpha(t)$; the second term there describes the infectivity an infective makes on the susceptible individuals in the same household. We remark that expressing infection by contact rate, the second term keeps the rate at which a given infective makes contact with other individuals in the same household unchanged relative to the household size.

Typically, we expect $\beta(t)$ to be larger than $\alpha(t)$, because of more frequent contacts in a household. In our model, infection rates and removal rates may vary from household to household, but if the epidemics in different households occur at the same time, they have the same rates. Since the total period of the epidemic in each household may be much shorter than the span of the epidemic in a region like a country or a large city, our model provides a way to describe the effect of control measures of the health authority during the epidemic in the region.

Because we define local infection in terms of $\bar{S}$, instead of S, which is different from that in [4], it is scaled by household size and hence the quantity $\eta(t) = \beta(t)/\gamma(t)$ may be regarded as the household reproduction number at time t and can be used to measure the seriousness of an infectious disease in households. For a homogeneous and uniformly mixed large community, a parameter of primary concern is the so-called basic reproduction number R_0, which is the average number of new infections caused by a "typical" infective during the early period of the epidemic; the threshold limit theorem roughly states that for an epidemic in this community, either only few individuals will ever become infected, or a positive proportion of the susceptibles will have been infected by the end of the epidemic; see, for example, [3]. Although it is risky to conclude anything like the threshold limit theorem in our situation from $\eta(t)$, it nonetheless seems appropriate to view it as a measure of the seriousness of the epidemic.

2.2 The Likelihood

Let τ denote a time point, possibly the time of observation. It follows from the likelihood formula for counting process (see, for example, [11, p. 187]) that, conditional on $(T_0^1, \ldots, T_0^M)$, the log-likelihood of $\{N_m(t), R_m(t) \mid t \le \tau, m = 1, \ldots, M\}$ is

$$\begin{aligned}\sum_{m=1}^{M}\Bigg[&\int_{T_0^m}^{\tau} \log\big(\alpha(t)S_m(t-) + \beta(t)I_m(t-)\bar{S}_m(t-)\big)\,dN_m(t)\\ &+ \int_{T_0^m}^{\tau} \log\big(\gamma(t)I_m(t-)\big)\,dR_m(t) - \int_{T_0^m}^{\tau} \alpha(t)S_m(t-)\,dt\\ &- \int_{T_0^m}^{\tau} \beta(t)I_m(t-)\bar{S}_m(t-)\,dt - \int_{T_0^m}^{\tau} \gamma(t)I_m(t-)\,dt\Bigg]. \qquad (4)\end{aligned}$$

This paragraph introduces some notation useful in the following Bayesian inference. Let $T_1^m < T_2^m < \cdots < T_{N_m(\tau)}^m$ denote all the infection times in the mth household in $(T_0^m, \tau]$ and let $T^m = (T_1^m, \ldots, T_{N_m(\tau)}^m)$. Similarly, let $Q_1^m < Q_2^m <$

$\cdots < Q^m_{R_m(\tau)}$ denote all the removal times in the mth household in $(T^m_0, \tau]$, and let $Q^m = (Q^m_1, \ldots, Q^m_{R_m(\tau)})$. We note that T^m and Q^m jointly satisfy the compatibility condition that $N_m(\tau) + 1 \geq R_m(\tau)$ and $T^m_{i-1} < Q^m_i$ for every $i = 1, \ldots, R_m(\tau)$. We note that, given $(T^1_0, \ldots, T^M_0)$, the data $\{N_m(t), R_m(t) \mid t \leq \tau, m = 1, \ldots, M\}$ are equivalent to the data $\{T^m, Q^m \mid m = 1, \ldots, M\}$. Thus, the conditional log-likelihood (4) can be expressed in terms of the data $\{T^m, Q^m \mid m = 1, \ldots, M\}$ and is denoted by $\sum_{m=1}^M l^c_m(T^m, Q^m \mid T^m_0, \theta)$, where $\theta = (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3)$ is in $\Theta = (\mathcal{R}_+ \times \mathcal{R}_+ \times \mathcal{R})^3$, with $\mathcal{R}_+$ being the set of positive real numbers. Let $T = \{T^1, \ldots, T^M\}$, $T_0 = \{T^1_0, \ldots, T^M_0\}$, $Q = \{Q^1, \ldots, Q^M\}$, $L^c_m = e^{l^c_m}$. Let $L^c(T, Q \mid T_0, \theta) = \prod_{m=1}^M L^c_m(T^m, Q^m \mid T^m_0, \theta)$, which is the complete data conditional likelihood given T_0.

3 Bayesian Inference Based on Removal Times

We propose to make Bayesian inference on the transmission parameters using the likelihood (4), assuming a prior density μ_m on T^m_0 for $m = 1, \ldots, M$, and treating the other infection times as missing values. We assume that the inference is conducted during the epidemic and hence we do not know if the epidemic is over at the time τ. We say the epidemic is over at t if $\sum_{m=1}^M I_m(t) = 0$.

Given a prior π on the parameter space Θ, our task is to sample from $V(\theta \mid Q)$, the posterior density of θ given Q. We assume π is a product measure; namely, $\pi = \pi_{\alpha_1} \times \cdots \times \pi_{\gamma_3}$, with π_{α_i} the prior of α_i, π_{β_i} the prior of β_i and π_{γ_i} the prior of γ_i, for $i = 1, 2, 3$.

Since both the joint density of (Q, θ) and the conditional density of Q given θ are hard to simulate, we propose to generate samples from $V(\theta \mid Q)$ by the following hybrid MCMC, also known as Metropolis-within-Gibbs algorithm. Our approach takes advantage of the fact that the complete data conditional likelihood given T_0 is computationally tractable.

Let $V_1(\theta \mid T_0, T, Q)$, $V_2(T_0 \mid \theta, T, Q)$ and $V_3(T \mid \theta, T_0, Q)$ denote respectively the conditional density of θ given $\{T_0, T, Q\}$, of T_0 given $\{\theta, T, Q\}$ and of T given $\{\theta, T_0, Q\}$. Let $(\theta^{(n)}, T^{(n)}_0, T^{(n)})$ be the current state of the Markov chain, where $\theta^{(n)}$ is in Θ, $T^{(n)}_0 = \{\ldots, T^{m(n)}_0, \ldots\}$ with $T^{m(n)}_0$ representing the first infection time in the mth household, and $T^{(n)} = \{\ldots, T^{m(n)}, \ldots\}$ with $T^{m(n)}$ representing all the infection times in the mth household other than the first one. The Gibbs sampler (see, for example, [27, p. 372]) suggests that we iteratively first generate $\theta^{(n+1)}$ according to $V_1(\theta \mid T^{(n)}_0, T^{(n)}, Q)$, then generate $T^{(n+1)}_0$ according to $V_2(T_0 \mid \theta^{(n+1)}, T^{(n)}, Q)$, and finally generate $T^{(n+1)}$ according to $V_3(T \mid \theta^{(n+1)}, T^{(n+1)}_0, Q)$; when n is large enough, $\theta^{(n)}$ would be an approximate sample of $V(\theta \mid Q)$. For generating data from $V_1(\theta \mid T_0, T, Q)$, $V_2(T_0 \mid \theta, T, Q)$ and $V_3(T \mid \theta, T_0, Q)$, we make use of Metropolis–Hastings algorithms. This is the basic idea of Metropolizing the Gibbs sampler, whose merits are discussed in [27, pp. 392–394]. See also [16] for more details.

Making use of

$$V_1(\theta \mid T_0, T, Q) \propto L^c(T, Q \mid T_0, \theta)\pi(\theta), \tag{5}$$

$$V_2(T_0 \mid \theta, T, Q) \propto L^c(T, Q \mid T_0, \theta)\mu(T_0), \tag{6}$$

$$V_3(T \mid \theta, T_0, Q) \propto L^c(T, Q \mid T_0, \theta), \tag{7}$$

we update the Markov chain by the following steps. Here μ is the product measure $\mu_1 \times \cdots \times \mu_M$. Care is taken so that the compatibility condition described in Sect. 2.2 is always satisfied.

(a) Metropolizing (5). Let $\theta^{(n)} = (\alpha_1^{(n)}, \alpha_2^{(n)}, \alpha_3^{(n)}, \beta_1^{(n)}, \beta_2^{(n)}, \beta_3^{(n)}, \gamma_1^{(n)}, \gamma_2^{(n)}, \gamma_3^{(n)})$ denote the current values of the parameters; we update these nine parameters one by one in the order of the coordinates. For example, when $\alpha_1^{(n)}, \alpha_2^{(n)}$ and $\alpha_3^{(n)}$ have been updated to $\alpha_1^{(n+1)}, \alpha_2^{(n+1)}$ and $\alpha_3^{(n+1)}$, the current value $\beta_1^{(n)}$ is updated by the following Metropolis–Hastings algorithm.

(i) Generate $\beta_1^{(n+1)'}$ according to π_{β_1}.

(ii) Let

$$\rho = \min\left\{ \frac{L^c(T^{(n)}, Q \mid T_0^{(n)}, \alpha_1^{(n+1)}, \alpha_2^{(n+1)}, \alpha_3^{(n+1)}, \beta_1^{(n+1)'}, \beta_2^{(n)}, \beta_3^{(n)}, \gamma_1^{(n)}, \gamma_2^{(n)}, \gamma_3^{(n)})}{L^c(T^{(n)}, Q \mid T_0^{(n)}, \alpha_1^{(n+1)}, \alpha_2^{(n+1)}, \alpha_3^{(n+1)}, \beta_1^{(n)}, \beta_2^{(n)}, \beta_3^{(n)}, \gamma_1^{(n)}, \gamma_2^{(n)}, \gamma_3^{(n)})}, 1 \right\};$$

we set $\beta_1^{(n+1)}$ to be $\beta_1^{(n+1)'}$ with probability ρ, and to be $\beta_1^{(n)}$ with probability $1-\rho$. Similar algorithms are used to update other parameters.

(b) Metropolizing (6). Let $T_0^{1(n)}, \ldots, T_0^{M(n)}$ be the current state of the initial infection times in these households. For $m = 1, \ldots, M$, we update $T_0^{m(n)}$ as follows.

(i) Generate $T_0^{m(n+1)'}$ according to the prior distribution μ_m.

(ii) Let

$$\rho = \min\left\{ \frac{L_m^c(T^{m(n)}, Q^m \mid T_0^{m(n+1)'}, \theta^{(n+1)})}{L_m^c(T^{m(n)}, Q^m \mid T_0^{m(n)}, \theta^{(n+1)})}, 1 \right\};$$

we set $T_0^{m(n+1)}$ to be $T_0^{m(n+1)'}$ with probability ρ, and to be $T_0^{m(n)}$ with probability $1-\rho$.

(c) Metropolizing (7). Because we do not know if the epidemic is over in a household, the number of infected at any given time point is only known to be no less than the number of removals up to that time point. This compatibility condition implies that we have a variable dimension situation; the reversible jump algorithm [19] is adapted to this part, which is similar to the algorithm in [26].

Some more notation is in order. Let δ be a real number in T^m, a finite increasing sequence of real numbers; we denote by $T^m - \delta$ the subset of T^m with δ excluded. Let ζ be a real number not in T^m; we denote by $T^m + \zeta$ the finite increasing sequence of real numbers consisting of T^m and ζ. Let $d(m(n))$ denote the number of infection times in $T^{m(n)}$ and call it its length. For each $m = 1, \ldots, M$, the transition from $T^{m(n)}$ to $T^{m(n+1)}$ is described in the following.

Our algorithm allows only three possible types of transitions: H, H^+ and H^- from $T^{m(n)}$ to $T^{m(n+1)}$. Here H is a transition satisfying $d(m(n+1)) = d(m(n))$, H^+ satisfying $d(m(n+1)) = d(m(n)) + 1$, and H^- satisfying $d(m(n+1)) = d(m(n)) - 1$. Let φ_{d_1,d_2} denote the probability of selecting $T^{m(n+1)}$ having length

d_2 given $T^{m(n)}$ having length d_1. Thus we consider φ_{d_1,d_2} that satisfies $\varphi_{d_1,d_2} = 0$ if $|d_2 - d_1| > 1$. We note that there are two constraints on the possible transition types: one is the compatibility condition and the other is the household size. In case all three types of transitions are possible, let $\varphi_{d_1,d_1} = \varphi_{d_1,d_1-1} = \varphi_{d_1,d_1+1} = \frac{1}{3}$; in case only two are possible, then each has probability $\frac{1}{2}$. For example, if $d_1 = 1$ and there are two removed individuals in this household then compatibility condition implies $d_2 = 1$ or 2 and hence $\varphi_{1,1} = \frac{1}{2}$, $\varphi_{1,2} = \frac{1}{2}$.

If the transition type H is selected, then a randomly chosen infection time in $T^{m(n)}$ is replaced by a number chosen randomly from Uniform$(T_0^{m(n+1)}, \tau)$. Denote the removed number by r and the added number by a. Let $T^{m(n+1)} = T^{m(n)} - r + a$. If $T^{m(n+1)}$ and Q^m together satisfy the compatibility condition, we set

$$\rho = \min\left\{ \frac{L_m^c(T^{m(n)} - r + a, Q^m \mid T_0^{m(n+1)}, \theta^{(n+1)})\varphi_{d(m(n+1)),d(m(n))}}{L_m^c(T^{m(n)}, Q^m \mid T_0^{m(n+1)}, \theta^{(n+1)})\varphi_{d(m(n)),d(m(n+1))}}, 1 \right\}.$$

Otherwise, $\rho = 0$.

We let $T^{m(n+1)} = T^{m(n)} - r + a$ with probability ρ, and $T^{m(n+1)} = T^{m(n)}$ with probability $1 - \rho$.

If the transition type H^- is chosen, we randomly remove one of the infection times in $T^{m(n)}$ and denote it by r. Let $T^{m(n+1)} = T^{m(n)} - r$. If $T^{m(n+1)}$ and Q^m together satisfy the compatibility condition, we set

$$\rho = \min\left\{ \frac{L_m^c(T^{m(n)} - r, Q^m \mid T_0^{m(n+1)}, \theta^{(n+1)})\varphi_{d(m(n+1)),d(m(n))} d(m(n))}{L_m^c(T^{m(n)}, Q^m \mid T_0^{m(n+1)}, \theta^{(n+1)})\varphi_{d(m(n)),d(m(n+1))}(\tau - T_0^{m(n+1)})}, 1 \right\}.$$

Otherwise, $\rho = 0$.

Let $T^{m(n+1)} = T^{m(n)} - r$ with probability ρ, and $T^{m(n+1)} = T^{m(n)}$ with probability $1 - \rho$.

If the transition type H^+ is chosen, we generate a according to Uniform$(T_0^{m(n+1)}, \tau)$. We define $T^{m(n+1)} = T^{m(n)} + a$ with probability

$$\rho = \min\left\{ \frac{L_m^c(T^{m(n)} + a, Q^m \mid T_0^{m(n+1)}, \theta^{(n+1)})\varphi_{d(m(n+1)),d(m(n))}(\tau - T_0^{m(n+1)})}{L_m^c(T^{m(n)}, Q^m \mid T_0^{m(n+1)}, \theta^{(n+1)})\varphi_{d(m(n)),d(m(n+1))}(d(m(n)) + 1)}, 1 \right\},$$

and $T^{m(n+1)} = T^{m(n)}$ with probability $1 - \rho$.

4 A Simulation Study

This section examines the numerical performance of the method in Sect. 3. Section 4.1 describes the way the data are generated and Sect. 4.2 reports the simulation results.

4.1 Data Generation for Point Processes

We describe a method to generate data from one household satisfying (1), (2), and (3); we omit the subscript m to simplify the notation.

Let Y_0 denote the first infection time in the household. For $n = 1, 2, \ldots$, let

$$Y_n = \inf\{t > Y_0 \mid N(t) + R(t) = n\},$$

$$Z_n = \begin{cases} 1 & \text{if } N(Y_n) - N(Y_n-) = 1, \\ 2 & \text{if } R(Y_n) - R(Y_n-) = 1, \end{cases}$$

$$X_{n+1} = \begin{cases} Y_{n+1} - Y_n & \text{if } Y_n < \infty, \\ \infty & \text{if } Y_n = \infty. \end{cases}$$

Without loss of generality, we assume $Y_0 = 0$ and $Z_0 = 1$. We note that Y_n is called the nth event time, Z_n the mark associated to the event time Y_n, and X_n an inter-occurrence time.

For $n = 1, 2, \ldots$, we generate the sequence (Y_n, Z_n) iteratively by first generating X_{n+1} and then Z_{n+1}. Given $Y_1, Z_1, \ldots, Y_n, Z_n$, we generate X_{n+1} by utilizing the fact that the conditional cumulative hazard of X_{n+1} at $t > Y_n$ is $\int_{Y_n}^{t} [\alpha(u)S(Y_n) + \beta(u)I(Y_n)\bar{S}(Y_n) + \gamma(u)I(Y_n)]\,du$. Let $G^{(n+1)}$ denote the conditional cumulative distribution of X_{n+1}, given Y_n. Then given $Y_1, Z_1, \ldots, Y_n, Z_n$, the probability of $Z_{n+1} = 2$ is equal to $\int_{Y_n}^{\infty} \gamma(t)I(Y_n)(1 - G^{(n+1)}(t - Y_n))\,dt$. Derivations of these formulas can be found in [11, p. 61].

4.2 Bayesian Inference

The parameters in this simulation study resemble those from the analysis of Taiwan SARS data. We assume there are $M = 400$ households, each household has six people, and the time of the first infection in each of these households is chosen randomly from Uniform[0, 100]. The analysis is based on the removal times observed in the interval [0, 100] and we do not assume the epidemic is over.

We carry out the inference by the MCMC algorithm outlined in Sect. 3. The priors are given by $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2 \sim \text{Exp}(10)$, $\alpha_3, \beta_3, \gamma_3 \sim N(0, 0.25)$. We note that these priors are meant to be more or less noninformative with large variances; although the prior variance of α_3 is 0.25, the corresponding variance of $\alpha(t)$ is more than 50 for t larger than 20. Table 1 gives means, medians, standard deviations of the prior distributions of $\alpha(t)$, $\beta(t)$ and $\gamma(t)$ for some time points t. The true parameter values for data generation are in the second row of Table 2. In the analysis of SARS data in Sect. 5, we introduce the prior of the first infection time in a household by means of the first fever date in that household. In this simulation study, the prior of the first infection time in a household is similarly assumed to be $t + 1 - \text{Exp}(1)$ distributed on the interval $(-\infty, t + 1]$, where t is the first infection time in the household, and $t + 1$ is the assumed fever date.

We followed the suggestions in [18, pp. 296–297], closely to implement the MCMC algorithm for a given randomly generated data set. Namely, we ran five Markov chains with randomly chosen initial values, updated each of the chains

Table 1 Means, medians and standard deviations of the prior distributions of $\alpha(t)$, $\beta(t)$ and $\gamma(t)$

Time (t)	20	40	60	80	100
prior mean	4.27	4.60	4.71	4.78	4.81
prior median	0.56	0.37	0.30	0.26	0.22
prior s.d.	7.64	8.05	8.19	8.26	8.30

Table 2 Estimation of parameters in simulation study based on 100 data sets

	α_1	α_2	α_3	β_1	β_2	β_3	γ_1	γ_2	γ_3
true values	1	15	−0.5	1	13	−0.03	1	20	0.03
posterior mean	2.54	17.25	−0.69	1.36	17.78	−0.022	1.05	23.65	0.034
posterior median	1.56	14.33	−0.66	1.09	14.72	−0.032	0.92	21.23	0.033
posterior s.d.	2.99	12.59	0.29	1.05	13.09	0.036	0.51	9.93	0.006

800,000 times and calculated the Gelman–Rubin statistics $\hat{R}$ for each of the nine parameters with the initial 400,000 updates as burn-ins. We found all the $\hat{R}$s are less than 1.1. Based on this experiment, we analyze each of the following 100 data sets by running only one chain with 800,000 updates and 400,000 burn-ins; the posterior distributions are based on the latter 400,000 updates. By the way, our program is coded in C, it takes about 4 hours to finish the analysis of one data set and we conducted parallel computing to finish this simulation study.

Tables 2 and 3 contain the results for 100 simulated data sets. The third row of Table 2 reports the average of the 100 posterior means of the coefficients; the fourth row reports the average of the 100 posterior medians; the fifth row reports the average of the 100 posterior standard deviations. The results for some of the values of the functions α, β, γ and $\eta = \beta/\gamma$ are contained in Table 3. Rows in these tables carry similar means as those in Table 2. For example, the third row in Table 3 provides the average of the 100 posterior means of the values of α at 20, 40, 60, 80 and 100. These tables indicate that although estimates regarding the function α are less accurate, those regarding functions β, γ and η are generally excellent, which are parameters of interest. We note that these values of α are larger than their true values, as expected, because only households having infectives are included. We also note that especially for the function values, the posterior standard deviations are much smaller than the corresponding prior standard deviations.

5 Application to Taiwan SARS Data

We now illustrate the method of this paper by analyzing Taiwan SARS data. Early in the global outbreak of the 2003 SARS, modes of transmission were unclear; control measures were implemented to contain this highly contagious emerging disease; in Taiwan, for example, quarantine started on March 18 and extensive fever screening started in April. On July 5, 2003, Taiwan was removed from the World Health Organization (WHO) list of SARS-affected countries. More detailed information regarding Taiwan SARS can be found in [29] and references therein.

Table 3 Estimation of $\alpha(t)$, $\beta(t)$, $\gamma(t)$ and $\eta(t)$ in simulation study based on 100 data sets

	Time (t)	20	40	60	80	100
α	true value	3.02×10^{-6}	1.37×10^{-10}	6.24×10^{-15}	2.83×10^{-19}	1.29×10^{-23}
	posterior mean	5.55×10^{-5}	2.59×10^{-6}	3.25×10^{-7}	6.70×10^{-8}	2.01×10^{-8}
	posterior median	4.34×10^{-6}	6.90×10^{-9}	1.45×10^{-11}	3.18×10^{-14}	6.85×10^{-17}
	posterior s.d.	1.92×10^{-4}	2.11×10^{-5}	4.41×10^{-6}	1.39×10^{-6}	6.46×10^{-7}
β	true value	0.041	0.023	0.013	0.007	0.004
	posterior mean	0.038	0.021	0.012	0.007	0.004
	posterior median	0.038	0.021	0.011	0.006	0.004
	posterior s.d.	0.006	0.003	0.003	0.003	0.003
γ	true value	0.083	0.142	0.232	0.355	0.501
	posterior mean	0.077	0.136	0.226	0.342	0.472
	posterior median	0.077	0.136	0.225	0.342	0.472
	posterior s.d.	0.007	0.008	0.015	0.026	0.056
η	true value	0.485	0.159	0.054	0.020	0.008
	posterior mean	0.499	0.152	0.052	0.020	0.009
	prior mean	1.57×10^{11}	8.01×10^{24}	4.39×10^{39}	3.24×10^{54}	2.40×10^{69}
	posterior median	0.495	0.151	0.051	0.019	0.008
	prior median	1.02	1.02	0.99	1.03	1.00
	posterior s.d.	0.085	0.026	0.015	0.010	0.007
	prior s.d.	7.56×10^{12}	5.87×10^{26}	4.31×10^{41}	3.24×10^{56}	2.40×10^{71}

Table 4 (SARS) The household sizes of the 399 households

Household size	1	2	3	4	5	6	7	8	9	10	11	15
number of households	50	66	59	107	58	30	15	6	3	2	2	1

As an illustration of the method of this paper, we analyze the data whose disease onset times are no earlier than March 18, because of the data tractability. For the period from February 24 to July 5, there are 664 reported probable cases of SARS in Taiwan. Among them, 440 cases appeared after March 18, have recorded date of admission to SARS-designated hospitals, and can be assigned unambiguously to 399 households with known size. Table 4 gives the frequencies of the sizes of these 399 households. Table 5 gives the frequencies of the final sizes of the outbreaks in these 399 households. For each case, we have available not only the time he/she is admitted to a SARS-designated hospital but also the time he/she starts to have fever; the former is used as the removal time and the latter is used to define the prior on the first infection time in each household. The following analysis is based on the data of these 440 patients.

We analyze the data using the same method in Sect. 4. The first infection time in a household is assumed to be $t - \text{Exp}(1)$ distributed on the interval $(-\infty, t]$, where t

Table 5 (SARS) The final sizes of the outbreaks in the 399 households

Final size	1	2	3	4
number of households	370	20	6	3

Table 6 (SARS) Estimation of parameters

	α_1	α_2	α_3	β_1	β_2	β_3	γ_1	γ_2	γ_3
posterior mean	1.68	17.80	−0.60	1.55	15.85	0.020	1.55	23.77	0.035
posterior median	0.85	14.73	−0.56	1.14	12.72	−0.029	1.35	21.30	0.033
posterior s.d.	2.33	13.08	0.30	1.52	12.57	0.18	0.88	11.17	0.012

Table 7 (SARS) Estimation of $\alpha(t)$, $\beta(t)$, $\gamma(t)$, and $\eta(t)$

	Time (t)	20(4/6)	40(4/26)	60(5/16)	80(6/5)	100(6/25)
α	posterior mean	1.02×10^{-4}	4.61×10^{-6}	4.81×10^{-7}	7.60×10^{-8}	1.59×10^{-8}
	posterior median	7.47×10^{-7}	9.42×10^{-12}	1.16×10^{-16}	1.49×10^{-21}	1.87×10^{-26}
	posterior s.d.	4.46×10^{-4}	3.58×10^{-5}	6.01×10^{-6}	1.40×10^{-6}	4.03×10^{-7}
β	posterior mean	0.051	0.029	0.018	0.012	0.008
	posterior median	0.050	0.029	0.017	0.010	0.005
	posterior s.d.	0.016	0.005	0.005	0.007	0.008
γ	posterior mean	0.115	0.206	0.345	0.523	0.724
	posterior median	0.114	0.206	0.344	0.524	0.726
	posterior s.d.	0.019	0.014	0.022	0.063	0.141
η	posterior mean	0.451	0.142	0.052	0.022	0.012
	posterior median	0.438	0.140	0.049	0.019	0.008
	posterior s.d.	0.163	0.027	0.016	0.014	0.012

is the first fever date of the cases in the household. We have analyzed the data using several sets of priors and obtained similar results. We report here the results using the priors in Sect. 4. We run five chains with 800,000 updates and 400,000 burn-ins for each of them, find that the Gelman–Rubin statistics for all the nine coefficients of the logistic functions α, β, γ are less than 1.1, and use the latter half of the five chains as samples from the posterior. The results are in Tables 6 and 7. Table 6 gives the posterior means, medians and standard deviations of the nine coefficients. Table 7 reports the posterior means, medians and standard deviations of the values of $\alpha, \beta, \gamma, \eta = \beta/\gamma$ at several time points; the scale of time is day and the time points are measured from March 18; note that the period from March 18 to July 5 is 110 days. Our analysis is based on all the data available at the 100th day, not assuming the epidemic is over.

Similarly to the results in the simulation study, these tables show that the standard deviations of the posterior distributions of $\alpha_1, \alpha_3, \beta_1, \beta_3, \gamma_1, \gamma_3$ are smaller than those of their prior distributions respectively and that the posterior standard deviations of

the function values are much smaller than their corresponding prior standard deviations. These seem to suggest that the results regarding the values of these functions are reliable. We note that it is these function values that are relevant in assessing the effect of control measures. We also note that similar remarks can be made for the analyses using different sets of priors; in addition, we find that although the posterior means and medians of some of the coefficients of the logistic functions α, β and γ may vary somewhat with the priors, those of the values of these functions make very little change with different priors. Our analysis shows that the within-household infection rate β decreases, the removal rate γ increases and their ratio η is less than 1 and decreases steadily and significantly during this period.

6 Discussion

We have presented a two-level mixing counting process S-I-R epidemic model for households that have at least one infective; this model allows the transmission parameters to vary over time; the parameters of interest in the model are within-household infectivity rate and removal rate. Bayesian methods for estimating these parameters, based on removal times, have been successfully illustrated in the simulation study. The simulation study indicates that estimates of within-household infectivity rate and removal rate are excellent, although the global infection rate is overestimated, as expected. We apply this method to study the Taiwan SARS data, which shows that the within-household infection rate decreases, the removal rate increases and their ratio is less than one and decreases significantly. This model allows the estimation of these parameters during the epidemic. For a rapidly transmitted disease, it provides a method to nearly real-time tracking of control measures within households.

We assume the global infection rate $\alpha(t)$, the local infection rate $\beta(t)$ and the removal rate $\gamma(t)$ are all monotone functions. While these are popular functions to use and may be reasonable in many practical situations, there are situations demanding other features. One desirable extension would allow $\alpha(t)$, $\beta(t)$ and $\gamma(t)$ to be any bounded positive deterministic functions, or bounded positive functions satisfying certain shape restrictions like monotonicity, convexity, unimodality, etc. In this context, we may make use of Bernstein polynomials; see, for example, [14, 15] for the use of Bernstein polynomials in shape-restricted regressions.

Our analysis is based on data from households having at least one infective. While this approach is useful in the study of within-household infection and removal rate, it overestimates the global infection rate and does not provide an estimate of the basic reproduction number. It is of interest to conduct a study that includes also households without any infective. With these data, it seems desirable and possible to study the basic reproduction number or effective reproduction number using both theoretical methods and simulation methods. It seems that while the simulation study may be carried in the way described in this paper, analyzing real data set requires judicious decision on which of the households having no infective are to be included in the study.

Suppose that the durations of outbreaks in households are much shorter than that in the whole community, then our model assumptions (1), (2) and (3) amount to considering classical general epidemic models, having the transmission parameters being

constant. These put certain restrictions on the transmission functions. We would like to relax the model to allow infectivity functions a general parametric form, with or without latent period. In fact, the more realistic SEIR (susceptible-exposed-infected-removed) model developed by [21] in the study of SARS deserves further investigation in the line of this paper. References regarding SEIR model in this context include [5, 23] and references therein. It seems that marked point process, which extends counting process, might be useful in this context. Another useful extension would be to incorporate observable covariates, like age, sex, health status, immune status, etc., in the model.

As we mentioned in the simulation study, it seems desirable to obtain a more efficient algorithm. In view of the fact that our algorithm requires the updates satisfying the compatibility condition so as to compute the likelihood, it seems a good idea to explore approximate Bayesian computation for the sampling of the posterior distribution; see [25], among others. Since the success of approximate Bayesian computation depends on the speed of generating data from the model and appropriate choice of summary statistics that capture information about the parameters, careful study is needed to design a more efficient algorithm. We note that generating data using the method of Sect. 4.1 is not fast enough.

Some of the above problems are under investigation; others will be taken up in the future.

Acknowledgements We thank the referee for comments and suggestions on an earlier version of the paper. This work is partially supported by the Commission of the European Communities EU sixth Framework Programme for research for policy support (contract SP22-CT-2004-511066), by Taiwan National Science Council (NSC 96-2118-M-400-002-MY2), and by Taiwan Centers for Disease Control (DOH-DC-SA02).

References

1. Addy CL, Longini IM, Haber M (1991) A generalized stochastic model for the analysis of infectious disease final size data. Biometrics 47:961–974
2. Andersen PK, Borgan, Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York
3. Andersson H, Britton T (2000) Stochastic epidemic models and their statistical analysis. Springer, New York
4. Ball FG, Mollison D, Scalia-Tomba G (1997) Epidemics with two levels of mixing. Ann Appl Probab 7:46–89
5. Bauch CT, Lloyd-Smith JO, Coffee MP, Galvani AP (2005) Dynamically modeling SARS and other newly emerging respiratory illnesses. Epidemiology 16:791–801
6. Becker NG (1989) Analysis of infectious disease data. Chapman and Hall, London
7. Becker NG, Dietz K (1995) The effect of the household distribution on transmission and control of highly infectious diseases. Math Biosci 127:207–219
8. Becker NG, Hall R (1996) Immunization levels for preventing epidemics in a community of households made up of individuals of different types. Math Biosci 132:205–216
9. Becker NG, Hopper JL (1983) The infectiousness of a disease in a community of households. Biometrika 70:29–39
10. Becker NG, Starczak DN (1997) Optimal vaccination strategies for a community of households. Math Biosci 139:117–132
11. Brèmaud P (1981) Point processes and queues: martingale dynamics. Springer, New York
12. Cauchemez S, Boelle P-Y, Donelly C, Furguson N, Thomas G, Leung G, Hedley A, Anderson R, Valleron A-J (2006) Real-time estimation in early detection of SARS. Emerg Infect Dis 12:110–113

13. Cauchemez S, Boelle P-Y, Thomas G, Valleron A-J (2006) Estimation in real time: the efficacy of measures to control emerging communicable diseases. Am J Epidemiol 164:591–597
14. Chang IS, Chien LC, Hsiung CA, Wen CC, Wu YJ (2007) Shape-restricted regression with random Bernstein polynomials. In: Liu R, Strawderman W, Zhang CH (eds) Complex data sets and inverse problems. IMS lecture notes—monograph series, vol 54, pp 187–202
15. Chang IS, Hsiung CA, Wu YJ, Yang CC (2005) Bayesian survival analysis using Bernstein polynomials. Scand J Stat 32:447–466
16. Chen M, Schmeiser B (1998) Towards black-box sampling. J Comput Graph Stat 7:1–22
17. Demiris N, O'Neill PD (2005) Bayesian inference for epidemics with two levels of mixing. Scand J Stat 32:265–280
18. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman and Hall/CRC, London
19. Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732
20. Islam MN, O'Shaughnessy CD, Smith B (1996) A random graph model for the final size distribution of household infections. Stat Med 15:837–843
21. Lipsitch M, Cohen T, Cooper B et al (2003) Transmission dynamics and control of severe acute respiratory syndrome. Science 300:1966–1970
22. Lipsitch M, Bergstrom CT (2004) Invited commentary: real-time tracking of control measures for emerging infections. Am J Epidemiol 160:517–519
23. Lloyd-Smith JO, Galvani AP, Getz WM (2003) Curtailing transmission of severe acute respiratory syndrome within a community and its hospital. Proc R Soc Lond B 270:1979–1989
24. Longini IM, Koopman JS (1982) Household and community transmission parameters from final distributions of infections in households. Biometrics 38:115–126
25. Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci USA 100:15324–15328
26. O'Neill PD, Roberts GO (1999) Bayesian inference for partially observed stochastic epidemics. J R Stat Soc A 162:121–129
27. Robert CP, Casella G (2004) Monte Carlo statistical methods, 2nd edn. Springer, New York
28. Wallinga J, Teunis P (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol 160:509–516
29. Wang TH, Wei KC, Hsiung CA, Maloney SA, Eidex RB, Posey DL, Chou WH, Shih WY, Kuo HS (2007) Optimizing severe acute respiratory syndrome response strategies: lessons learned from quarantine. Am J Public Health 97:S98–S100
30. White LF, Pagano M (2008) A likelihood based method for real-time estimation of the serial interval and reproductive number of an epidemic. Stat Med 27:2999–3016