Mathematical Statistics
Stockholm University

Epidemic models; inference

Tom Britton

Research Report 2003:1

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.matematik.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2003:1**,
http://www.matematik.su.se/matstat

Epidemic models; inference

Tom Britton[*][†]

January 2003

1 Introduction

Statistical inference for epidemics is most often based on stochastic epidemic models (see **Epidemic Models, stochastic**). A special property of such models is that individuals are dependent in that the chance of getting infected depends on whether or not other individuals are infected. When making inference another complicating property is that most often the underlying epidemic process is only partially observed. It is very rare that information about who infected whom is available. The most common type of data actually consists of only knowing who was infected and who was not, i.e. having no information about the time evolution of the spread. This type of data is called final size data.

In the present overview we present inference procedures for what is known as the *general epidemic model* which assumes a homogeneous community, and a model for a structured community (see **Epidemic models, structured population**) in which the community is partitioned into households. Which inference procedure to use depends on the underlying model, but also on the type of available data. Below, both maximum-likelihood and martingale methods are used on the general epidemic model, depending on the type of data. Further, in a separate section Markoc chain Monte Carlo (MCMC) methods for more complex models, having other structured communities or partial observations, are discussed.

2 Outbreak in a homogeneous community

Below we present inference procedures for the general epidemic model. It assumes a community of homogeneous individuals that mixes uniformly. One way to relax the assumption of homogeneity is to allow for different types of individual, where different types may have different susceptibility, infectivity and/or mixing patterns. Inference procedures for such extended models can for example be found in [10], where inference for a multitype epidemic in a closed community is considered, or Farrington *et al.* [18], who consider estimation procedures for an endemic situation where types corresponds to age-cohorts.

The general epidemic is an SIR model (see **SIR epidemic models**) for a closed community. Let $S(t)$, $I(t)$ and $R(t)$ respectively denote the number of susceptible, infectious

[*] Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: tom.britton@matematik.su.se.

[†] To appear in *Encyclopedia of Biostatistics*

and removed (= recovered and immune), at time t, and let n denote the community size. One way to define the general epidemic is by specifying the intensities for the two counting processes $N(t) = n - S(t)$ (the number of individuals who have been infected) and $R(t)$: given the process at time t, the rate of new infections (the intensity for $N(t)$) is $\lambda_N(t) = \beta\bar{S}(t)I(t)$, where $\bar{S}(t) = S(t)/n$, and the rate of removals (the intensity for $R(t)$) is $\lambda_R(t) = \gamma I(t)$. See [2] for theory on counting processes. The parameter β is hence the rate at which an infectious individual has contact with other individuals, so $betaS(t-)$ is the rate at which he or she infects other individuals since only susceptivle individuals can get infected. The parameter γ is the recovery rate of infectious individuals, and $1/\gamma$ is the average length of the infectious period.

2.1 Complete data

First we sketch how to perform inference assuming the epidemic process is observed continuously – so called complete data. If $(S(u), I(u), R(u))$, or equivalently $(N(u), R(u))$ is observed continuously up to time t, then the log-likelihood is given by

$$\ell(\beta,\gamma) = \int_0^t \left[\log\left(\beta\bar{S}(u)I(u)\right)dN(u) - \beta\bar{S}(u)I(u)du\right] + \int_0^t \left[\log\left(\gamma R(u)\right)dR(u) - \gamma R(u)du\right]. \tag{1}$$

The first term of each integral above is actually a sum. The counting process $N(u)$ increase one unit at a time making $dN(u) = 1$ at these time instants and $dN(u) = 0$ otherwise. The first term of the first integral is hence the sum of $\log\left(\beta\bar{S}(u)I(u)\right)$ evaluated at these time instants, and similarly for the first term of the second integral.

From this the maximum likelihood estimates can be derived and shown to equal:

$$\hat{\beta}_{\text{ML}} = N(t)/\int_0^t \bar{S}(u)I(u)du, \tag{2}$$

$$\hat{\gamma}_{\text{ML}} = R(t)/\int_0^t I(u)du. \tag{3}$$

Standard errors can also be derived using large population results from the general epidemic (e.g. [8]). The most important parameter, the basic reproduction number R, for the general epidemic is given by $R = \beta/\gamma$ (see **Reproduction number**), so the maximum likelihood estimator of R, given complete data, is

$$\hat{R}_{\text{ML}} = \frac{\hat{\beta}_{\text{ML}}}{\hat{\gamma}_{\text{ML}}} = \frac{N(t)\int_0^t I(u)du}{R(t)\int_0^t \bar{S}(u)I(u)du}. \tag{4}$$

The critical vaccination coverage v^*, the community proportion necessary to vaccinate in order to obtain herd immunity assuming a 100% effective vaccine, is given by $v^* = 1 - 1/R$ (see **Epidemic models, thresholds**). Accordingly v^* is estimated by

$$\hat{v}^*_{\text{ML}} = 1 - 1/\hat{R}_{\text{ML}}. \tag{5}$$

Standard errors for $\hat{R}_{\text{ML}}$ and $\hat{v}^*_{\text{ML}}$ can be obtained using the delta-method.

2.2 Final size data

As mentioned in the introduction, the most common type of data is final size data in which only the final state of the outbreak is observed, i.e. how many were infected and how many were not. It is not possible to get an estimate of β and γ separately for this type of data, since both parameters are related to time, and final size data contains no information about the time evolution of the epidemic. In fact, the log-likelihood in (1) is not observable for final size data. Instead we use that $M_1 = N(t) - \int_0^t \beta \bar{S}(u)I(u)du$ and $M_2 = R(t) - \int_0^t \gamma I(u)du$ are martingales (see [2] for the underlying theory behind this). From M_1 and M_2 we can form a new martingale such that the unobservable quantities of M_1 and M_2 cancel out. It turns out that the rightmartingale is

$$\begin{aligned} M(t) &= \int_0^t \frac{1}{\bar{S}(u-)} dM_1(t) - \frac{\beta}{\gamma} M_2(t) = \int_0^t \frac{1}{\bar{S}(u-)} dN(t) - \frac{\beta}{\gamma} R(t) & (6) \\ &= \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{S(t)+1} - \frac{\beta}{\gamma} R(t) \approx n \log(n/S(t)) - \frac{\beta}{\gamma} R(t). & (7) \end{aligned}$$

The second equality relies on the assumption that initially one individual was infectious and the rest were susceptible, i.e. $(S(0),\ I(0),\ R(0)) = (n-1,\ 1,\ 0)$. At the end of the epidemic $(t = \tau)$ there are no infectious individuals present, so $R(\tau) = n - S(\tau)$ and $M(\tau) \approx -n \log(1-\tilde{p}) - n\frac{\beta}{\gamma}\tilde{p}$, where $\tilde{p} = R(\tau)/n$ is the observed final proportion infected. Since M is a zero mean martingale we can apply the method of moments to get an estimate of $R = \beta/\gamma$ from final size data:

$$\hat{R}_{\text{FSD}} = \left(\frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{n - R(\tau) + 1} \right) / R(\tau) \approx \frac{-\log(1-\tilde{p})}{\tilde{p}}. \qquad (8)$$

This is the same estimator as if estimation would be based on the deterministic limit of the general epidemic (see **Epidemic models, deterministic**) where the final proportion infected p is known to solve the equation $1 - p = \exp(-Rp)$. However, in the stochastic setting we can also obtain standard errors for the estimator using martingale theory (e.g. [24]):

$$s.e.(\hat{R}_{\text{FSD}}) = \left[\frac{1}{(n-1)^2} + \frac{1}{(n-2)^2} + \cdots + \frac{1}{(n-R(\tau)+1)^2} + \frac{\hat{R}^2_{\text{FSD}}}{n}\tilde{p} \right]^{1/2} \Big/ \tilde{p}. \qquad (9)$$

The critical vaccination coverage $v^* = 1 - 1/R$ is of course estimated by $\hat{v}^*_{\text{FSD}} = 1 - 1/\hat{R}_{\text{FSD}}$ from final size data. Standard errors can as before be obtained by applying the delta-method.

The maximum likelihood estimate of R, and hence also of v^* given final size data can in principle be derived using formulae for the final size distribution (e.g. Bailey [5]). However, these formulae quickly become cumbersome for large communities, making such inference computationally involved and numerically unstable.

3 Outbreak in a community of households

We now present inference procedures in a different setting where individuals reside in households and where it is believed that infection rates are much higher between individuals of the same households than between individuals of different households. We do this for a fairly simple model originating from Longini and Koopman [20] where households are treated as if they were independent. Since then these ideas have been refined in several ways, for example by allowing individuals of different types and/or treating a fully stochastic model where households are dependent (e.g. [1], [6], [21] and [11].

The key idea in the Longini-Koopman model [20] is to treat the probability of getting infected from outside the household during the course of the epidemic as a parameter. In reality this probability depends on the number of individuals who get infected and is hence a stochastic quantity, but the simplifying assumption reduces computational complexities tremendously. Further, by estimating the parameter it will be close to its "correct"value.

3.1 A simple household model

Individuals reside in households. An individual who gets infected has infectious contacts with other individuals in the household independently and with equal probability $p_W = 1 - q_W$. Additionally, each individual receives an infectious contact from outside the household with probability $p_B = 1 - q_B$ (the indices stand for within and between households). Individuals who receive at least one infectious contact from infected household members, or from outside the household, get infected. Only those who escape infectious contacts both from within and outside the household avoid getting infected during the epidemic outbreak. Let $p_h(j); j = 0, \dots, h$ denote the probability that j individuals get infected in a household having h (initially susceptible) individuals. Then these probabilities can be derived recursively from the following equations:

$$p_h(j) = \binom{h}{j} q_W^{j(h-j)} q_B^{h-j} - \sum_{r=0}^{j-1} \binom{h-r}{j-r} p_h(r) q_W^{(h-j)(j-r)} \qquad j = 0, \dots, h, \tag{10}$$

e.g. [1]. For example, $p_h(0) = q_B^h$ and $p_h(1) = \binom{h}{1}(1 - q_B) q_B^{h-1} q_W^{h-1}$ which can easily be explained. No one gets infected if everyone escapes infection from outside. One individual gets infected if 1 out of h gets infected from outside, and the remaining $h - 1$ individuals escape infection both from outside and from the infected household member. The probabilities quickly become complicated as the requested number of infected increases, but for households smaller than say 5 or even 10 they can be computed algebraically using a computer.

3.2 Inference for the simple household model

Inference is quite straightforward once the relevant $p_h(j)$'s have been calculated, since households were assumed independent. Let $\{n_h(j)\}$ denote the collected data, where $n_h(j)$ denotes the observed number of households of size h in which j individuals got infected during the epidemic. Then the log-likelihood for the data is simply

$$\ell(q_W, q_B) = \sum_{h,j} n_h(j) \log(p_h(j)), \tag{11}$$

where the dependence on the parameters is implicit from the definition of $\{p_h(j)\}$ in (10). The parameters are simply estimated by maximizing the log-likelihood with respect to q_W and q_B. Because households are assumed independent, standard large population theory is applicable when the number of households is large, and the maximum likelihood estimators are consistent. Standard errors for the estimates can be obtained from the observed information matrix by differentiating the log-likelihood twice (e.g. [13]).

As the model is defined, there is no basic reproduction number R, because households behave independently. In Ball *et al.* [6] a related fully stochastic model is considered, enabling estimation of the basic reproduction number R.

4 Inference using MCMC methods

In previous sections we have mainly treated models and data for which it was possible to derive expressions for outcome probabilities. In more realistic (i.e. complex) settings this may not be practically possible. Often the detail in the data does not allow for straightforward estimation of parameters. Then some missing-data method can sometimes be helpful. There are a few examples where the EM-algorithm can be helpful (see e.g. [4]), but here we focus on Markov chain Monte Carlo (MCMC) methods (e.g. [17]). This methodology has been successfully applied in a few situations but its real breakthrough in epidemic inference still lies ahead.

The main idea of MCMC analysis in epidemic inference is to explore the outcome space of unobserved (latent) variables for which inference procedures would have been much easier, had these variables been observed. Most often uninformative priors are used for model parameters, but in specific cases prior knowledge can of course be expressed into informative prior distributions. Below we list some inference problems where MCMC methods have been applied, and refer to listed references for details.

Inference is non-trivial even for the general epidemic model when the removal times, but not the infection times, are observed. This type of data is quite common since the removal time of an individual is approximately the same as detection time, which is quite often known. The reason for the complication is that the likelihood then has to be integrated over all possible infection times, a time-consuming task even for very small community sizes. In O'Neill and Roberts [23] this problem is analysed using MCMC methods in which the Markov chain explores the space of possible infection times. (A different approach, using martingales, is performed in [9].)

Also for household data, detection times but not infection times may sometimes be available. For a model allowing a fairly general distribution for the infectious period, perhaps preceded by a latency period, inference is complicated even for households of size two and when treated as independent. In O'Neill et al. [22] this type of data is analysed using MCMC methods, where the unobserved infection times and latency periods are explored in the Markov chain.

It is of course hard to include all heterogeneities into a model. For example, to determine all social connections between individuals in a community is impossible. A way out of this problem is to model unknown social structures by introducing unobserved random social contacts. In Britton and O'Neill [12] a first step in this direction was taken by modelling the social structure using a random graph, and assuming that transmission may only occur between neighbouring individuals of the graph. Inference is performed without assuming any information about the social graph, and the Markov chain explores the possible graphs, where detection times close in time increase the probability of a social link between the corresponding pair of individuals.

5 Concluding remarks

The emphasis of this article has been on inference procedures for epidemic models in general, rather than on models for specific diseases. The methods are suited for diseases in which transmission occurs by person-to-person contact, and not for vector-borne diseases like malaria or infectious diseases caused by contaminated water or food like salmonella. Examples of such diseases are childhood diseases like measles and mumps, smallpox, HIV (although heterogenous structures tend to be very complex here), influenza and common cold.

We have described inference procedures for a few stochastic epidemic models. In many applications the underlying setting is too complicated to enable inference from stochastic models, for example when long term endemic situations are considered and the community changes dynamically, or when there are too many types of heterogeneities. Then data can be calibrated to deterministic models thus giving parameter estimates. A thorough treatment of many such situations is given in Anderson and May [3] (see also **Epidemic models, deterministic**). Inference using stochastic models, as opposed to deterministic, has the advantage that it provides uncertainty estimates of parameters. Stochastic models are also better suited for situations where small social units, such as households, play an important role in the disease spread. In this case deterministic models, relying on large population results, may give misleading results. Deterministic models on the other hand, have the clear advantage of being simpler to analyse, thus permitting more complex models to be used.

The practical problem to estimate the effect of a vaccine against an infectious disease, the vaccine efficacy, is not treated in the present article. Clearly this is an important inferential problem within infectious disease epidemiology, but it is left out from the presentation as epidemic models play a minor role in such analyses. Estimation procedures for such problems can for example be found in [19] and [15] and the references therein.

For more detailed presentations on statistical inference for epidemic models we recommend the monographs by Becker [7] and Andersson and Britton [4], and the survey paper [8]. More on epidemic models in general can be found in [5], [3] [14] and [16].

Referenser

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

[2] Andersen, P. K., Borgan, Ø, Gill, R. D., Keiding N. (1993). *Statistical models based on counting processes.* New York: Springer.

[3] Anderson, R. M. and May, R. M. (1991). *Infectious diseases of humans; dynamic and control.* Oxford: Oxford University Press.

[4] Andersson, H. and Britton, T. (2000). *Stochastic models and their statistical analysis.* Springer Lecture Notes in Statistics **151**. Springer, New York.

[5] Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications.* London: Griffin.

[6] Ball, F., Mollison, D. and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Ann. Appl. Prob.* **7**, 46-89.

[7] Becker, N. G. (1989). *Analysis of Infectious Disease Data.* Chapman and Hall, London.

[8] Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *J. R. Statist. Soc. B* **61**, 287-307.

[9] Becker, N. G. and Hasofer, A. M. (1997). Estimation in epidemics with incomplete observations. *J. Roy. Statist. Soc. B,***59**, 415-429.

[10] Britton, T. (2001). Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage. *J. R. Statist. Soc. B,***63**, 705-715.

[11] Britton, T. and Becker, N. G. (2000). Estimating the immunity coverage to prevent epidemics in a community of households. *Biostatistics* **1**, 389-402.

[12] Britton, T. and O'Neill, P. D. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scand. J. Stat.,* **29**, 375-390.

[13] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman & Hall, London.

[14] Daley, D. J. and Gani, J. (1999). *Epidemic Modelling: an introduction.* Cambridge University Press, Cambdridge.

[15] Datta, S., Halloran M. E. and Longini I. M. (1999). Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: randomization by individual or household. *Biometrics,* **55**, 792-798.

[16] Diekmann, O. and Heesterbeek, J.A.P. (2000): *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation.* Wiley, Chichester.

[17] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (Eds) (1996). *Markov chain Monte Carlo in practice.* Chapman and Hall, London.

[18] Farrington, C. P., Kanaan, M. N. and Gay, N. J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Appl. Statist.* **50**, 251-292.

[19] Halloran, M. E., Haber, M. and Longini, I. M. (1992). Interpretation and estimation of vaccine efficacy under heterogeneity. *Amer. J. Epidemiol.* **136**, 328-343.

[20] Longini, I. M. and Koopman, J. S. (1982). Household and community transmission parameters from final distributions of infections in households. *Biometrics* **38**, 115-126.

[21] Lyne, O. D. and Ball, F. G. (1999). Parameter estimation for SIR epidemics in households. Bull. Int. Statist. Inst. 52nd Session, Contributed Papers, Vol. LVIII, Book 2, p 251.

[22] O'Neill, P., Balding, D., Becker, N. G., Eerola, M. and Mollison, D. (2000): Analyses of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods. *Applied Statistics,* **49**, 517-542.

[23] O'Neill, P. and Roberts, G. (1999): Bayesian inference for partially observed stochastic epidemics, *J. Roy. Statist. Soc. A.***162**, 121-129.

[24] Rida, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *J. R. Statist. Soc. B* **53**, 269-283.