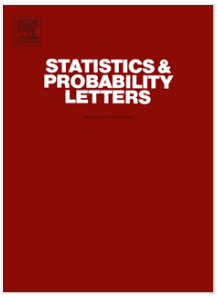
A note on maximum likelihood estimation of the initial number of susceptibles in the general stochastic epidemic model

Theodore Kypraios



A note on maximum likelihood estimation of the initial number of susceptibles in the general stochastic epidemic model

Theodore Kypraios

PII: S0167-7152(09)00215-6
DOI: 10.1016/j.spl.2009.06.003
Reference: STAPRO 5444

To appear in: *Statistics and Probability Letters*

Received date: 16 June 2008
Revised date: 2 June 2009
Accepted date: 9 June 2009

Please cite this article as: Kypraios, T., A note on maximum likelihood estimation of the initial number of susceptibles in the general stochastic epidemic model. *Statistics and Probability Letters* (2009), doi:10.1016/j.spl.2009.06.003

A Note on Maximum Likelihood Estimation of the Initial Number of Susceptibles in the General Stochastic Epidemic Model.

Theodore Kypraios

School of Mathematical Sciences, University Park, University of Nottingham, NG7 2RD, United Kingdom

Abstract

The initial number of susceptible individuals in a population is usually assumed to be known and statistical inference for some of the quantities of interest, such as the *basic reproductive number* R_0, is straightforward. However, in any epidemic, there may exist a number of individuals who may not be involved in the transmission of the disease. In this note we show how maximum likelihood estimators can be derived for the parameters of interest. The proposed methodology is then applied to the Abakaliki smallpox data in Nigeria.

1 Introduction

Understanding the spread of an infectious disease is a crucial issue in order to prevent major outbreaks of an epidemic. The analysis of outbreak data can be more effective when it is based on a model for the actual process which generates the data. Models could be used to provide a better understanding of the transmission dynamics, the infection process, and the epidemiologically quantities of interest. There exists a comprehensive literature on deterministic and stochastic epidemic modelling; (see for example, Bailey, 1975, Becker, 1989, Daley and Gani, 1999, and the references therein). Many researchers have focused on estimating key quantities of interest, such as the rate at which an infected individual makes contacts with the rest individuals in the population, and the basic reproduction number R_0 (Becker, 1989, Becker and Hasofer, 1997, Riley et al., 2003).

Email address: theodore.kypraios@nottingham.ac.uk (Theodore Kypraios).
URL: http://www.maths.nott.ac.uk/∼tk (Theodore Kypraios).

Estimating the size of a population is a common problem in many scientific fields, for example, in ecology (Huggins, 1989, Yip, 1989). Traditional inference for stochastic epidemic models depends on the knowledge of the initial number of susceptible individuals. Huggins et al. (2004) encountered the problem of estimating the initial number of individuals that are susceptible to a disease. Being able to estimate the number of initially uninfected individuals enable us to determine if there are individuals in the population with either a natural immunity to the disease or are for some reason not exposed to the disease (e.g. isolated).

Huggins et al. (2004) provided an estimator of the number of initially susceptible individuals as well as its approximate variance, by adopting a martingale framework (see, for example, Becker, 1989). Here we show that if the epidemic is fully observed, i.e. infection and removal times are available, a maximum likelihood estimator (MLE) can be derived in a straightforward manner. A simulation study is conducted to compare our MLE and its properties (i.e. its standard error) with the martingale estimator (ME) by Huggins et al. (2004). We then apply our methodology to the Abakaliki smallpox data in Nigeria that has been previously considered by Huggins et al. (2004) too.

2 Notation

We adopt a very similar notation as Huggins et al. (2004). A closed population (i.e. no births/ deaths/ immigration) of size $N + a$ is considered; we assume that at time $t = 0$ there are a initially infective individuals. Denote by $S(t)$, $I(t)$ and $R(t)$ the number of susceptible, infective and removed individuals respectively at present time t. The infectious periods of different individuals are independent and identically distributed according to some random variable D, which can have any arbitrary but specified distribution. In addition, we assume that the epidemic is observed up to a certain time, say τ. Denote by $n_I \leq N$ and $n_R \leq N$, the number of individuals who became infected and removed by time T respectively. In general, $n_R \leq n_I \leq N$. Note that when an individual becomes infected, we also assume that he/she becomes infective at the same time (i.e. able to spread the disease).

The epidemic process $(S(t), I(t))$ is Markov if and only if the infectious period has the lack-of-memory property. This is the special (Markovian) case where the infectious periods follow an Exponential distribution. Such a model is known as the *general stochastic epidemic* (GSE). Then, the process $(S(t), I(t))$ can be fully described in terms of continuous time Markov chains with the following transition rates:

$$(S(t)=i, I(t)=j) \rightarrow (S(t+\delta t)=i-1, I(t+\delta t)=j+1) : \frac{\beta}{N} S(t) I(t)$$
$$(S(t)=i, I(t)=j) \rightarrow ((S(t+\delta t)=i, I(t+\delta t)=j-1) : \gamma I(t)$$

while the transition probabilities turn out to be:

$$\begin{aligned}
\mathbb{P}[S(t+\delta t) - S(t) = -1,\ I(t+\delta t) - I(t) = 1 \mid \mathcal{H}_t] &= \frac{\beta}{N} S(t) I(t) \delta t + o(\delta t) \\
\mathbb{P}[S(t+\delta t) - S(t) = 0,\ I(t+\delta t) - I(t) = -1 \mid \mathcal{H}_t] &= \gamma I(t) \delta t + o(\delta t) \\
\mathbb{P}[S(t+\delta t) - S(t) = 0,\ I(t+\delta t) - I(t) = 0 \mid \mathcal{H}_t] &= 1 - \frac{\beta}{N} S(t) I(t) \delta t \\
&\quad - \gamma I(t) \delta t + o(\delta t)
\end{aligned}$$

where $\mathcal{H}_t$ is the sigma-algebra generated by the history of the process up to time t, i.e. $\mathcal{H}_t = \sigma\{(S(f), I(f)) : 0 \leq f \leq t\}$, with $\mathcal{H}_0 = \sigma\{S(0) = N, I(0) = \alpha)\}$ specifying the initial conditions. Therefore, the probability of an infection or a removal at the time interval $[t, t+\delta t)$ are $\beta S(t) I(t) + o(\delta t)$ and $\gamma I(t) + o(\delta t)$ respectively. The correction term $o(\delta t)$ becomes negligible for small δ, i.e. $\frac{o(\delta t)}{\delta t} \rightarrow 0$ as $\delta t \rightarrow 0$. Note that by assuming a closed population, we only need to keep track only (instead of the three) processes, since it holds $I(t) + S(t) + R(t) = N$ for any t.

The form of the transition probabilities show that the probability of infection at time t is proportional to the total number of infectives and susceptibles at time t. The constant of proportionality, β, is referred to as the *infection* rate. The transition probability of a removal shows that the length of the infectious periods are independent, identically distributed exponential random variables with mean $1/\gamma$, and therefore γ is referred as the *removal* rate for each individual. The epidemic continues until there are no more infective individuals left circulating in the population.

We assume that we observe the times at which individuals become infected are known i.e. $I(t), 0 \leq t \leq \tau$ is fully observed. However, we don't observe the initial number of susceptible individuals in the population, N, and therefore the process $S(t), 0 \leq t \leq T$ is unobserved.

3 Maximum Likelihood Estimation

We are interested in drawing inference for the unknown number of initially susceptible individuals in the population (N), given that we have the infection process $I(t)$ and also know that the epidemic has ceased. In this section we show that a maximum likelihood estimator $\widehat{N}$ can be obtained.

First, we consider the likelihood of the data given the parameters of interest, β, γ and N, using counting process theory (see for instance, Andersson and Britton, 2000). Letting $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_{n_I})$, to denote the (ordered) successive removal times observed during $[0, T]$. In other words, τ_i refers to the ith removal time. Denote by ϕ_1 the initial infection time and $\boldsymbol{\phi} = (\phi_2, \ldots, \phi_{n_I})$ the remain successive infection times during (ϕ_1, T); n_I denotes the number of total number of individuals (including the initial one) who contracted the disease.

$$L(\boldsymbol{\tau}, \boldsymbol{\phi}|\beta, \gamma, N) \propto \prod_{j=1}^{n_I} \gamma I(\tau_j^-) \times \prod_{j=2}^{n_I} \frac{\beta}{N} S(\phi_j^-) I(\phi_j^-) \times \exp\left\{ -\int_{\phi_1}^{T} \left(\frac{\beta}{N} S(t) I(t) + \gamma I(t) \right) dt \right\} \tag{1}$$

where the notation ϕ_j^- denotes the left hand limit, so for example $I(\phi_j^-)$ denotes the $\lim_{\uparrow \phi_j}(I(s))$, or in other words the time immediately prior to ϕ_j. Note that although $I(t)$ only depends on the infection and removal times, $S(t)$ depends on N too.

We are interested in maximising the log-likelihood of the observed data given the parameters. Therefore, taking the logarithm of (1) gives:

$$\begin{aligned} \log L(\boldsymbol{\tau}, \boldsymbol{\phi}|\beta, \gamma, N) = (n_I - 1)(\log \beta - \log N) + \log \left(\prod_{j=2}^{n_I} S(\phi_j^-) I(\phi_j^-) \right) \\ - \frac{\beta}{N} \int_{\phi_1}^{T} S(t) I(t) \ \mathrm{dt} - \gamma \int_{\phi_1}^{T} I(t) \ \mathrm{d}t + n_I \log \gamma \\ + \log \left(\prod_{j=1}^{n_I} I(\tau_j^-) \right) \end{aligned} \tag{2}$$

The removal rate γ is easily derived by differentiating (2) with respect to γ and set the derivative equal to zero. Then we are left with:

$$\frac{\partial_2 \log L(\boldsymbol{z}|N, \beta)}{(\partial \beta)(\partial N)} = 0.$$

Although there is not available a closed expression for $\widehat{\beta}$ and $\widehat{N}$, we can maximise (2) with respect to β and N numerically (see details in Section 4). The inverse of the matrix of the second-order (partial) derivatives (also called the

Hessian matrix) would then give us the variance-covariance matrix of the estimators which in turn will lead to the computation of their (approximate) standard errors.

4 Simulation Study

In this section we conduct a very similar simulation study to the one described in Huggins et al. (2004) to compare the efficiency of the proposed MLE to the ME as derived there.

We also consider populations of $N = 100$, $N = 250$, $N = 1000$ and $N = 5000$ susceptibles with $\alpha = 1$ initially infective. We chose two different values for the infection rates, $\beta = 1.3$ and $\beta = 1.5$ and one value for the removal rate, $\gamma = 1$. We follow Huggins et al. (2004) and we only considered simulated epidemics where more than 20% of the individual were infected; for $\beta = 1.3$ we also considered epidemics were more than 40% of the individuals were infected. For each combination of the parameters, 1000 epidemics were simulated; we then used the statistical language R and the function `optim` to maximise numerically the log-likelihood with respect to β and N.

The simulation results presented in Table 1 suggest that, overall, the unknown number of initially susceptible individuals in the population is estimated well. Although this is the case for the ME too (Huggins et al., 2004), the maximum likelihood approach offers higher precision. For each different scenario, i) we compute the standard deviation of the maximum likelihood estimates of N (sd($\widehat{N}$)), ii) the average standard error of each of the estimates (av(se($\widehat{N}$))), iii) the coverage and the average final size of the simulated epidemics.

The standard deviation of the $1,000$ estimates of N simulations is significantly lower than the corresponding error of the ME. Furthermore, the average standard errors of the MLE are much smaller especially for small values of N. Another advantage of the MLE as compared to the martingale approach is that there were no cases (out of the 1,000 simulations) where an estimate could not be derived. Nevertheless, it seems that for small values of N and for minor outbreaks the martingale estimator performs better as it offers better coverage. One possible reason why the maximum likelihood approach provides low coverage in such circumstances could be due to the fact that, for small values of N and minor outbreaks, the likelihood function is flat. This fact combined with the way the standard errors are calculated (see also Section 3) could reflect an underestimation of the standard errors. In addition, it seems that the MLE exhibits negative bias for small population sizes (so as the ME). However, it seems that when the disease is reasonably infective (for instance, $\beta = 1.3$ - major outbreak) then this bias is decreasing.

Table 1
Simulation Study

N	av($\widehat{N}$)	sd($\widehat{N}$)	av(se($\widehat{N}$))	Coverage	av(I_τ)
$\beta = 1.5$					
100	88.09	22.05	15.04	77.8	58.10
250	233.17	38.44	28.73	84.3	144.03
1000	979.65	72.76	63.48	91.4	580.17
5000	4961.20	124.68	144.17	95.6	2908.46
$\beta = 1.3$					
100	81.49	25.52	19.69	69.7	47.36
250	218.35	52.89	42.62	76.7	112.64
1000	934.89	130.63	112.62	86.2	427.32
5000	4877.42	252.93	285.57	93.7	2108.32
$\beta = 1.3$ (Major outbreak)					
100	96.22	18.16	19.90	90.2	57.64
250	247.86	36.12	41.97	94.7	131.76
1000	997.27	75.87	105.38	97.8	473.56
5000	4974.06	161.15	269.90	98.6	2207.41

5 Application

We illustrate the above methodology on the Abakaliki smallpox data in Nigeria. A total of 30 cases was observed and Becker (1989, pg. 111) provides us with the 29 time intervals between the detected cases. From these we can obtain the corresponding infection time of each individual having assumed a certain value for the infectious period. In this sequel we have assumed the infectious period to be fixed and the same for every individual equal to 14 days (Mack, 1972). It is then relatively straightforward to maximise the log-likelihood (2) with respect to the parameters of interest N and β. The function `optim` was used in R to obtain the parameter estimates and their the corresponding standard errors of them (see Table 2).

We obtained $\widehat{N}_{MLE} = 35.27$ which is relatively close to the value which is reported in Huggins et al. (2004), $\widehat{N} = 42.1$. Our estimator though has a

much smaller standard error (6.70) then the one obtained in Huggins et al. (2004) (37.15). It is reported in Bailey (1975) that there were 120 individuals who had close contacts regularly. Obviously, our estimate is much smaller than 120; a possible explanation could be the fact a significant proportion of these individuals were immune to the disease at the start of the epidemic.

Figures 1 illustrates the shape of the log-likelihood and Figure 2 the profile likelihood of β (right-hand plot) for $N = 120$ which is the value of N which has been assumed in other attempts of modelling this dataset (see for example, O'Neill and Roberts, 1999).

Table 2
Parameter estimates for the Abakaliki smallpox data

Parameter	**Estimate**	**Standard Error**	**95% CI**
N	35.27	6.70	(22.12, 48.41)
β	0.14	0.038	(0.065, 0.214)

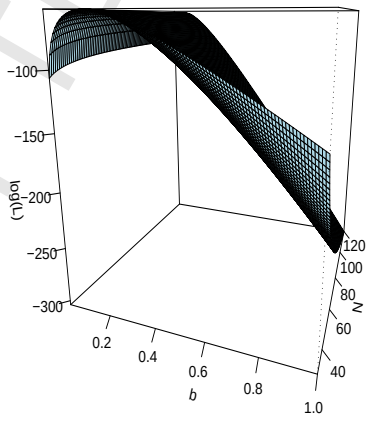


Fig. 1. Profile log-likelihood of β assuming $N = 120$ for the smallpox data

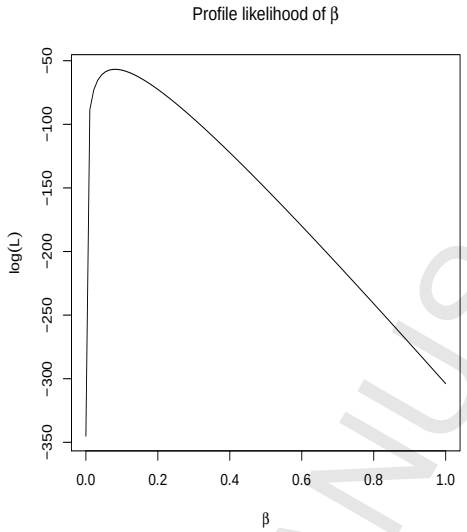


Fig. 2. Profile log-likelihood of β assuming $N = 120$ for the smallpox data

6 Discussion

We are often interested in estimating the number of initially susceptible individuals in a population. In this short note we showed that by assuming the infection and removal times of the individuals are known, maximum likelihood estimates can be derived for both N and the infection rate (β) in a straightforward manner. Overall, the estimator performs well and especially where the is large outbreak. There is simulation-based evidence that in a number of cases these estimators perform better than the ones obtained by using martingale methods as presented in Huggins et al. (2004).

Although the simulation and the application presented in this note refer to homogenously mixing model, maximum likelihood approaches can be easily used for more complex settings given that the infection times (as well as the removal times) are known. Unfortunately, it is very rare in practice that these will be available. Usually, only the removal times are know while it is often the case that someone may only have the final size of the epidemic. Nevertheless, in this case, it has already been illustrated that *data augmentation* techniques (Tanner and Wong, 1987) seem to be a natural framework for partially observed epidemics (O'Neill and Roberts, 1999, Gibson and Renshaw, 1998, Kypraios, 2007). A natural extension of the current maximum likelihood approach is to employ Markov Chain Monte Carlo (MCMC) algorithms draw

inference the parameters β, γ and N within a Bayesian framework.

7 Acknowledgments

The author would like to thank Dr Nikos Demiris for comments on earlier versions of this manuscript. In addition, the author is grateful to the editors and the reviewer for their valuable comments and suggestions which have led to improvements in the paper. Finally, funding from Wellcome Trust (Ref:076850) is gratefully acknowledged.

References

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, volume 151 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*. Hafner Press [Macmillan Publishing Co., Inc.] New York, second edition.

Becker, N. G. (1989). *Analysis of infectious disease data*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Becker, N. G. and Hasofer, A. M. (1997). Estimation in epidemics with incomplete observations. *J. Roy. Statist. Soc. Ser. B*, 59(2):415–429.

Daley, D. J. and Gani, J. (1999). *Epidemic modelling: an introduction*, volume 15 of *Cambridge Studies in Mathematical Biology*. Cambridge University Press, Cambridge.

Gibson, G. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov Chain methods. *IMA J. Math. Appl. Med. Biol.*, 15:19–40.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140.

Huggins, R. M., Yip, P. S. F., and Lau, E. H. Y. (2004). A note on the estimation of the initial number of susceptible individuals in the general epidemic model. *Statist. Probab. Lett.*, 67(4):321–330.

Kypraios, T. (2007). *Efficient Bayesian Inference for Partially Observed Stochastic Epidemics and A New class of Semi−Parametric Time Series Models*. PhD thesis, Department of Mathematics and Statistics, Lancaster University, Lancaster.

Mack, T. M. (1972). Smallpox in europe, 1950-1971. *J. Infect. Dis*, 125:161–169.

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

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L. M., Lam, T. H., Thach, T. Q., Chau, P., Chan, K. P., Lo, S. V., Leung, P. Y., Tsang, T., Ho, W., Lee, K. H., Lau, E. M., Ferguson, N. M., and Anderson, R. M. (2003). Transmission dynamics of the etiological agent of sars in hong kong: impact of public health interventions. *Science*, 300(5627):1961–1966.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.*, 82(398):528–550. With discussion and with a reply by the authors.

Yip, P. (1989). An inference procedure for a capture and recapture experiment with time-dependent capture probabilities. *Biometrics*, 45(2):471–479.