Johns Hopkins University, Dept. of Biostatistics Working Papers

6-21-2005

A Partial Likelihood for Spatio-temporal Point Processes

Peter J. Diggle
Department of Mathematics & Statistics, Lancaster University, United Kingdom and Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, U.S.A., p.diggle@lancaster.ac.uk

Suggested Citation
Diggle, Peter J., "A Partial Likelihood for Spatio-temporal Point Processes" (June 2005). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 75.
http://biostats.bepress.com/jhubiostat/paper75

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

A Partial Likelihood for Spatio-temporal Point Processes

Peter J Diggle
(Department of Mathematics and Statistics, Lancaster University
and
Department of Biostatistics, Johns Hopkins University School of Public Health)

June 21, 2005

Suumary. Spatio-temporal point process data arise in many fields of application. An intuitively natural way to specify a model for a spatio-temporal point process is through its conditional intensity at location x and time t, given the history of the process up to time t. Typically, this results in an analytically intractable likelihood. Likelihood-based inference therefore relies on Monte Carlo methods which are computationally intensive and require careful tuning to each application. We propose a partial likelihood alternative which is computationally straightforward and can be applied routinely. We apply the method to data from the 2001 foot-and-mouth epidemic in the UK, using a previously published model for the spatio-temporal spread of the disease.

Keywords: foot-and-mouth disease; infectious disease epidemiology; partial likelihood; spatio-temporal point process.

1 Introduction

In this paper, we consider spatio-temporal point process data of the form $(x_i, t_i) : i = 1, ..., n$, in which each (x_i, t_i) identifies the location and time of occurrence of an event of interest; in Section 3 we describe an application to the 2001 UK foot-and-mouth epidemic, where the x_i are farm-locations and the t_i the dates on which a case of foot-and-mouth was reported for the farm in question. We assume that the data consist of all relevant events in a pre-specified spatial region A and time-interval $[0, T]$, that a parametric model for the underlying point process has been specified and that our goal is to make inferences about the model parameters.

In Section 2 of the paper we develop an approach to inference which uses a form of partial likelihood (Cox, 1975). The approach benefits from the generally attractive

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

properties of likelihood-based methods whilst avoiding the computational complexity of full likelihood inference. In Section 3 we analyse data from the UK 2001 foot-and-mouth epidemic using a model proposed by Keeling et al (2001).

2 A partial likelihood

We assume that our point process is orderly, meaning roughly that coincident points cannot occur; for a rigorous discussion, see for example Daley and Vere-Jones (1988, Chapter 2). We denote by $\mathcal{H}_t$ the complete history of the process up to time t, and by $\lambda(x, t|\mathcal{H}_t)$ the conditional intensity for an event at location x and time t, given $\mathcal{H}_t$.

For data $(x_i, t_i) \in A \times [0, T] : i = 1, ..., n$, with $t_1 < t_2 < ... < t_n$, in principle the log-likelihood function can be expressed as

$$L(\theta) = \sum_{i=1}^{n} \log \lambda(x_i, t_i|\mathcal{H}_{t_i}) - \int_0^T \int_A \lambda(x, t|\mathcal{H}_t) dx dt. \tag{1}$$

See, for example, Daley and Vere-Jones (1988, Chapter 13). Two major obstacles to the routine use of (1) are that the form of the conditional intensity may itself be intractable, and that even when the conditional intensity is available direct evaluation of the integral term in (1) may be impractical. Monte Carlo methods are becoming more widely available for problems of this kind (Geyer, 1999; Moller and Waagepetersen, 2004). However, in practice these methods often need careful tuning to each application and the associated cost of developing and running reliable code can be an obstacle to their routine use.

As an alternative, computationally simpler approach to inference for models which are defined through their conditional intensity we propose a partial likelihood, which we obtain by conditioning on the locations x_i and times t_i and considering the resulting log-likelihood for the observed time-ordering of the events $1, ..., n$. To allow for right-censored event-times, we denote by $\mathcal{R}_i$ the risk-set at time t_i; in the absence of censoring, $\mathcal{R}_i = \{i, i+1, ..., n\}$. Now let

$$p_i = \lambda(x_i, t_i|\mathcal{H}_{t_i}) / \sum_{j \in \mathcal{R}_i} \lambda(x_j, t_i|\mathcal{H}_{t_i}). \tag{2}$$

Then, the partial log-likelihood is

$$L_p(\theta) = \sum_{i=1}^{n} \log p_i. \tag{3}$$

As discussed in Cox (1975), estimates obtained by maximising the partial likelihood inherit the general asymptotic properties of maximum likelihood estimators, although they may entail some loss of efficiency by comparison with full maximum likelihood estimates. Also, some parameters of the original model may be unidentifiable from

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper75

the partial likelihood. The loss of identifiability can be advantageous if the non-identified parameters are nuisance parameters as often applies, for example, to the baseline hazard function in the classic proportional hazards model for survival data (Cox, 1972). Otherwise, and again as exemplified by the proportional hazards model for survival data, other methods of estimation are needed to recover the unidentified parameters; see, for example, Andersen, Borgan, Gill and Keiding (1992).

3 Application: the 2001 foot-and-mouth epidemic

Foot-and-mouth disease (FMD) is a highly infectious viral disease of farm livestock. The virus can be spread directly between animals over short distances in contaminated airborne droplets, and indirectly over longer distances, for example via the movement of contaminated material. The UK experienced a major FMD epidemic in 2001, which resulted in the slaughter of more than 6 million animals. Its estimated cost to the UK economy was around £8 billion (UK National Audit Office, 2002).

3.1 Data

We shall analyse data from the two counties most severely affected by the epidemic, Cumbria in the north-west of England and Devon in the south-west. Because the two counties are geographically separated, we shall treat them informally as two replicates of a natural experiment, so allowing us to compare parameter estimates and pool as appropriate. A map of the at-risk farms at the start of the epidemic in both counties, distinguishing between farms which did and did not subsequently experience the disease, shows the typical pattern of an infectious disease process, with strong spatial aggregation of cases resulting from sequences of short-range transmission between neighboring farms, an effect which is all the more obvious when the data are plotted dynamically as a space-time map.

Information available on each farm includes: the numbers of cattle and/or sheep held; the date, if any, on which the disease was reported; and the date, if any, on which the farm's animals were slaughtered.

3.2 Model

The basic form of the model follows Keeling et al (2001); we discuss possible extensions in Section 3.5. Let $\lambda_{ij}(t)$ denote the conditional rate of transmission from farm i to farm j, given the history $\mathcal{H}_t$. Let n_{1i} and n_{2i} denote the numbers of cows and sheep held on farm i. Let $I_{ij}(t)$ denote an at-risk indicator for transmission of infection from farm i to farm j at time t; we assume that $I_{ij}(t) = 1$ if farm j is not infected and not slaughtered by time t, and farm i is infected and not slaughtered by time t. In the basic form of the model we assume that the reporting date is the infection date plus a

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

constant time τ, corresponding to the latent period of the disease plus any reporting delay. We ignore the complication that farm animals may be slaughtered after they become infected but before the disease is diagnosed.

The central feature of the model is a transmission kernel,

$$f(u) = \nu \exp\{-(u/\phi)^{\kappa}\} + \rho, \tag{4}$$

in which the powered exponential term corresponds to direct transmission of the infection over short distances, whilst the parameter ρ allows for occasional, apparently spontaneous cases occurring far from all currently infectious farms. For identifiability we set $\nu = 1$, so that ρ in (4) measures the relative importance of long-range transmission in the spread of the disease.

With the above definitions in place, the model specifies that

$$\lambda_{ij}(t) = \lambda_0(t) A_i B_j f(||x_i - x_j||) I_{ij}(t) \tag{5}$$

where $\lambda_0(t)$ is an arbitrary baseline hazard,

$$A_i = (\alpha n_{1i} + n_{2i}) \tag{6}$$

and

$$B_j = (\beta n_{1i} + n_{2i}). \tag{7}$$

The parameters α and β represent the relative infectiousness and susceptibility, respectively, of cows to sheep.

3.3 Fitting the model

For any farm k, we define $\lambda_k(t) = \sum_j \lambda_{jk}(t)$, from which we obtain the conditional intensities

$$\lambda(x_i, t_i | \mathcal{H}_{t_i}) = \lambda_i(t_i) / \sum_k \lambda_k(t_i)$$

and the partial log-likelihood follows by substitution of the conditional intensities into (2) and (3). To maximise the partial log-likelihood we use the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) as implemented in the `R` functon `optim()`, which provides a numerical estimate of the Hessian matrix.

3.4 Results

In the model for the transmission kernel, the parameters κ and ρ are poorly identified because the cases which appear to correspond to long-range transmission are few in number, and can be explained empirically either by including a small, positive value of ρ or by adjusting the value of κ. Because ρ corresponds formally to what is known to be a real effect, namely the indirect spread of infection via the movement of farm

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper75

equipment and staff, we retain ρ as a positive-valued parameter to be estimated, but fix $\kappa = 0.5$ to correspond to the observation in Keeling et al (2001) that the transmission kernel is more sharply peaked than exponential.

We first investigated whether the data in Cumbria and Devon support the assumption of a common set of parameters in the two counties. The likelihood ratio test statistic for common versus separate parameters is 2.98 on 4 degrees of freedom, hence $p = 0.56$ and we therefore accepted the hypothesis of common parameter values. We then obtained common parameter estimates $(\hat{\alpha}, \hat{\beta}, \hat{\phi}, \hat{\rho}) = (4.92, 30.68, 0.39, 9.9 \times 10^{-5})$. For all practical purposes, $\hat{\rho} \approx 0$, although a likelihood ratio test formally rejects $\rho = 0$ because the likelihood is sensitive to the precise probabilties which the model assigns to rare events.

One question of specific interest is whether the infectivities and susceptibilities for individual farms, A_i and B_j, are linear or sub-linear in the numbers of animals. To investigate this, we extend (6) and (7) to $A_i = (\alpha n_{1i}^{\gamma} + n_{2i}^{\gamma})$ and $B_j = (\beta n_{1i} + n_{2i})$, respectively, where γ is an additional parameter to be estimated. Fitting this five-parameter model results in a large increase in the maximised log-likelihood, from -6196.3 to -5861.4.

Another possible extension of the model would be to include farm-level covariates by defining $A_i = (\alpha n_{1i}^{\gamma} + n_{2i}^{\gamma}) \exp(z_i' \delta)$, where z_i is a vector of covariates for farm i, with a similar expression for the susceptibilties B_j. The z_i might, for example, codify management practices or other measured characteristics of individual farms which could affect their propensity to transmit, or succumb to, the disease. By way of illustration, we consider adding a log-linear effect of farm area to the model. The likelihood ratio statistic for the covariate effect is 3.26 on 1 degree of freedom, corresponding to $p = 0.07$. However, we can expect this test to be rather weak, because the observed distribution of farm area is extremely skewed, and the few farms with large areas will therefore have high leverage.

Estimates for the five-parameter model are shown in Table 1, together with approximate 95% confidence limits deduced from the numerical estimate of the Hessian matrix. Optimisation was conducted on the log-scale for all parameters, which is why the confidence limits are not symmetric about the point estimates. Estimated correlations amongst the parameter estimates were all small, the largest being 0.25 between $\log \phi$ and $\log \rho$. The results in Table 1 point strongly to a sub-linear dependence of infectivity and susceptibility on the numbers of animals. Note, however, that under the weak form of dependence implied by the estimate of γ, the parameter β is estimated very imprecisely.

These results are qualitatively similar to those reported in Keeling et al (2001), although they only considered the case $\gamma = 1$. They reported point estimates $\tilde{\alpha} = 1.61$ and $\tilde{\beta} = 15.2$. They did not specify a parametric model for the transmission kernel but their Figure 1B shows similar behaviour to our fitted model, with a sharper-than-exponential mode at zero and a long upper tail. Their Figure 1B shows a decay from 1 at $u = 0$ to a value of approximately 0.1 at $u = 1$km, compared with our

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Parameter	Estimate	95% confidence interval	
α	1.42	1.13	1.78
β	36.17	0.19	692.92
ϕ	0.41	0.36	0.48
ρ	1.3×10^{-4}	8.5×10^{-5}	2.1×10^{-4}
γ	0.13	0.09	0.21

Table 1: Parameter estimation for the five-parameter model fitted to combined data from Cumbria and Devon

$\hat{f}(1) = 0.21$.

Finally, we use a simple adaptation of the Nelson-Aalen estimator (Andersen, Borgan, Gill and Keiding, 1992, Chapter 4) to obtain a non-parametric estimate of the cumulative base-line hazard,

$$\hat{\Lambda}_0(t) = \int_0^t \hat{\lambda}_0(u)du.$$

We re-write (5) as $\lambda_{ij}(t) = \lambda_0(t)\rho_{ij}(t)$ and define $\rho(t) = \sum_i \sum_j \rho_{ij}(t)$. The Nelson-Aalen estimator is now given by

$$\begin{aligned} \hat{\Lambda}_0(t) &= \int_0^t \hat{\rho}(u)^{-1} dN(u) \\ &= \sum_{i:t_i \leq t} \hat{\rho}(t_i)^{-1}, \end{aligned} \tag{8}$$

where $\hat{\rho}(t)$ is the parametric estimate of $\rho(t)$ implied by the fitted model. Figure 1 shows the Nelson-Aalen estimates obtained from the Cumbria and Devon data. The generally lower estimates for Devon are consistent with the lower overall prevalence of the disease (137 cases out of 8182 at-risk farms in Devon, 657 cases out of 5090 at-risk farms in Cumbria). Both estimates are approximately linear over the first two to three months, by which time the epidemic in Devon has almost run its course. The slope of the Cumbria estimate increases thereafter. This does not necessarily imply a failure of the culling strategies being applied, since the model already takes account of their effects, but rather suggests that external environmental effects, for example the increase in animal movements outdoors in spring and summer, may have promoted an increase in the virulence of the disease process.

4 Discussion

In this short paper, our analysis of the FMD data has shown the feasibility of using the partial likelihood approach to answer a variety of questions relevant to an understanding of the disease process. A fuller analysis of these data will be reported separately.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper75

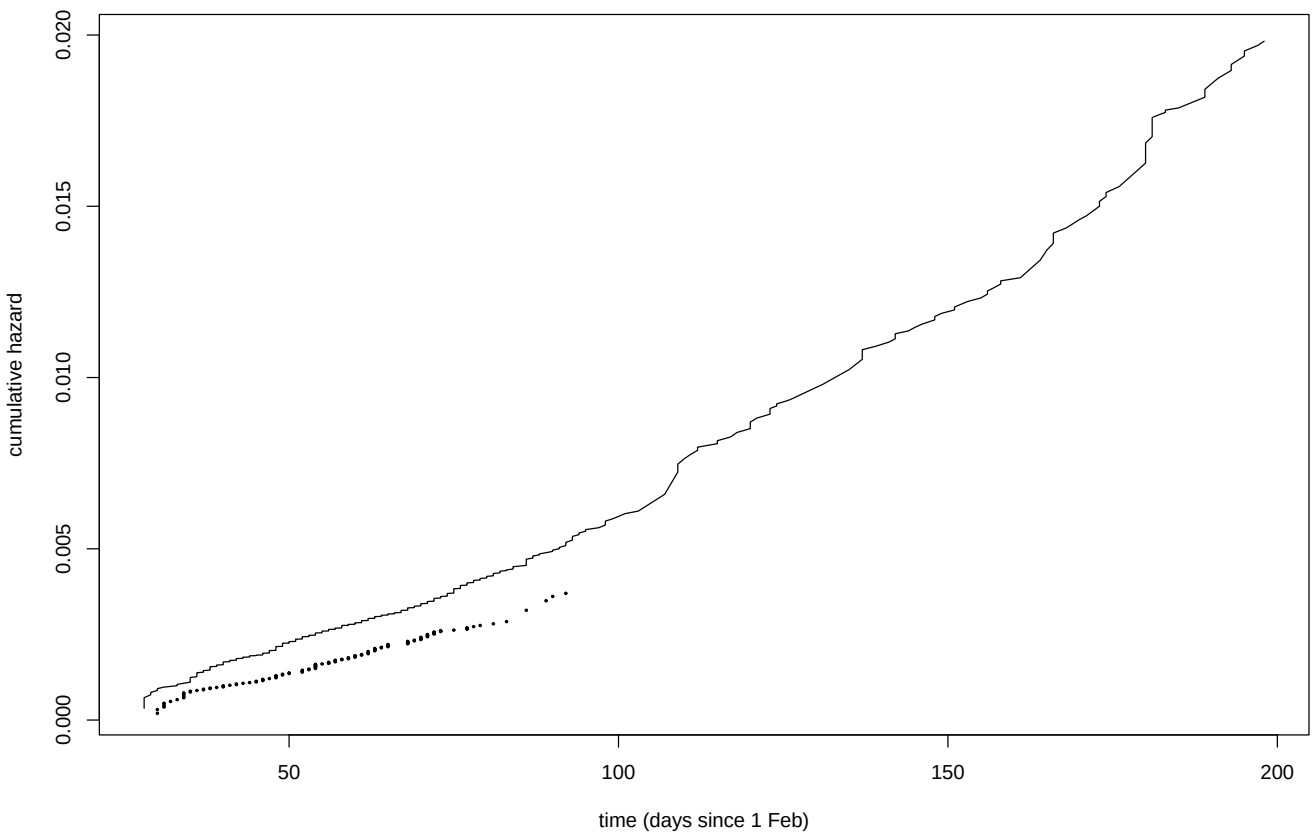

Figure 1: Estimated cumulative baseline hazards in Cumbria (solid line) and in Devon (dotted line) for the five-parameter model.

We suggest that the proposed partial likelihood provides a useful method for analysing spatio-temporal point process models specified via the conditional intensity function. The method is based on a generally accepted principle of inference with known asymptotic properties, whilst being computationally straightforward and therefore well-suited to routine use.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council through the award of a Senior Fellowship to Peter Diggle (Grant number GR/S48059/01), and by the UK Department for Environment, Food and Rural Affairs, grant number SE2930.

I thank Laura Green, Matt Keeling, Graham Medley, Sir David Cox and Hugues Lassalle for very helpful discussions.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

References

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1992). *Statistical Models Based on Counting Processes*. New York : Springer.

Cox, D.R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society* B, **34**, 187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269–75.

Daley D J and Vere-Jones D (1988). *An Introduction to the Theory pf Point Processes.* New York : Springer.

Geyer, C.(1999). Likelihood inference for spatial point processes. In *Stochastic Geometry: likelihood and computation*, ed O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout, 79–140.

Keeling, M.J., Woolhouse, M.E.J., Shaw, D.J., Matthews, L., Chase-Topping, M., Haydon, D.T., Cornell, S.J., Kappey, J., Wilesmith, J. and Grenfell, B.T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, **294**, 813–7.

Moller, J. and Waagepetersen, R.P. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* London : Chapman and Hall.

Nelder, J.A. and Mead, R. (1965). A simplex algorithm for function minimisation. *Computer Journal*, **7**, 308–13.

UK National Audit Office (2002). *The 2001 Outbreak of Foot and Mouth Disease.* Report by the Comptroller and Auditor General, HC 939, Session 2001–2002. London: The Stationery Office.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper75