Statistics and Probability Letters 118 (2016) 190–196

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Likelihood based inference for partially observed renewal processes

M.N.M. van Lieshout *

CWI, P.O. Box 94079, NL-1090 GB Amsterdam, The Netherlands
Department of Applied Mathematics, University of Twente, P.O. Box 217, NL-7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 26 January 2016
Received in revised form 28 June 2016
Accepted 4 July 2016
Available online 8 July 2016

Keywords:
Markov chain Monte Carlo
Renewal process
Sequential point process
State estimation

ABSTRACT

This paper is concerned with inference for renewal processes on the real line that are observed in a broken interval. For such processes, the classic history-based approach cannot be used. Instead, we adapt tools from sequential spatial point process theory to propose a Monte Carlo maximum likelihood estimator that takes into account the missing data. Its efficacy is assessed by means of a simulation study and the missing data reconstruction is illustrated on real data.

1. Introduction

Inference for point processes on the real line has been dominated by a dynamic approach based on the stochastic intensity (Brémaud, 1972; Karr, 1991; Last and Brandt, 1995) which relates the likelihood of a point at any given time to the history of the process. Such an approach is quite natural in that it is the mathematical translation of the intuitive idea that information becomes available as time passes. Furthermore, the approach allows the utilization of powerful tools from martingale theory, and, since the stochastic intensity is closely related to the hazard functions of the inter-arrival times distributions conditional on the past, a likelihood is immediately available (Daley and Vere-Jones, 2003).

Censoring in the sense of truncation at a random time independently of the point process can be dealt with (Andersen et al., 1993). However, as shown in Section 3, the dynamic approach does not seem capable of dealing with situations in which the flow of time is interrupted. In such cases, combined state estimation techniques are needed that are able to simultaneously carry out inference and reconstruct the missing points.

The aim of this paper is to apply ideas from sequential point process theory (Lieshout, 2006a,b), in particular the Papangelou conditional intensity which describes the probability of finding a point at a particular time conditional on the remainder of the process. Thus, the concept is related to the stochastic intensity, except that the future is taken into account as well as the past. It is this last feature that allows the incorporation of missing data. Moreover, for hereditary point processes at least, the sequential Papangelou conditional intensity defines a likelihood.

We focus on renewal processes on the positive half line and begin by recalling their definition in Section 2. Section 3 shows that a substantial bias may be incurred by the classic approach to parameter estimation when the process is only partially observed. An alternative method based on Markov chain Monte Carlo maximum likelihood (Geyer, 1999) and the sequential Papangelou conditional intensity (Lieshout, 2006b) is proposed in Section 4; Section 5 presents a simulation study

* Correspondence to: Department of Applied Mathematics, University of Twente, P.O. Box 217, NL-7500 AE Enschede, The Netherlands.
E-mail address: M.N.M.van.Lieshout@cwi.nl.

http://dx.doi.org/10.1016/j.spl.2016.07.002
0167-7152/

to quantify the bias reduction. It should be noted that the method applies equally to other partially observed point processes. Section 6 illustrates this point by analysing data about calls to a medical helpline.

2. Renewal processes

A renewal process on $[0, T]$, $T > 0$, is defined as follows (Karr, 1991, Chapter 8). Starting at time 0, let U_i be a sequence of (non-defective) independent and identically distributed inter-arrival times with probability density function π, cumulative distribution function F and hazard function h. Set $S_0 = 0$ and $S_i = S_{i-1} + U_i$ for $i \in \mathbb{N}$. Then those S_i, $i \geq 1$, that fall in $(0, T]$ form a simple sequential point process Y (Lieshout, 2006b). For simplicity, we assume that the process starts at time 0, but other initial distributions such as the forward recurrent time may also be accommodated.

Due to the independence assumptions in the model, the stochastic intensity $h^*(\cdot)$ of Y is particularly appealing. Write V_t for the backward recurrence time at t, that is, the difference between t and the last event falling before or at time t. Then, by Karr (1991, Prop. 8.10), $h^*(t) = h(V_{t-})$ and Y admits a density f with respect to $\nu_{[0,T]}$, the distribution of a unit rate Poisson process on $[0, T]$, that can be written as

$$f(t_1, \ldots, t_n) = e^T \exp\left[-\int_0^T h^*(t)\,dt\right] \prod_{i=1}^n h^*(t_i) = e^T (1 - F(T - t_n)) \prod_{i=1}^n \pi(t_i - t_{i-1}) \tag{1}$$

for $(t_1, \ldots, t_n) \in H_n([0, T]) = \{(t_1, \ldots, t_n) \in [0, T]^n : t_1 < \cdots < t_n\}$, cf. Karr (1991, Thm 8.17), under the conventions that an empty product is set to one and that $t_0 = 0$.

By definition, the stochastic intensity h^* is a function of the past of the process. A more versatile concept of conditional intensity is the sequential Papangelou conditional intensity (Lieshout, 2006b) for inserting t at position $k \in \{1, \ldots, n+1\}$ into the vector $(t_1, \ldots, t_n)$. It is defined by

$$\lambda_k(t|t_1, \ldots, t_n) = \frac{f(t_1, \ldots, t_{k-1}, t, t_k, \ldots, t_n)}{f(t_1, \ldots, t_n)}$$

whenever both $t_{k-1} < t < t_k$ (with $t_0 = 0$ and $t_{n+1} = \infty$ by convention) and $f(t_1, \ldots, t_n) > 0$; it is set to zero otherwise. Note that λ_k depends on both the past and the future. Conversely, provided $f(\cdot)$ is hereditary in the sense that $f(t_1, \ldots, t_n) > 0$ implies that $f(s_1, \ldots, s_m) > 0$ for all sub-sequences $(s_1, \ldots, s_m)$ of $(t_1, \ldots, t_n)$, the factorization

$$f(t_1, \ldots, t_n) = f(\emptyset) \prod_{i=1}^n \lambda_i(t_i|t_1, \ldots, t_{i-1})$$

holds.

Specializing to renewal processes, note that (1) is not necessarily hereditary, for example when π has small bounded support. Therefore we will assume that $\pi > 0$ and set $t_0 = 0$. Then (1) is hereditary and the sequential Papangelou conditional intensity $\lambda_k(t|t_1, \ldots, t_n)$ for $t_1 < \cdots < t_n$ reads

$$\lambda_k(t|t_1, \ldots, t_n) = \begin{cases} \pi(t - t_{k-1}) \dfrac{\pi(t_k - t)}{\pi(t_k - t_{k-1})} & \text{if } 1 \leq k \leq n \text{ and } t_{k-1} < t < t_k \\ \pi(t - t_{k-1}) \dfrac{1 - F(T - t)}{1 - F(T - t_{k-1})} & \text{if } k = n+1 \text{ and } t_n < t \leq T \end{cases} \tag{2}$$

and zero otherwise. Clearly, (2) depends on the vector $(t_1, \ldots, t_n)$ only through the two immediate neighbours t_{k-1} and t_k of t, provided such neighbours exist.

Next, consider the conditional distribution of Y on (T_1, T_2), $0 < T_1 < T_2 < T$, given $Y \cap [0, T_1] = \vec{\mathbf{r}}$ and $Y \cap [T_2, T] = \vec{\mathbf{s}}$. Then, for $t \in (T_1, T_2)$ and $(t_1, \ldots, t_n) \in H_n((T_1, T_2))$,

$$\lambda_k(t|t_1, \ldots, t_n; \vec{\mathbf{r}}, \vec{\mathbf{s}}) = \begin{cases} \pi(t - t_{k-1}) \dfrac{\pi(t_k - t)}{\pi(t_k - t_{k-1})} & \text{if } 1 \leq k \leq n \text{ and } \max(T_1, t_{k-1}) < t < t_k \\ \pi(t - t_{k-1}) \dfrac{\pi(s_1 - t)}{\pi(s_1 - t_{k-1})} & \text{if } k = n+1,\ \vec{\mathbf{s}} \neq \emptyset \text{ and } t_n < t < T_2 \\ \pi(t - t_{k-1}) \dfrac{1 - F(T - t)}{1 - F(T - t_{k-1})} & \text{if } k = n+1,\ \vec{\mathbf{s}} = \emptyset \text{ and } t_n < t < T_2 \end{cases} \tag{3}$$

and zero otherwise. Now, we use the convention that $t_0 = \max(\vec{\mathbf{r}})$ if $\vec{\mathbf{r}} \neq \emptyset$ and $t_0 = 0$ otherwise. As (2), (3) depends on the vector $(t_1, \ldots, t_n)$ only through the two immediate neighbours t_{k-1} and t_k of t, provided such neighbours exist. Moreover, assuming that $\vec{\mathbf{s}} \neq \emptyset$, the density of the conditional distribution of Y on (T_1, T_2) with respect to a unit rate Poisson process on (T_1, T_2) is, for $n \in \mathbb{N}$ and $(t_1, \ldots, t_n) \in H_n((T_1, T_2))$,

$$f(t_1, \ldots, t_n|\vec{\mathbf{r}}, \vec{\mathbf{s}}) = f(\emptyset|\vec{\mathbf{r}}, \vec{\mathbf{s}}) \frac{\pi(t_1 - \max(\vec{\mathbf{r}}))\pi(s_1 - t_n)}{\pi(s_1 - \max(\vec{\mathbf{r}}))} \prod_{i=2}^n \pi(t_i - t_{i-1}) \tag{4}$$