Journal of Statistical Planning and Inference 142 (2012) 2710–2718

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Nonparametric estimation of the number of components of a superposition of renewal processes

Anup Dewanji [a,*], Subrata Kundu [b], Tapan K. Nayak [b]

[a] Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India
[b] Department of Statistics, George Washington University, Washington, DC 20052, USA

ARTICLE INFO

Article history:
Received 15 July 2011
Received in revised form
2 March 2012
Accepted 3 March 2012
Available online 17 March 2012

Keywords:
Asymptotic normality
Bias
Kaplan–Meier estimator
Profile likelihood
Software reliability

ABSTRACT

Suppose all events occurring in an unknown number (ν) of iid renewal processes, with a common renewal distribution F, are observed for a fixed time τ, where both ν and F are unknown. The individual processes are not known a priori, but for each event, the process that generated it is identified. For example, in software reliability application, the errors (or bugs) in a piece of software are not known a priori, but whenever the software fails, the error causing the failure is identified. We present a nonparametric method for estimating ν and investigate its properties. Our results show that the proposed estimator performs well in terms of bias and asymptotic normality, while the MLE of ν derived assuming that the common renewal distribution is exponential may be seriously biased if that assumption does not hold.

1. Introduction

Suppose signals are received, in continuous time, from an unknown number (ν) of sources, where for each signal, its source is also observed. Thus, any specific source is detected when its first signal is received. Obviously, any source that does not send a signal during the observation period remains undetected. While each source induces a point process, the observed event (signal receiving) times come from the superposition of all the processes. The situation considered here may arise in continuous time capture–recapture experiments, where each animal or species acts as one source. In software testing, each error (or bug) is a source for causing software failures and the overall failure process is a superposition of an unknown number of point processes.

We shall consider the case where the overall process is observed for a fixed time τ. Formally, we observe the values of the following random variables: R is the number of detected processes (or sources), M_i the number of events in the ith detected process, $i=1,\dots,R$, and T_{ij} the jth inter-event (renewal) time in the ith detected process, $i=1,\dots,R, j=1,\dots,M_i$. Then, $S_i=\tau-\sum_{j=1}^{M_i}T_{ij}$ is the last censored event time for the ith detected process. We shall let $M=\sum_{i=1}^{R}M_i$ denote the total number of events observed. Following the standard convention we shall use upper case letters to denote random variables and lower case letters to denote their observed values.

It is most convenient to assume that all component processes are iid Poisson processes with a common but unknown rate λ (e.g., Becker, 1984; Nayak, 1988, 1991; Chao and Lee, 1993). Under the Poisson assumption, the maximum likelihood estimators (MLE) of ν and λ as well as their distributions have been obtained in Nayak (1988, 1991), which are reviewed briefly in Section 2. More generally, Dewanji et al. (1995) assumed that the component processes are iid renewal processes

* Corresponding author. Tel.: +91 33 2575 2808; fax: +91 33 2577 3104.
E-mail address: dewanjia@isical.ac.in (A. Dewanji).

0378-3758/$ - see front matter © 2012 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.jspi.2012.03.008

with a common renewal density $f_\theta(\cdot)$ and cdf $F_\theta(\cdot)$, where θ is an unknown parameter vector. They discussed maximum likelihood estimation of ν and θ and derived asymptotic normality of a class of estimators, which includes the MLE. We shall review some of their results in the next section. Several authors have discussed the estimation of a population size based on continuous time capture–recapture data under various settings; we refer to Chao and Lee (1993), Lloyd (1994), Wilson and Anderson (1995), Yip et al. (1996), Xi et al. (2007), and Yoshida et al. (1996) for further reading and additional references.

In this paper, we focus on estimating ν, assuming that all components are iid renewal processes. In contrast to Dewanji et al. (1995), which assumes a parametric family for the renewal density, we present a nonparametric estimator. Naturally, any estimator of ν derived for a parametric family of renewal distributions is likely to be sensitive to the assumptions about the renewal distribution. Because of mathematical and computational simplicity one may be tempted to use the MLE ($\hat{\nu}_{exp}$) of ν derived assuming that the renewal distribution is exponential. A natural question is: Can the simple estimate based on exponential distribution be used regardless of the true renewal distribution? Thus, we also explore the bias of $\hat{\nu}_{exp}$ when the true distribution is not exponential. We should note that both $\hat{\nu}_{np}$ and $\hat{\nu}_{exp}$ are infinite if and only if $M = R$. However, $P\,(M = R)$ is typically very small. Nevertheless, we compare them by their simulated mean, median and standard deviation, when they are finite.

In Section 2, we describe briefly some results of Nayak (1988) and Dewanji et al. (1995) for Poisson and parametric renewal processes. In Section 3, we introduce our nonparametric method along with two computational methods for obtaining the nonparametric MLE ($\hat{\nu}_{np}$) of ν. In Section 4, we discuss some asymptotic properties of $\hat{\nu}_{np}$, as $\nu \to \infty$. In Section 5, we investigate, by means of simulation, the performance of both $\hat{\nu}_{np}$ and $\hat{\nu}_{exp}$ under different renewal distributions. We find that the nonparametric estimator performs well in terms of bias and asymptotic normality, while $\hat{\nu}_{exp}$ can be severely biased when the true renewal distribution is not exponential.

2. Preliminaries

In a parametric setup, Dewanji et al. (1995) derived the likelihood function for the observed data described in the previous section as

$$L(\nu,\theta) = \frac{\nu!}{(\nu-r)!}[\overline{F}(\tau;\theta)]^{\nu-r}\prod_{i=1}^{r}\left[\left\{\prod_{j=1}^{m_i} f(t_{ij};\theta)\right\}\overline{F}(s_i;\theta)\right], \quad (1)$$

where $f(\cdot;\theta)$ is the renewal density function with $\overline{F}(\cdot;\theta)$ being the corresponding survival function. They also noted that for any given θ (with $F(\tau;\theta)<1$), the likelihood in (1) is maximized (with respect to ν) by

$$\hat{\nu}(\theta) = \left\lfloor \frac{r}{1-\overline{F}(\tau;\theta)} \right\rfloor, \quad (2)$$

where $\lfloor x \rfloor$ is the largest integer not exceeding x. Thus, if θ is known, $\hat{\nu}(\theta)$ in (2) is the MLE of ν. For most distribution families, the maximization of (1) with respect to θ, for given ν, can be done only numerically. For the joint estimation of ν and θ, Dewanji et al. (1995) suggested the following iterative procedure. Starting with an initial estimate $\nu^{(0)}$ of ν, calculate an estimate $\theta^{(0)}$ of θ by maximizing $L(\nu^{(0)},\theta)$. Then, use $\theta^{(0)}$ in (2) to obtain $\nu^{(1)} = \hat{\nu}(\theta^{(0)})$, and then calculate $\theta^{(1)}$ by maximizing $L(\nu^{(1)},\theta)$ with respect to θ. Repeat this process until convergence is achieved. In practice, one should use several initial estimates $\nu^{(0)}$ to be confident that a global maxima is obtained. We may also note that the MLE of ν and θ may not exist. For example, if the renewal distribution is exponential, the MLE exists if and only if $m>r$, where $m = \sum_{i=1}^{r} m_i$.

As a simpler alternative, Dewanji et al. (1995) suggested a conditional maximum likelihood (CML) method, where θ is estimated by maximizing (with respect to θ) the conditional likelihood, given $R{=}r$, which is given by

$$r!\prod_{i=1}^{r}\left[\frac{f(t_{i1};\theta)}{F(\tau;\theta)}\right]\prod_{i=1}^{r}\left[\left\{\prod_{j=2}^{m_i} f(t_{ij};\theta)\right\}\overline{F}(s_i;\theta)\right]. \quad (3)$$

The maximizer of (3), denoted by $\hat{\theta}_c$, is a conditional maximum likelihood estimator (CMLE) of θ. The CMLE of ν is then obtained by using $\theta = \hat{\theta}_c$ in (2).

Dewanji et al. (1995) also derived certain asymptotic properties (as $\nu \to \infty$) of the MLE and CMLE. In particular, (i) if (3) is maximized uniquely, then under certain mild conditions, $\hat{\theta}_c$ is a consistent estimator of θ, (ii) the CMLE and MLE are asymptotically equivalent and (iii) for a broad class of estimators $(\hat{\nu},\hat{\theta})$, which includes the MLE and the CMLE, as $\nu \to \infty$,

$$[\nu^{1/2}(\hat{\theta}-\theta), \nu^{-1/2}(\hat{\nu}-\nu)] \xrightarrow{L} \mathcal{N}(0,\Sigma), \quad (4)$$

where

$$\Sigma = \begin{pmatrix} I(\theta) & -\delta \\ -\delta' & F(\tau;\theta)/\overline{F}(\tau;\theta) \end{pmatrix}^{-1},$$