Journal of Statistical Planning and Inference 140 (2010) 992–1002

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Estimation procedures with delayed observations

Alicja Jokiel-Rokita, Ryszard Magiera *

Wrocław University of Technology, Poland

ARTICLE INFO

Article history:
Received 22 June 2007
Received in revised form
28 September 2009
Accepted 28 September 2009
Available online 4 October 2009

MSC:
primary 62L12, 62L15
secondary 62C10

Keywords:
Bayesian approach
Sequential procedure
Stopping time
Exponential family
Reliability

ABSTRACT

The statistical model is considered in which the collection of data from several independent populations is available only at random times determined by order statistics of lifetimes of a given number of objects. Each of the populations is distributed according to a general multiparameter exponential family. The problem is to estimate the mean value vector parameter of the multiparameter exponential family of distributions of the forthcoming observations. Under the loss function involving a weighted squared error loss, the cost proportional to the events appeared and a cost of observing the process, a class of optimal sequential procedures is established. The procedures are derived in two situations: when the lifetime distribution is completely known and in the case when it is unknown but assumed to belong to an exponential subfamily with an unknown failure rate parameter.

1. Introduction

We consider a class of estimation problems for data available at random times. For example, in studying the effectiveness of experimental safety devices of mobile constructions relevant data may become available only as a result of accident. Medical data, such as data on drug abuse or an asymptomatic disease, can sometimes only be obtained when patients voluntarily seek help or are somehow otherwise identified and examined, at random times. Other examples are data resulting from market research or from the scheduling of production sequence when the orders arrive at random times.

Let $Y_i,\ i=1,\ldots,m$, be random variables (vectors) such that the distribution of Y_i has a density (with respect to a certain dominating measure) proportional to

$$\exp[\vartheta_i^T G_i(y_i) - \Phi_i(\vartheta_i)], \tag{1}$$

i.e., we suppose that the distribution of Y_i belongs to an n_i−parameter canonical exponential family generated by $\Phi_i(\vartheta_i)$ and $G_i(y_i)$, $\vartheta_i=(\vartheta_{i,1},\ldots,\vartheta_{i,n_i})^T \in \Theta_i \subseteq \mathbb{R}^{n_i}$ (T denotes transposition), $i=1,\ldots,m$; $G_i(y_i)=(G_{i,1}(y_i),\ldots,G_{i,n_i}(y_i))^T$, $G_{i,j}(y_i):\mathbb{R}^{k_i}\to\mathbb{R}$, $j=1,\ldots,n_i$. We assume that Θ_i is the interior of the natural parameter space of the respective exponential family defined by (1).

Let $Y(J)=(Y_1(J),\ldots,Y_m(J))$, $J=1,\ldots,N$, be independent random vectors, where for each J the component $Y_i(J)$ is distributed according to the exponential family of (1).

* Corresponding author.
E-mail address: Ryszard.Magiera@pwr.wroc.pl (R. Magiera).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2009.09.023

It is supposed that $Y(J)$ is observed at time t_J, $J=1,\ldots,N$, where $t_1,\ldots,t_N$ are the order statistics, $t_J=X_{J:N}$, of positive exchangeable random variables $X_1,\ldots,X_N$ which are independent of $Y(1),\ldots,Y(N)$. The random variables $X_1,\ldots,X_N$ can be interpreted as the lifetimes of N objects. We assume that the random variables $X_1,\ldots,X_N$ are independent and have a common distribution function F.

Denote by

$$K(t)=\sum_{J=1}^{N}\mathbf{1}_{[0,t]}(X_J) \tag{2}$$

the number of observations which have been made by time $t\geq 0$. Thus, $(Y(1),\ldots,Y(K(t)))$ is a sample of random size $K(t)$ with components obtainable successively at random times $t_1,\ldots,t_{K(t)}$ and such that the distribution of $Y(J)$ has density

$$p(y(J);\vartheta)\propto\exp\left\{\sum_{i=1}^{m}[\vartheta_i^T G_i(y_i(J))-\Phi_i(\vartheta_i)]\right\}, \tag{3}$$

where $\vartheta=(\vartheta_1^T,\ldots,\vartheta_m^T)^T$.

The problem is to estimate the mean value vector of the exponential family of distributions from which the observations are arriving. It is assumed that the loss due to estimation error is a weighted squared error loss and that the cost of sampling involves a cost proportional to the number of arrivals and a cost of observing the process up to a stopping time.

A class of Bayes and minimax sequential estimation procedures from the delayed observations of the exponential family defined by (1) will be derived under some natural assumptions on the distribution of the random variables $X_1,\ldots,X_N$. The results will be given in the case when the common distribution of $X_1,\ldots,X_N$ is known exactly as well as in the case when this distribution is unknown but belongs to an exponential subfamily with an unknown failure rate parameter. The solutions constitute a generalization of the respective results on estimation obtained by Starr et al. (1976), who considered the problem of estimating a mean of delayed normally distributed observations with known variance in the case when the lifetime distribution is known completely or it is the commonly used exponential distribution, but with an unknown parameter.

2. Optimal stopping with known *F*

Let h be a given real valued function on $E_N=\{0,1,\ldots,N\}$, and such that $0\leq h(k)<\infty$ for each $k\in E_N$, and let

$$\mathfrak{L}_h(t)=\mathfrak{L}_h(K(t),t)=h(K(t))+c(t),$$

for $t\geq 0$. The function $c(t)$ is assumed to be a differentiable and non-decreasing function with $c(0)=0$.

The function $\mathfrak{L}_h(t)$ can be interpreted as the total loss incurred if the process is stopped at time t. In particular, when $h(k)$ is of the form

$$h(k)=l(k)+c_Ak,\quad c_A\geq 0,$$

then

$$\mathfrak{L}_h(t)=l(K(t))+c_AK(t)+c(t)$$

consists of the loss associated with the error of estimation $l(K(t))$, the cost proportional to the events appeared, $c_AK(t)$, and the cost $c(t)$ of observing the process up to time t.

Let

$$\mathcal{F}_t=\sigma\{K(s),s\leq t,Y(1),\ldots,Y(K(t))\}$$

be the information which is available at time t.

We search for a stopping time τ with respect to $\mathcal{F}_t,\ t\geq 0$, which will minimize the value

$$\mathcal{V}_h(\tau)=E[\mathfrak{L}_h(\tau)]=E[h(K(\tau))+c(\tau)].$$

Suppose that the lifetime distribution function F satisfies the following conditions: $F(0)=0$; $F(t)>0$ for $t>0$; F is absolutely continuous with density f; and f is the right hand derivative of F on $(0,\infty)$. The class of such F will be denoted by $\mathcal{G}$.

Let $\zeta=\sup\{t:F(t)<1\}$, and $\rho(t)=f(t)[1-F(t)]^{-1}$, $0\leq t<\zeta$, denote the failure rate. The process $K(t)$, $0\leq t\leq\zeta$, is a non-stationary Markov chain with respect to $\mathcal{F}_t$, $0\leq t\leq\zeta$, and its infinitesimal operator is

$$\mathcal{A}_th(k)=(N-k)[h(k+1)-h(k)]\rho(t)$$

for $k\in E_N=\{0,1,\ldots,N\}$ and all real valued functions h on E_N (see, for instance, Starr et al., 1976, p. 104).

The following lemma is a slight modification of Theorem 2.1 of Starr et al. (1976) for the cost function $c(t)$ instead of ct.

Lemma 1. *Suppose that $h(k)-h(k+1)$ is non-increasing for $k\leq N-1$ and that $F\in\mathcal{G}$. Moreover, assume that the function $c'(t)/\rho(t)$ is non-decreasing. Then the stopping time*

$$\tau_h=\inf\{t\geq 0:\mathcal{A}_th(K(t))+c'(t)\geq 0\}=\inf\{t\geq 0:[N-K(t)][h(K(t))-h(K(t)+1)]\leq c'(t)/\rho(t)\}$$

minimizes $\mathcal{V}_h(\tau)$.