Journal of Multivariate Analysis 147 (2016) 273–284

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Semi-parametric survival function estimators deduced from an identifying Volterra type integral equation

Gerhard Dikta[a,*], Martin Reißel[a], Carsten Harlaß[a,b]

[a] Fachhochschule Aachen, Heinrich-Mußmann Straße 1, D-52428 Jülich, Germany

[b] University of Wisconsin-Milwaukee, Department of Mathematical Sciences, PO Box 413, Milwaukee, WI 53201-0413, USA

ARTICLE INFO

Article history:
Received 15 October 2015
Available online 16 February 2016

AMS 2010 subject classifications:
62N01
62N02
62E20

Keywords:
Survival analysis
Censored data
Semi-parametric random censorship model
Asymptotic efficiency
Product-integration
Volterra integral equation

ABSTRACT

Based on an identifying Volterra type integral equation for randomly right censored observations from a lifetime distribution function F, we solve the corresponding estimating equation by an explicit and implicit Euler scheme. While the first approach results in some known estimators, the second one produces new semi-parametric and pre-smoothed Kaplan–Meier estimators which are real distribution functions rather than sub-distribution functions as the former ones are. This property of the new estimators is particular useful if one wants to estimate the expected lifetime restricted to the support of the observation time.

Specifically, we focus on estimation under the semi-parametric random censorship model (SRCM), that is, a random censorship model where the conditional expectation of the censoring indicator given the observation belongs to a parametric family. We show that some estimated linear functionals which are based on the new semi-parametric estimator are strong consistent, asymptotically normal, and efficient under SRCM. In a small simulation study, the performance of the new estimator is illustrated under moderate sample sizes. Finally, we apply the new estimator to a well-known real dataset.

1. Introduction

Lifetime or failure time data analysis is frequently based on incomplete observations where incompleteness is caused by some type of censoring. To handle censored observations, certain assumptions about the underlying censoring mechanism are necessary. One type of assumptions, which is widely accepted in practice, is described by the random censorship model (RCM). Under this model, one has two independent sequences of independent and identically distributed (IID) random variables: the survival times $X_1, \ldots, X_n$ and the censoring times $Y_1, \ldots, Y_n$. These sequences define the observations $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$, where $Z_i = \min(X_i, Y_i)$ and δ_i indicates whether the observation time Z_i is a survival time ($\delta_i = 1$) or a censoring time ($\delta_i = 0$). Here we assume that these sequences are defined over some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and we denote the distribution functions (DF) of X, Y, and Z by F, G, and H, respectively. Furthermore, we assume that all DFs are continuous. Note that the continuity of H guarantees that the observation times $Z_1, \ldots, Z_n$ are almost surely distinct.

Nonparametric statistical inference of F under RCM is usually built on the time-honored Kaplan–Meier (KM) or product limit estimator, see [18], defined by

$$F_n^{KM}(t) = 1 - \prod_{i:\, Z_i \le t} \left(1 - \frac{\delta_i}{n - R_{i,n} + 1}\right),$$

where $R_{i,n}$ denotes the rank of Z_i among the Z-sample.

[*] Corresponding author.
E-mail addresses: dikta@fh-aachen.de (G. Dikta), reissel@fh-aachen.de (M. Reißel), charlass@uwm.edu (C. Harlaß).

http://dx.doi.org/10.1016/j.jmva.2016.02.008

Besides other approaches, the KM-estimator can be derived from product-integration. As pointed out by Gill and Johansen in their survey paper, see [15], the KM-estimator is the product integral corresponding to the Nelson–Aalen estimator of the cumulative hazard function of F, see [22,1,17]. Precisely, the identifying Volterra type integral equation

$$F(t) = \int_0^t \bar{F}(s-) \Big/ \bar{H}(s-)\, H^1(ds), \tag{1}$$

where $H^1(s) = \mathbb{P}(\delta = 1, Z \le s)$, $\bar{F}(s) = 1 - F(s)$, and $\bar{H}(s) = 1 - H(s)$, is the starting point for this approach. Here $\bar{F}(s-)$ and $\bar{H}(s-)$ denote the corresponding left-hand limits. Note that due to the assumed continuity of the DFs, it is not necessary to take the left continuous version of the integrand in our setup. Therefore, we omit it in the text below.

Let $1_A(x) \equiv 1(x \in A)$ be the indicator function of the set A, and denote by

$$H_n(t) = \sum_{i=1}^n 1(Z_i \le t) \Big/ n, \qquad \bar{H}_n(t) = 1 - H_n(t), \qquad H_n^1(t) = \sum_{i=1}^n \delta_i 1(Z_i \le t) \Big/ n$$

the empirical counterparts of H, $\bar{H}$, and H^1. Furthermore, we will use $\hat{F}$ to express a generic estimator of F. The Kaplan–Meier estimator is then derived as the solution of the corresponding estimating equation

$$\hat{F}(t) = \int_0^t \bar{\hat{F}}(s-) \Big/ \bar{H}_n(s-)\, H_n^1(ds),$$

where $\bar{\hat{F}}(s-)$ and $\bar{H}_n(s-)$ denote the corresponding left-hand limits. As stated in [15], this is simply the result of applying an explicit Euler scheme with node points given by the ordered Z-sample, that is by $Z_{1:n}, \dots, Z_{n:n}$, for the approximated numerical solution of the original identifying integral equation. Historically, the application of an Euler scheme to obtain an approximate solution in product form of an initial value problem, or the equivalent Volterra integral equation, dates back to Volterra [30].

We will now outline this approach using a particular representation of the integrating measure H^1 in the identifying Eq. (1). As pointed out in [8], H^1 has a Radon–Nikodym density with respect to H, namely $H^1(dt) = m(t)H(dt)$, where $m(t) = \mathbb{E}(\delta \mid Z = t)$ is the conditional expectation of δ given $Z = t$. Therefore, we can rewrite Eq. (1) to get

$$F(t) = \int_0^t \bar{F}(s) \Big/ \bar{H}(s)\, m(s)\, H(ds) \tag{2}$$

which gives the estimating equation

$$\hat{F}(t) = \int_0^t (\bar{F}(s) \big/ \bar{H}(s))_n\, m_n(s)\, H_n(ds),$$

where $(\bar{F}(s) \big/ \bar{H}(s))_n$ and $m_n(s)$ are estimators of $\bar{F}(s) \big/ \bar{H}(s)$ and $m(s)$, respectively. Note that

$$\hat{F}(Z_{k:n}) = \hat{F}(Z_{k-1:n}) + \int_{]Z_{k-1:n}, Z_{k:n}]} (\bar{F}(s) \big/ \bar{H}(s))_n\, m_n(s)\, H_n(ds). \tag{3}$$

Substitute $(\bar{F}(s) \big/ \bar{H}(s))_n$ with $\bar{\hat{F}}(Z_{k-1:n}) \big/ \bar{H}_n(Z_{k-1:n})$ in the integrand (explicit Euler scheme) to get

$$\hat{F}(Z_{k:n}) = \hat{F}(Z_{k-1:n}) + \bar{\hat{F}}(Z_{k-1:n}) \big/ (n\bar{H}_n(Z_{k-1:n}))\, m_n(Z_{k:n}),$$

for $k = 1, \dots, n$, and conclude

$$1 - \hat{F}(Z_{k:n}) = \left(1 - \hat{F}(Z_{k-1:n})\right) \left(1 - \frac{m_n(Z_{k:n})}{n - k + 1}\right). \tag{4}$$

Obviously, Eq. (4) yields the typical product form:

$$1 - \hat{F}(Z_{k:n}) = \prod_{i=1}^k \left(1 - \frac{m_n(Z_{i:n})}{n - i + 1}\right).$$

But we still have to specify m_n to get the final estimator of F. This should be done using the available information about m. If nothing is known besides RCM, we can only use $\delta_{(k:n)}$, the associated indicator of $Z_{k:n}$, to estimate $m(Z_{k:n})$. In this case, Eq. (4) is exactly the Kaplan–Meier estimator F_n^{KM}. If we know, that m is a smooth function, we can use a nonparametric estimator of m, and get a pre-smoothed Kaplan–Meier estimator $F_{1,n}^{PR}$, see [32,4]. We can also interpret $(\delta_1, Z_1), \dots, (\delta_n, Z_n)$ as observations from a binary regression model, where the regression function is given by m. If we have good reasons to assume that m belongs to a parametric family, that is,

$$m(t) = m(t, \theta_0),$$