Journal of the Korean Statistical Society 45 (2016) 568–580

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Some heuristics about bandwidth selection for the smooth Kaplan–Meier estimator

É. Youndjé

Laboratoire Raphaël Salem UMR 6085 CNRS Université de Rouen, France

ARTICLE INFO

Article history:
Received 2 October 2015
Accepted 11 April 2016
Available online 5 May 2016

AMS 2000 subject classifications:
primary 62G86
secondary G2N86

Keywords:
Censored data
Survival analysis
Kaplan–Meier estimator
Kernel smoothing
Bandwidth selection

ABSTRACT

Estimation of the distribution function in the censorship model is very important in survival analysis. In this work, three bandwidth selection methods for the smooth Kaplan–Meier estimator are introduced. The simulations conducted in the paper reveal that the methods are very promising. Thus, this paper offers hopeful solutions to this "old" problem.

1. Introduction

Let X^a be a nonnegative random variable with distribution function F^a and density function f^a, likewise let X^b be a nonnegative random variable independent of X^a with f^b and F^b respectively as density and distribution function. Assume that the observations

$$(X_i, \delta_i^a), \quad i = 1, \dots, n$$

are an i.i.d. sample from the random pair defined by

$$X = \min(X^a, X^b), \qquad \delta^a = \mathbb{1}_{[X^a \le X^b]}.$$

Let us set

$$\delta^b = 1 - \delta^a \quad \text{and} \quad \delta_i^b = 1 - \delta_i^a.$$

We are interested in the estimation of F^a and F^b using the data $(X_i, \delta_i^a)_{i=1,\dots,n}$. For $t \in \{a,\ b\}$ the most widely used/known estimator of F^t is the Kaplan–Meier estimator defined by

$$1 - F_n^t(x) = \prod_{i:\ X_{(i)} \le x} \left(1 - \frac{\delta_{(i)}^t}{n - i + 1}\right). \tag{1}$$

E-mail address: Elie.Youndje@univ-rouen.fr.

http://dx.doi.org/10.1016/j.jkss.2016.04.002
1226-3192/© 2016 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

Here $(X_{(i)}, \delta^t_{(i)}),\ i = 1, \ldots, n$ are the n pairs of observations ordered by the $X_{(i)}$ i.e. $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. It is shown in the literature that the estimator $F^t_n(x)$ shares many of the interesting properties of the classical empirical distribution function. To cite a few of these properties, $F^t_n(x)$ is the maximum likelihood estimator of $F^t(x)$ (Kaplan & Meier, 1958), is asymptotically normal (Stute, 1995), is strongly consistent (Peterson, 1977), and is uniformly strongly consistent (Stute & Wang, 1993). Despite these good properties, the estimator F^t_n is a step function and can be less appealing when the target distribution function F^t is continuous. A smooth version of the Kaplan–Meier estimator of F^t was introduced in Földes, Rejtő, and Winter (1981), and was further studied in Ghorai and Susarla (1990) and Kulasekera, Williams, Coffin, and Manatunga (2001). Let us conclude this paragraph by mentioning a few recent works on nonparametric estimation with right censored data. For quantile (hence cumulative distribution function) estimation we can cite Hong, Kim, and Kim (2013), for hazard estimation we cite Ouadah (2013), and, finally for density function estimation we refer to Jácome and Cao (2008) and Jomhoori, Fakoor, and Azarnoosh (2012).

Since it is assumed that X^a and X^b have f^a and f^b as densities, the random variable $X = \min(X^a, X^b)$ has a density. Let F and f denote the distribution and density function of X respectively. The kernel estimator of F based on $X_1, \ldots, X_n$ is given by (see Nadaraya, 1964)

$$\begin{aligned} F_h(x) &= \int \frac{1}{h} K\left(\frac{x-y}{h}\right) F_n(y)\mathrm{d}y \\ &= \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x-X_i}{h}\right), \end{aligned}$$

where F_n is the empirical estimator of F, K a nonnegative kernel function and H its distribution function i.e.

$$H(x) = \int_{-\infty}^{x} K(t)\mathrm{d}t,$$

and $h = h(n) > 0$, is the smoothing parameter. Likewise, the kernel estimator of F^a based on the censored data $(X_i, \delta^a_i)_{i=1,\ldots,n}$ is obtained by replacing the empirical distribution function by the Kaplan–Meier estimator i.e. (see for example Ghorai & Susarla, 1990)

$$F^a_{h_1}(x) = \int \frac{1}{h_1} K\left(\frac{x-y}{h_1}\right) F^a_n(y)\mathrm{d}y$$

where $h_1 = h_1(n)$ is the bandwidth. Similarly the smooth Kaplan–Meier estimator of F^b with bandwidth $h_2 = h_2(n)$ is defined by

$$F^b_{h_2}(x) = \int \frac{1}{h_2} K\left(\frac{x-y}{h_2}\right) F^b_n(y)\mathrm{d}y.$$

The aim of this article is to suggest a way to use the observations to determine the bandwidths h_1 and h_2. In the subsequent sections of this paper three approaches based on cross-validation ideas are introduced and studied.

The next section outlines a generalization of the Bowman, Hall, and Prvan (1998) method to censored data. Section 3 introduces the basic ideas behind the two other methods of this article. Section 4 studies a method based on the Average Square Errors (ASE) measure of accuracy, while Section 5 presents the Average Absolute Errors (AAE) approach. Section 6 discusses some of the lessons drawn from the simulations. The remainder of the paper is comprised of proofs.

2. Extension of the Bowman et al. method to the censorship model

We will use the following quantities in the sequel

$$\begin{aligned} \tau_a &= \tau_{X^a} = \inf\{x \mid F^a(x) = 1\} \\ \tau_b &= \tau_{X^b} = \inf\{x \mid F^b(x) = 1\} \\ \tau &= \tau_X = \inf\{x \mid F(x) = 1\}. \end{aligned}$$

It follows immediately from the independence of X^a and X^b and from the equality $X = \min(X^a, X^b)$ that $\tau = \min(\tau_a, \tau_b)$, and

$$1 - F(x) = (1 - F^a(x))(1 - F^b(x)). \tag{2}$$

The goal of this section is to show what motivates us to develop the main methods of this paper. These methods are presented in Sections 3–5 and are applicable when F^a and F^b are continuous. Despite the fact that the methods introduced in this section are not theoretically fully justified, they are very interesting and we believe (see Theorem 1) that these methods can be used when one of the distribution functions is not continuous. For example if F^b is not continuous the criterion $CV(F^a_{h_1})$ in (5) below can be used to obtain a bandwidth for $F^a_{h_1}$. Also, as will be seen in Section 4, these methods give good results when n is large enough and F^a and F^b are continuous.