Computational Statistics and Data Analysis 93 (2016) 308–323

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Parametrically guided nonparametric density and hazard estimation with censored data

Majda Talamakrouni, Ingrid Van Keilegom *, Anouar El Ghouch

Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université Catholique de Louvain, Belgium

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form 9 January 2015
Accepted 15 January 2015
Available online 29 January 2015

Keywords:
Cox model
Density estimation
Kaplan–Meier estimator
Kernel smoothing
Maximum likelihood
Right censoring

ABSTRACT

The parametrically guided kernel smoother is a promising nonparametric estimation approach that aims to reduce the bias of the classical kernel density estimator without increasing its variance. Theoretically, the estimator is unbiased if a correct parametric guide is used, which can never be achieved by the classical kernel estimator even with an optimal bandwidth. The estimator is generalized to the censored data case and used for density and hazard function estimation. The asymptotic properties of the proposed estimators are established and their performance is evaluated via finite sample simulations. The method is also applied to data coming from a study where the interest is in the time to return to drug use.

1. Introduction

Censored data appear in a broad variety of research studies with practical applications. Random right censoring is one of the most common types of censoring. For example in medical, economic or engineering studies, it frequently happens that the variable of interest T is only partially observed due to the earlier occurrence of a censoring event. In such studies, the estimation of the probability density and hazard function of T has received considerable attention in the literature, as it allows to visualize and explore the distribution of data.

In this paper we wish to estimate the density and hazard function when T is subject to right censoring, by using a hybrid estimation method that has at the same time nonparametric and parametric ingredients. These two extremal estimation approaches have rather opposite characteristics. The fully parametric approach is accurate and powerful when the parametric family is correctly chosen, otherwise it can lead to incorrect inference. The fully nonparametric approach includes several methods, among which the popular kernel smoothing procedure. It is very flexible, since it does not rely on any restrictive assumptions about the form of the underlying density or hazard function. However, the resulting estimator has typically a slower rate of convergence.

In the case where the data are not subject to censoring, there is a large variety of approaches to estimate the density and the hazard function that are either semiparametric or that use aspects from both the nonparametric and the parametric school, and that are hence situated in between these two extreme approaches. One of these approaches is the parametrically guided nonparametric estimator proposed by Hjort and Glad (1995). Apart from reducing the bias compared to the classical kernel approach, the parametrically guided nonparametric approach allows for a theoretically unbiased estimator, which is

* Correspondence to: Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium.

E-mail address: ingrid.vankeilegom@uclouvain.be (I. Van Keilegom).

http://dx.doi.org/10.1016/j.csda.2015.01.009

impossible with the classical kernel approach. The basic idea of this approach is to start with any parametric density estimator and then to adjust this first stage parametric approximation using a nonparametric kernel-type estimator of a particular correction factor. More precisely, the key identity underlying the parametrically guided nonparametric approach is

$$f(t) = f_{\widehat{\theta}}(t) r_{\widehat{\theta}}(t),$$

where $r_{\widehat{\theta}}(t) = \frac{f(t)}{f_{\widehat{\theta}}(t)}$, $f_{\widehat{\theta}}(t)$ is a first stage parametric density approximation and $\widehat{\theta}$ is an estimator of the least false value θ^* according to a certain distance measure between $f(\cdot)$ and $f_\theta(\cdot)$ (see Assumption 3.3, below). Hjort and Glad (1995) defined the parametrically guided nonparametric estimator by

$$\widehat{f}_{\widehat{\theta}}(t) = f_{\widehat{\theta}}(t) \widehat{r}_{\widehat{\theta}}(t), \tag{1.1}$$

where $\widehat{r}_{\widehat{\theta}}(\cdot)$ is a kernel-type nonparametric estimator of the correction factor $r_{\widehat{\theta}}(\cdot)$. Essentially, this multiplicative correction does not affect the variance but can reduce the bias. The intuitive idea behind this approach is that if the parametric estimator $f_{\widehat{\theta}}(\cdot)$ is close to the true density $f(\cdot)$, the multiplicative correction function $r_{\widehat{\theta}}(\cdot)$ will be smoother than the true density $f(\cdot)$ and therefore simpler to estimate using kernel smoothing, resulting in an improved $\widehat{f}_{\widehat{\theta}}(\cdot)$ compared to the traditional kernel estimator. If the true density is far from the parametric estimator, then there is not much loss in accuracy for the parametrically guided nonparametric estimator.

The aim of this paper is to extend their method to the case of censored data. To the best of our knowledge, except for the recent work of Talamakrouni et al. (in press), who studied a guided local linear estimator of a regression function when the response is subject to censoring, the parametrically guided nonparametric method has never been investigated in the context of censored data. In addition to studying the estimation of the density function, we also propose and study a parametrically guided nonparametric estimator of the hazard rate function in the presence of censoring.

Apart from the above parametrically guided nonparametric estimator of Hjort and Glad (1995), there have been other proposals in the literature that combine the nice features of both the parametric and the nonparametric approach. These methods are quite different but can also achieve bias reduction compared to the fully nonparametric method. As far as we are aware of, except for the paper of Copas (1995) who adapted a local maximum likelihood estimator to censored data, none of them has been considered so far in the context of censored data. First of all, we find the projection pursuit density estimation developed by Friedman et al. (1984) for a multivariate density using a similar multiplicative correction. Hjort (1986) and Buckland (1992) introduced similar ideas using an estimated orthogonal expansion for the multiplicative correction factor. Hjort and Jones (1996) proposed a local parametric density estimator based on a local kernel smoothed likelihood function. This approach has a similar intention as the approach of Copas (1995) but is somehow more general. Another class of local likelihood methods has been discussed by Eguchi and Copas (1998). Efron and Tibshirani (1996) combined the maximum likelihood and the kernel estimator by putting an exponential family through a kernel estimator. Other semiparametric estimators involving an extra parameter have been proposed in the literature. For example, Olkin and Spiegelman (1987) and Faraway (1989) considered a convex combination of a parametric and a nonparametric estimate, and afterwards, Naito (2004) constructed a class of semi-parametric estimators using a local L_2-fitting criterion to estimate the correction factor. Finally, more recently, Veraverbeke et al. (2014) discussed a parametrically pre-adjusted nonparametric density estimator.

Parallel to this vast literature on parametrically guided density estimation, there also exists a large literature on parametrically guided nonparametric regression. We mention for example Glad (1998), Martins-Filho et al. (2008) and Fan et al. (2009), among others.

The paper is organized as follows. The next section explains in detail the proposed methodology. Section 3 provides some asymptotic results for the proposed estimators, while Section 4 investigates the finite sample properties of the new estimators. In Section 5 we apply the proposed method to data on the time to return to drug use from a study of the AIDS research unit of the University of Massachusetts. Finally, some general conclusions are drawn in Section 6. The proofs are collected in the Appendix.

2. Methodology

Let T be a variable of interest with density f and distribution function F, and let C be a censoring variable with continuous distribution function G. We assume throughout our paper that T is independent of C. Under random right censoring, the variable T is not completely observed. One can only observe (X, δ), where $X = \min(T, C)$, $\delta = I(T \leq C)$ and $I(\cdot)$ is the indicator function. Our first objective is to estimate the probability density function f using the observed i.i.d. sample (X_i, δ_i), $i = 1, \ldots, n$ of (X, δ).

The kernel-based density estimator that we are currently investigating has been extended to censored data by Blum and Susarla (1980), among others. The estimator is based on the Kaplan and Meier (1958) estimator $\widehat{F}$ of the distribution function F and is defined as follows:

$$\widehat{f}(t) = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{h}\right) d\widehat{F}(s), \tag{2.2}$$