Journal of Statistical Planning and Inference 184 (2017) 25–47

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Adaptive estimation of the hazard rate with multiplicative censoring

G. Chagny [a,*], F. Comte [b], A. Roche [c]

[a] LMRS, UMR CNRS 6085, Université de Rouen Normandie, France
[b] MAP5 UMR CNRS 8145, Université Paris Descartes, France
[c] CEREMADE UMR CNRS 7534, Université Paris Dauphine, France

ARTICLE INFO

Article history:
Received 9 September 2016
Accepted 23 November 2016
Available online 8 December 2016

MSC:
62G05
62N01

Keywords:
Adaptive procedure
Model selection
Hazard rate estimation
Multiplicative censoring model

ABSTRACT

We propose an adaptive estimation procedure of the hazard rate of a random variable X in the multiplicative censoring model, $Y = XU$, with $U \sim \mathcal{U}([0, 1])$ independent of X. The variable X is not directly observed: an estimator is built from a sample $\{Y_1, \ldots, Y_n\}$ of copies of Y. It is obtained by minimisation of a contrast function over a class of general nested function spaces which can be generated e.g. by splines functions. The dimension of the space is selected by a penalised contrast criterion. The final estimator is proved to achieve the best bias–variance compromise and to reach the same convergence rate as the oracle estimator under conditions on the maximal dimension. The good behavior of the resulting estimator is illustrated over a simulation study.

1. Introduction

In this paper, our aim is to estimate the hazard rate associated with a nonnegative random variable X, defined by

$$h = f_X / \bar{F}_X$$

where $\bar{F}_X(x) = 1 - F_X(x) = \mathbb{P}(X \geq x)$ (resp. f_X) is the survival function (resp. the density) of X. However, instead of having at our disposal an independent and identically distributed (*i.i.d.* in the sequel) sample $X_1, \ldots, X_n$ with the distribution of X, we assume that we observe $\{Y_1, \ldots, Y_n\}$ such that

$$Y_i = X_i U_i, \quad \text{for all } i = 1, \ldots, n, \tag{1}$$

where X_i is a non-negative unobserved random variable and U_i is also unobserved and follows the uniform distribution on the interval $[0, 1]$. The quantity of interest, X_i, is supposed to be independent of U_i, for all $i = 1, \ldots, n$. The model $Y = XU$ is called a multiplicative censoring model by Vardi (1989), but we emphasise that this kind of censoring is very different from more standard right-censoring. It represents incomplete observations when the lifetime at hand is only known to belong to a random time interval; this happens in AIDS studies in particular.

Model (1) has been introduced by Vardi (1989) as a common model for different statistical problems such as inference for renewal processes, deconvolution with exponential noise and density estimation under decreasing constraint. Until now,

* Corresponding author.
E-mail addresses: gaelle.chagny@univ-rouen.fr (G. Chagny), fabienne.comte@parisdescartes.fr (F. Comte), angelina.roche@dauphine.fr (A. Roche).

http://dx.doi.org/10.1016/j.jspi.2016.11.002
0378-3758/

only density and survival function estimation have been studied in this model. First, a maximum likelihood estimation procedure has been introduced by Vardi (1989) and shown to be consistent. Then, Vardi and Zhang (1992) have proved its uniform consistency and asymptotic normality, but in both papers, it is assumed that two samples $\{Y_1, \dots, Y_n\}$ and $\{X_1, \dots, X_m\}$ are observed and that $m/(m+n)$ converges to a positive constant $c > 0$. More recently, Asgharian et al. (2012) have proposed a kernel density estimator and established conditions for strong uniform consistency. All previous results are not applicable in our context as we assume $m = 0$.

In our setting, where X is not directly observable, no estimation procedure of the hazard rate has been proposed, to our knowledge. Nonparametric hazard rate estimation has been developed in the context of direct or right-censored observations, mainly with quotient of functional estimators, built by kernel methods as in Patil (1993a,b), wavelet strategies as in Antoniadis et al. (1999), or projection and model selection techniques as in Brunel and Comte (2005). In the present paper, our references on the topic are two studies dealing with nonparametric estimation of hazard rate in the context of direct observations or right-censored data developed in Comte et al. (2011) and in Plancade (2011). We show how to generalise the method proposed in these papers to model (1); their specificity is to propose an adaptive regression estimator built by direct contrast minimisation (no quotient) and model selection. We do not provide exhaustive bibliography on the subject, but the interested reader is referred to the recent paper of Efromovich (2016) and references therein.

Concerning the specific model considered here, we obtain the following relationship between the density f_Y of Y and the density f_X from the link between the random variables given by (1),

$$f_Y(y) = \int_y^{+\infty} \frac{f_X(x)}{x} dx, \quad y > 0.$$

This formula indicates that estimating the density of X from the density of the observed variable Y is an inverse problem. Based on this observation, Andersen and Hansen (2001) have proposed an estimation procedure of the density f_X by a series expansion approach. Convergence rates for the mean integrated squared error are derived. Van Es et al. (2005) have proposed an estimation procedure of the density of $\log(X^2)$ in the non *i.i.d.* case, under a different assumption on the law of U. Abbaszadeh et al. (2012, 2013), Chesneau (2013) and Chaubey et al. (2015) have proposed adaptive wavelet estimators of the density f_X, when the observations follow related – yet different – models, for instance with an additional bias on X (Abbaszadeh et al., 2012), in the non *i.i.d.* case (Chesneau, 2013) or under the assumption that X follows a mixing distribution (Chaubey et al., 2015). They obtain convergence rates for the $\mathbb{L}^2$-risk (or even the $\mathbb{L}^p$-risk Abbaszadeh et al., 2013) on $[0, 1]$. Brunel et al. (2016) have proposed an adaptive estimation procedure for both the density f_X and the survival function $\bar{F}_X$ in the case where X can take negative values. They also obtain rates of convergence, for both integrated and pointwise quadratic risk which are similar to the ones obtained by Andersen and Hansen (2001), though under different regularity assumptions on the functions to estimate. These rates are proved to be optimal in the minimax sense. Comte and Dion (2016) have proposed an adaptive estimation procedure for the density function in a different context, the noise is supposed to be uniform over an interval $[1 - a, 1 + a]$ $(a > 0)$. None of the previous works considers hazard rate estimation, while this function is widely used in survival analysis.

In this paper, we provide a projection strategy for the estimation of the hazard rate function h, following the ideas developed by Comte et al. (2011) and by Plancade (2011). To this end, we take into account the specific model (1) and propose an original minimum contrast estimator. We first build a collection of projection estimators over linear models, and then choose an estimate in the collection, by using model selection. In Section 2, we detail the estimation procedure for a fixed model and justify the choice of our contrast. We give theoretical results in Section 3. In Section 4, we define the empirical criterion for choosing the model dimension and provide theoretical results (oracle-inequality and rates of convergence) for the selected estimator. Finally, in Section 5, we study the numerical behavior of the proposed estimator. Section 6 is devoted to the proofs.

2. Estimation procedure

2.1. Notations

We estimate the target function h on a compact subset $A = [0, \mathbf{a}]$ of $[0, +\infty[$. Let $(\mathbb{L}^2(A), \|\cdot\|, \langle ., . \rangle)$ be the space of square integrable functions on A, equipped with its classical Hilbert structure: $\langle f, g \rangle = \int_A f(t)g(t)dt$ and $\|f\| = \sqrt{\langle f, f \rangle}$, for all $f, g \in \mathbb{L}^2(A)$. We also introduce $\|\cdot\|_{\bar{F}_X}$, a reference semi-norm that naturally appears in our estimation problem, given by $\|t\|^2_{\bar{F}_X} := \langle t, t \rangle_{\bar{F}_X}$, $\langle s, t \rangle_{\bar{F}_X} := \int_A s(x)t(x)\bar{F}_X(x)dx$, for $s, t \in \mathbb{L}^2(A)$. It satisfies $\|t\|_{\bar{F}_X} \leq \|t\|$. We also denote $\|f\|_{\infty,I} := \sup_{x \in I} |f(x)|$, and $\|f\|_{p,I}$ the classical $\mathbb{L}^p$-norm of a function f on an interval $I \subset \mathbb{R}$.

We consider a collection $(S_m)_{m \in \{1,\dots,N_n\}}$ of linear subspaces such that

$$S_m = \text{Span}\{\varphi_j,\ j \in \mathbb{J}_m\},$$

where $\mathbb{J}_m \subset \mathbb{N} \setminus \{0\}$, $N_n \geq 1$, $\{\varphi_j,\ j \in \mathbb{J}_m\}$ is a basis of the subspace, and φ_j has support in A. We denote by D_m the dimension of S_m, which means that $D_m = |\mathbb{J}_m|$, where $|B|$ denotes the cardinality of a set B. The following properties are required for the models.