SEMI-RECURSIVE KERNEL CONDITIONAL DENSITY ESTIMATORS UNDER RANDOM CENSORSHIP AND DEPENDENT DATA

Sihem Semmar, Ali Laksaci, Salah Khardani
Djilali Liabes University. Fac. of Science, Algeria, doc.semmar@gmail.com
King Khalid University, Abha, Kingdom of Saudi Arabia.
Ecole Nationale des Sciences et Technologies, Tunisia

ABSTRACT

In this work, we extend to the case of the strong mixing data the results of Khardani and Semmar. A kernel−type recursive estimator of the conditional density function is introduced. We study the properties of these estimators and compare them with Rosemblatt’s nonrecursive estimator. Then, a strong consistency rate as well as the asymptotic distribution of the estimator are established under an $\alpha$-mixing condition. A simulation study is considered to show the performance of the proposed estimator.

Keywords: Asymptotic normality, censored data, conditional density, mixing sequences, recursive kernel estimators, survival data, strong consistency.

1. INTRODUCTION

It is well known that censored data are, usually, used to describe many practical applications in a wide variety of fields, including economics, medicine, biology, and biostatistics. For example, let $T$ be the survival time of individuals who are involved in a clinical study or a possible monotone transformation of the survival time. Consider a random covariable $X$, such as the age, the dose of a drug or the cholesterol level. As often occurs in practice, the response $T$ is subject to random right censoring. In other words, instead of observing $T$, one observes the pair of variables $(Y, \delta)$, where $Y = \min(T, C)$, $\delta = 1_{T \leq C}$, and $C$ represents the censoring time, which is expected to be conditionally independent of $T$ given $X$.

In this article we focus on the nonparametric estimation of the conditional density function of response censored variable $T$ given a vectorial variable $X$. We estimate this model by a recursive kernel approach and we derive its asymptotic properties when the observations are dependent. We note that this subject of the conditional density estimation plays an important role not only in the prediction of response variable given a regressor but it is a fundamental tool to explore all relationships between responses and covariates.

The main purpose of this contribution is to study the recursive estimator of the conditional density when the response variable is subject to random right censoring and the observations are $\alpha$− mixing process. We establish the almost complete consistency as well as the asymptotic normality of the constructed estimator. All these asymptotic results are obtained with precision of the convergence rate and are established under some standard conditions. It should be noted that the present contribution generalize to the dependent case the results of Khardani and Semmar (2014) in the iid case. Recall that this generalization to time series case is not a simple extension but it requires some alternative and additional tools. We point out the density estimation in censored time series data has been studied by Khardani, Lemdani, and Ould Saïd (2010, 2011) by using...
the classical kernel method. Alternatively in the present work, we use the recursive approach and we treat the conditional case. The benefit of this consideration is double: First, the recursivity offers, rapid, accurate and robust estimator. Second, conditioning allows to provide predictors with more specification. From practical point of view, our approach is very useful in big data analysis which has a lot of interest in the last years. In particular the Internet network provide massive data arriving in streams (including Twitter activity, the Facebook news stream, Internet packet data, stock market activity, credit card transactions and Internet and phone usage), and if they are not processed immediately or stored, then they are lost forever. Thus, data streams have become an increasingly important filed of research of statistical analysis. In this context, building a semi-recursive estimator which does not require to store all the data in memory and can be updated easily in order to deal with online data is of great interest. Indeed, the recursive estimator can be updated with each new observation which permits to save computational time and storage memory, unlike to the non-recursive case where the estimator needs to be recalculated completely when a new data set is observed. This gain is more important in the conditional density estimation, since the number of points at which the function is estimated is usually very large.

2. RECURSIVE KERNEL CONDITIONAL DENSITY UNDER A RANDOM CENSORSHIP

Let \((X_i, T_i)\) be a \(\mathbb{R}^d \times \mathbb{R}\) valued stationary strongly mixing or \(\alpha\)-mixing process defined on probability space \((\Omega, \mathcal{A}, P)\) : The object of this article is to study the conditional density of \(T_i\) given \(X_i = x\) which is obtained by

\[
\phi(t|x) = \frac{g(x,t)}{\ell(x)}
\]

(1)

where \(g(\ldots)\) denotes the joint density of \((T_i, X_i)\) and \(\ell(.)\) denotes the marginal density of \(X_i\). Typically, we consider the case of random right censorship. Formally we suppose that the \((T_i)_{1 \leq i \leq n}\) generated form a stationary \(\alpha-\)mixing sequence with unknown continuous distribution function (df) \(F\) and density \(g\). In this situation of censored lifetimes model we assume that there exist a sample of independent and identically distributed (iid) censoring random variables (rvs) \((C_i)_{1 \leq i \leq n}\) with continuous df \(G\), such that \(T_i\) is observed only if \(T_i \leq C_i\). Then we proceed, with the \(n\) pairs \((Y_i, \delta_i)\) with

\[
Y_i = T_i \wedge C_i \quad \text{and} \quad \delta_i = 1_{T_i \leq C_i}, \quad 1 \leq i \leq n.
\]

(2)

where \(1_A\) denotes the indicator function of the set \(A\). To follow the convention in biomedical studies, we assume that \((C_i)_{1 \leq i \leq n}\) and \((T_i, X_i)_{1 \leq i \leq n}\) are independent; this condition is plausible whenever the censoring is independent of the modality of the patients.

From the observations \((X_1, Y_1, \delta_1), \ldots, (X_n, Y_n, \delta_n)\) of \((X, Y, \delta)\), the kernel estimate of the conditional density \(\phi(t|x)\) denoted \(\hat{\phi}_n(t|x)\) is defined by

\[
\hat{\phi}_n(t|x) = \frac{\sum_{i=1}^{n} h_n^{-1} \delta_i g_n^{-1}(Y_i) K \left( \frac{x - X_i}{h_n} \right) L \left( \frac{t - Y_i}{h_n} \right)}{\sum_{i=1}^{n} K \left( \frac{x - X_i}{h_n} \right)} \cdot \frac{g_n(x,t)}{\ell_n(x)}
\]

(3)

where \(K, L\) are kernels and \(h_n\) is a sequence of positive real numbers.
A recursive version of the previous kernel estimator is defined by

$$
\hat{g}_n(t|x) = \frac{\sum_{i=1}^{n} h_i^{-d} \delta G_n^{-1}(Y_i)K \left( \frac{x - X_i}{h_i} \right) L \left( \frac{t - Y_i}{h_i} \right)}{\sum_{i=1}^{n} h_i^{-d} K \left( \frac{x - X_i}{h_i} \right)}
$$

$$
\hat{g}_n(x,t) = \frac{\hat{g}_n(x,t)}{\ell_n(x)}
$$

where

$$
\hat{g}_n(x,t) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i^{d+1}} \delta \hat{G}_n^{-1}(Y_i)K \left( \frac{x - X_i}{h_i} \right) L \left( \frac{t - Y_i}{h_i} \right), \quad \forall x \in \mathbb{R}^d.
$$

and

$$
\ell_n(x) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i^{d}} K \left( \frac{x - X_i}{h_i} \right), \quad \forall x \in \mathbb{R}^d.
$$

3. SIMULATIONS RESULTS

In this section, we discuss the feasibility and the performance of the recursive nonparametric method through an empirical study. Precisely, our main purpose is to compare the ratio (efficiency/rapidity) between recursive and classical kernel conditional density estimation in the time series case.
FIGURE 2 – Recursive and non-recursive estimators for a skewed unimodal distribution case.

FIGURE 3 – Recursive and non-recursive estimators for a skewed bimodal distribution.

4. THE BIBLIOGRAPHY


ICMA2021-4


