Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5800–5812
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Plug-in bandwidth selection in kernel hazard estimation from dependent data

Alejandro Quintela-del-Río*

Departamento de Matemáticas, Facultad de Informática, Universidad de A Coruña, Campus de Elviña, s/n, 15071 A Coruña, Spain

Received 17 January 2005; received in revised form 14 October 2006; accepted 15 October 2006
Available online 7 November 2006

Abstract

The plug-in bandwidth selection method in nonparametric kernel hazard estimation is considered, and a weak dependence on the sample data is assumed. A general result of asymptotic optimality for the plug-in bandwidth is presented, that is valid for the hazard function, as well as for the density and distribution functions. In a simulation study, this method is compared with the "leave more than one out" cross-validation criterion under dependence. Simulations show that smaller errors and much less sample variability can be reached, and that a good selection of the pilot bandwidth can be done by means of "leave one out" cross-validation. Finally, an application to an earthquake data set is made.

Keywords: Kernel estimation; Hazard; Distribution function; Density function; Strong mixing processes; Bandwidth selection; Plug-in

1. Introduction

Let us consider a real random variable X, and assume that X has a continuous distribution function F and a probability density function f. In this context, we also can describe the distribution of X by means of other equivalent functions, such as the failure-rate function or hazard function

$$r(x) = \frac{f(x)}{1 - F(x)}, \tag{1}$$

where $1 - F(x) > 0$, that is, $r(\cdot)$ is defined in the set $S = \{x \in \mathbb{R}/1 - F(x) > 0\}$. Given a random sample $X_1, \ldots, X_n$, each X_i having the same distribution as X, one among the conventional nonparametric estimators of $r(\cdot)$ is the kernel estimator, defined by

$$r_h(x) = \frac{f_h(x)}{1 - F_h(x)}. \tag{2}$$

* Tel.: +34 98 1167000; fax: +34 98 1167160.
E-mail address: aquintela@udc.es.

0167-9473/$ - see front matter
doi:10.1016/j.csda.2006.10.010

In this expression

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3}$$

is the known Parzen–Rosenblatt estimator of $f(\cdot)$, and $F_h(\cdot)$ is the kernel estimator of $F(\cdot)$, defined by

$$F_h(x) = \frac{1}{n}\sum_{i=1}^{n} K^*\left(\frac{x - X_i}{h}\right) = \int_{-\infty}^{x} f_h(t)\,\mathrm{d}t, \tag{4}$$

with $K(\cdot)$ a kernel function, $K^*(x) = \int_{-\infty}^{x} K(u)\,\mathrm{d}u$, and $h = h(n) \in \mathbb{R}^+$ is the smoothing parameter, or bandwidth.

Hazard estimation is of interest in several fields of applied statistics (medicine, reliability,...). Nonparametric estimation of hazard function was introduced in the statistical literature by Watson and Leadbetter (1964). Other authors, among which are Ahmad (1976), Singpurwalla and Wong (1983) and Hassani et al. (1986), work in this field, but all considering independence in the data.

One of the most appealing applications of hazard estimation is to analyze the structure of earthquakes, by considering the random variable X as the difference between the time of occurrence of two consecutive earthquakes, and studying its hazard function, or in this case, the failure rate of X (Estévez et al., 2002a).

A practical use of the hazard estimate, and for the density and distribution also, requires establishing a criterion to select the bandwidth or smoothing parameter h. Estévez and Quintela (1999) work with the cross-validation criterion in a dependence context—strong mixing (Rosenblatt, 1956)—and Estévez et al. (2002b) extend this method by using a penalizing approach that works better in finite samples. In both papers, practical applications are made to earthquakes data from different regions.

The plug-in methodology works in this way: to assess the global performance of $g_h(\cdot)$ as an estimator of $g(\cdot)$ (where g can be the density, distribution, or hazard function and g_h are (3), (4) and (2), respectively) we will consider the following quadratic measures of accuracy:

The integrated squared error (*ISE*)

$$ISE(h) = \int (g_h(x) - g(x))^2 w(x) f(x)\,\mathrm{d}x, \tag{5}$$

and the mean integrated squared error (*MISE*)

$$MISE(h) = E\left(\int (g_h(x) - g(x))^2 w(x) f(x)\,\mathrm{d}x\right). \tag{6}$$

If g is the hazard function, *MISE* could be nonexisting, and we can work with the function

$$MISE^*(h) = E\int\left[(r_h(x) - r(x))\frac{1 - F_h(x)}{1 - F(x)}\right]^2 w(x) f(x)\,\mathrm{d}x \tag{7}$$

(Vieu, 1991). This author proved that such measures are asymptotically equivalent for the three functions considered here. As usual, w is a bounded and compactly supported weight function.

In density and hazard estimation, *MISE* (or *MISE**) can be written as

$$MISE(h) = C_1(nh)^{-1} + C_2(h^{2k}) + \mathrm{o}(MISE(h)), \tag{8}$$

where C_1 and C_2 are constants. Minimizing the two first terms of this function, we obtain that the asymptotically optimal bandwidth has the form

$$h_{AMISE} = C(K, g, k)n^{-1/(2k+1)}, \tag{9}$$

with $C(K, g, k)$ a constant depending on the kernel K and the unknown function g, and being k the number of continuous derivatives of the density f.