The Annals of Statistics
1985, Vol. 13, No. 2, 818–820

A NOTE ON DATA-ADAPTIVE KERNEL ESTIMATION OF THE HAZARD AND DENSITY FUNCTION IN THE RANDOM CENSORSHIP SITUATION

BY HELMUT SCHÄFER

University of Heidelberg

In a recent paper, Tanner (1983) proves pointwise consistency of a variable bandwidth kernel estimator for the hazard function. In the present note, a simplified proof of uniform consistency of a data-adaptive kernel estimator with certain additional advantages is given.

1. Introduction. Let T_i, $i \in \mathbb{N}$, be independent positive random variables with identical distribution function F and density function f, for example lifetimes of the patients in a medical study. Assume that the T_i are censored by independent and identically distributed positive random variables C_i, $i \in \mathbb{N}$, C_i being independent of T_i for all i. Thus, one only observes $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i < C_i)$.

Recently, several authors (Ramlau-Hansen, 1983; Tanner and Wong, 1983; Tanner, 1983; Yandell, 1983) have investigated asymptotic properties of kernel estimators for the hazard function $h(x) = f(x)/(1 - F(x))$ and the density function f obtained by convolution smoothing from H_n and F_n, the empirical cumulative hazard function introduced by Nelson (1969) and Kaplan and Meier's (1958) estimator for F, respectively. Taking the example of hazard function estimation in the remainder (replace H_n by F_n systematically for density estimation), in a more general form the estimators may be written as

$$h_n(x) = \sum_{i=1}^{n} (\delta_i/N_n(X_i)) \cdot R_{n,i}^{-1} \cdot K((x - X_i)/R_{n,i})$$

with random variables $R_{n,i}$, $i \leq n$, where $N_n(x) = \sum_{i=1}^{n} I(X_i \geq x)$ denotes the number of individuals at risk at time $x - 0$ and the kernel K is a density function on $\mathbb{R}$.

Tanner (1983) proves pointwise strong consistency of a data-adaptive estimator with stochastic bandwidths $R_{n,i} = R_n$ depending on the point of interest x, defined by R_n = distance of x to its k_nth nearest neighbour among the failure points X_i, $\delta_i = 1$, $i \leq n$. These bandwidths have the disadvantage of being "biased" by the censoring distribution in the sense that they adapt to the conditional density of the random variables T_i under the condition $T_i < C_i$ of being uncensored, rather than to the density function f or the hazard function h to be estimated. Furthermore, in order to obtain a function integrating to 1 in the case of density estimation, the bandwidths of data-adaptive kernel estimators should

Received January 1984.

AMS 1980 *subject classifications*. Primary 62P10, secondary 62G05, 65D10.

Key words and phrases. Nonparametric estimation, density function, hazard function, kernel estimators, sample-point-dependent bandwidths, random censorship model.

not depend on the point of interest x, but on the sample point X_i (Breiman et al., 1977).

We therefore define

$$R_{n,i} = \inf\{r \mid H_n(X_i + r) - H_n(X_i - r) \geq p_n\}$$

with a sequence $p_n \to 0$ of positive real numbers. These bandwidths not only have the desired properties, but also simplify the proof and yield uniform consistency, because they allow straightfoward application of the convergence results for H_n (F_n in the case of density estimation) (Aalen, 1978) to derive the required asymptotic behaviour. We like to emphasize that the usual technique of integral representation and integration by parts does not apply to variable kernel estimators with bandwidths depending on the sample point. This may be one of the reasons why no attention has been paid to asymptotic properties of these estimators. In the following proof, we use Riemannian sums together with sharper asymptotic expressions for the bandwidths.

2. Uniform consistency.

THEOREM. *Choose p_n such that $p_n \log(n)/n^{1/2} \to \infty$ and define $R_{n,i}$ as above. Let $K(\cdot)$ be a bounded Riemannian integrable function with compact support on the interval* $[-1, 1]$. *Let the density functions of both the C_i and T_i be continuous everywhere, let $0 < a < b$ with $P(C_i < b) < 1$. Then* $\sup_{x \in [a,b]} | h_n(x) - h(x) | \to_{a.s.} 0$.

PROOF. We put $H_n(I) = H_n[u, o] = H_n(o) - H_n(u)$ for an interval $I = [u, o]$ for abbreviation, and $R_n(t) = \inf\{r \mid H_n[t - r, t + r] \geq p_n\}$. Throughout the proof, we suppose to deal with a sample fulfilling the convergence result of Proposition 3i of Aalen (1978) on a compact interval $[0, c]$ with $c > b$.

Let $\varepsilon > 0$. We first consider the set $S_1 = \{x \in [a, b] \mid h(x) > \varepsilon\}$. Define $r_n(x) = p_n/2h(x)$ and an interval $I_n(x) = [x - 4r_n(x), x + 4r_n(x)]$ for every $x \in S_1$. By a first application of Aalen (1978), $H_n[x - 4r_n(x), x] > p_n$ and $H_n[x, x + 4r_n(x)] > p_n$ uniformly in $x \in S_1$ for large n, which implies $R_{n,i} < | x - X_i |$ for $X_i \notin I_n(x)$ and hence

$$h_n(x) = \textstyle\sum_{X_i \in I_n(x)} (\delta_i/N_n(X_i)) R_{n,i}^{-1} K((X_i - x)/R_{n,i}).$$

Now choose a partition (t_j) of the interval $[-4, 4]$ and a positive $\delta < \varepsilon$ such that the "enlarged" Riemannian upper sum $\sum_j | t_{j+1} - t_j | \sup_{t \in [t_j - 4\delta, t_{j+1} + 4\delta]} K(t) < 1 + \varepsilon$, and consider the corresponding partition of $I_n(x)$ defined by the intervals $I_n^j(x) = [x + t_j r_n(x), x + t_{j+1} r_n(x)]$. Clearly,

$$h_n(x) \leq \textstyle\sum_j H_n(I_n^j(x)) \sup_t \dfrac{1}{R_n(t)} \sup_t k\left(j, \dfrac{r_n(x)}{R_n(t)}\right),$$

where $k(j, \alpha) = \sup_{\tau \in [\alpha t_j, \alpha t_{j+1}]} K(\tau)$ and where $\sup_t$ is taken over $t \in I_n(x)$.

Taking into account $r_n(x) < p_n/2\varepsilon$, uniform continuity of h on $[a, b]$ and Proposition 3i of Aalen (1978) yield $H_n(I_n^j(x)) \leq h(x) r_n(x) | t_{j+1} - t_j | (1 + \varepsilon)$

uniformly in $x \in S_1$ and j for large n. By the same arguments,

$$H_n[t - r_n(x)/(1 + \delta), t + r_n(x)/(1 + \delta)] < p_n$$

and

$$H_n[t - r_n(x)/(1 - \delta), t + r_n(x)/(1 - \delta)] > p_n$$

uniformly in $x \in S_1$, $t \in I_n(x)$ for large n, which, by definition of $R_n(t)$, implies $\sup_{t \in I_n(x)} |(r_n(x)/R_n(t)) - 1| < \delta$ for all $x \in S_1$. The application of these bounds in the above sum gives $h_n(x) \le h(x)(1 + \varepsilon)^3$ by construction of (t_j). The lower bound is obtained similarly.

For x from $S_2 = \{x \in [a, b] \mid h(x) \le \varepsilon\}$, define the interval $I_n(x)$ to be the smallest interval $[a_n(x), b_n(x)]$ with $H[a_n(x), x] \ge 2p_n$ and $H[x, b_n(x)] \ge 2p_n$. Let I_n denote the union of all $I_n(x)$, $x \in S_2$. Supposing $H(0) < H(a)$ and $H(b) < H(c)$ for sake of simplicity, $I_n \subset [0, c]$ and, by an argument making use of uniform continuity of h, $I_n \subset \{x \mid h(x) < 2\varepsilon\}$ for large n.

Now, repeated application of the convergence of H_n as in the first part of the proof yields

$$h_n(x) \le H_n(I_n(x))\sup_{t \in I_n}(1/R_n(t))\sup K, \quad H_n(I_n(x)) < 6p_n$$

and

$$\inf_{t \in I_n} R_n(t) > p_n/6\varepsilon$$

for large n, which completes the proof.

REFERENCES

AALEN, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *Ann. Statist.* **6** 534–545.

BREIMAN, L., MEISEL, W. and PURCELL, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19** 135–144.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

NELSON, W. (1969). Hazard plotting for incomplete failure data. *J. Qual. Techn.* **1** 27–52.

RAMLAU-HANSEN, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11** 453–466.

TANNER, M. A. and WONG, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* **11** 989–993.

TANNER, M. A. (1983). A note on the variable kernel estimator of the hazard function from randomly censored data. *Ann. Statist.* **11** 994–998.

YANDELL, B. S. (1983). Nonparametric inference for rates and densities with censored serial data. *Ann. Statist.* **11** 1119–1135.

INSTITUT FÜR MEDIZINISCHE
STATISTIK UND DOKUMENTATION
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 325
D-6900 HEIDELBERG
WEST GERMANY