

A comparison of hazard rate estimators for left truncated and right censored data

BY ÜLKÜ UZUNOĞULLARI

Faculty of Industrial Engineering, Bilkent University, Ankara, Turkey

AND JANE-LING WANG

Division of Statistics, University of California, Davis, California 95616-8705, U.S.A.

SUMMARY

Left truncation and right censoring arise frequently in practice for life data. This paper is concerned with the estimation of the hazard rate function for such data. Two types of nonparametric estimators based on kernel smoothing methods are considered. The first one is obtained by convolving a kernel with a cumulative hazard estimator. The second one is in the form of a ratio of two statistics. Local properties including consistency, asymptotic normality and mean squared error expressions are presented for both estimators. These properties facilitate locally adaptive bandwidth choice. The two types of estimators are then compared based on their theoretical and empirical performances. The effect of overlooking the truncation factor is demonstrated through the Channing House data.

Some key words: Consistency; Kernel estimator; Mean squared error; Optimal bandwidth; Weak convergence.

1. INTRODUCTION

In medical follow-up or in engineering life testing studies one may not be able to observe the variable of interest, referred to hereafter as the lifetime. Among the different forms in which incomplete data appear, right censoring and left truncation are two common ones. Left truncation may occur if the time origin of the lifetime precedes the time origin of the study. Only subjects that fail after the start of the study are being followed, otherwise they are left truncated. Those that are followed are further subject to right censoring during the follow-up period. Left truncation and right censoring may thus arise together. See for example, an AIDS study by Struthers & Farewell (1989) where the 'lifetime' is the incubation period, and the truncation variable is the time from infection until the entry to the study.

Formally, let X be the random variable of interest, called the lifetime variable, with continuous distribution function $F(x) = \text{pr}(X \leq x)$. Let T and C be the random variables for the left truncation and right censoring time respectively. It is assumed that T , C , X , are independent and, without loss of generality, that they are nonnegative. Let $Y = X \wedge C \equiv \min(X, C)$ and $\delta = I(Y = X)$, where $I(\cdot)$ is the indicator function. Under the left truncation and right censoring model one observes the triplets (Y, T, δ) only if $T \leq Y$, otherwise nothing is observed. The observations are thus taken from the conditional distribution of (Y, T, δ) given that $T \leq Y$. Note that the above reduces to the right censoring model when $\text{pr}(T = 0) = 1$ and to the left truncation model when $\text{pr}(C = +\infty) = 1$.

The nonparametric product-limit estimator of F for left truncated and right censored data was given by Tsai, Jewell & Wang (1987) and shown to be the nonparametric maximum likelihood estimator of F by Wang (1987). Our focus in the present paper is to estimate the hazard rate $\lambda(x)$ defined by $\lambda(x) = f(x)/\{1 - F(x)\}$, where f is the probability density function of F . The left truncation factor is sometimes overlooked by researchers. This will result in underestimating the hazard rate as demonstrated in § 7, Fig. 5, using the Channing House data of Hyde (1977).

The nonparametric estimation of the hazard rates was initiated by Watson & Leadbetter (1964) who proposed three types of estimators for independent and identically distributed observations. Two of these types are shown to be equivalent by Rice & Rosenblatt (1976). We extend both remaining types of estimators in § 2 to accommodate left truncated and right censored data. The first type aims at estimating both the hazard function and its derivatives, and is constructed by convolving a kernel with a cumulative hazard estimator; see (2.8). The second type aims only at the hazard function itself and is in the form of the ratio of a kernel subdensity estimator and an empirical function; see (2.12).

Local properties of both types of estimators are studied in §§ 3 and 4 including consistency, asymptotic normality and mean squared error expressions. The derivation of the mean squared error for the ratio estimator, i.e. the second one, is nonstandard for kernel estimators. The optimal bandwidth which minimizes the leading term of the mean squared error is derived for both types of estimators. In particular, the optimal bandwidth for the ratio estimator is the same as the optimal bandwidth of the kernel subdensity estimator in the numerator. The kernel estimator which employs the optimal bandwidth is called the optimal kernel estimator. It is shown in Theorems 3.3 and 4.3 that the kernel hazard estimator employing a bandwidth which is a consistent estimator of the optimal bandwidth has the same limiting distribution as the optimal kernel estimator. This is referred to as locally adaptive bandwidth choice. Choices of consistent bandwidth estimators are suggested for both types of estimators.

The choice of bandwidths is crucial for the quality of the resulting kernel curve estimator. This is even more so for hazard rate estimation than for density estimation since the variance of the kernel estimator for the former tends to infinity for large values of lifetime. The truncation and censoring scheme further complicate the situation; in particular, both the bias and variance may blow up near both tails; see equations (3.6), (4.1) and (4.2).

The two types of estimators are then compared in § 5. They have approximately the same variance but different biases. The ratio estimators usually yield very jagged curves unless a smooth empirical function is employed for the denominator. The global performance of the locally adaptive bandwidth choices is compared to that of a fixed global bandwidth and the optimal bandwidth through a simulated sample in § 6. The procedures are applied in § 7 to the Channing House data given by Hyde (1977).

Our results in § 4 extend some of the earlier results for right censored data by Ramlau-Hansen (1983), Tanner & Wong (1983), Yandell (1983) and Müller & Wang (1990). As for truncated data, except for a technical report by the authors and a Ph.D. dissertation by the first author, there are no published results on hazard function estimation to our knowledge.

2. DERIVATIONS OF THE TWO HAZARD ESTIMATORS

Refer to the left truncation and right censoring model in § 1. Let (Y_i, T_i, δ_i) ($i = 1, \dots, n$) be the independent and identically distributed random vectors which one observes, where

$Y_i = X_i \wedge C_i$, $\delta_i = I(Y_i = X_i)$, $T_i \leq Y_i$ and X_i , C_i , T_i are independent with marginal distributions F , H and G respectively. For any function K with values in $[0, 1]$ denote $\bar{K}(t) = 1 - K(t)$. Under the independence assumption, the distribution function W of Y_i satisfies $\bar{W} = \bar{H}\bar{F}$. Note that the number of observations n is a random variable. All the arguments hereafter in this paper are conditional on n .

If K is a distribution function, let $a_K = \inf \{t: K(t) > 0\}$ and $b_K = \sup \{t: K(t) < 1\}$ be the lower and upper endpoint of support of K . Then under the current model, as discussed by Woodroffe (1985), one can expect to estimate $F(x)$ only if

$$a_G \leq a_W = \min(a_F, a_H), \quad b_G \leq b_W = \min(b_F, b_H).$$

Otherwise only the conditional distribution $F_{X|\tau}(x) = F(x|X \geq \tau)$ can be estimated, for any $\tau \geq a_G$. Note, however, that $\lambda(x)$ is always identifiable for $x \geq \tau$, since $\lambda_{X|\tau}(x) = \lambda(x)$, for $x \geq \tau$. Therefore we do not need to be concerned with the identifiability problem in this paper. Assume that $\alpha = \text{pr}(T \leq Y) > 0$. The subdistribution function $W_1^*(t)$ of the uncensored observations is

$$\begin{aligned} W_1^*(y) &= \text{pr}(Y \leq y, \delta = 1 | T \leq Y) \\ &= \alpha^{-1} \left\{ \int_0^y \int_0^s G(x) dF(x) dH(s) + \bar{H}(y-) \int_0^y G(x) dF(x) \right\}. \end{aligned}$$

Hence

$$dW_1^*(y) = \alpha^{-1} G(y) \bar{H}(y-) dF(y). \quad (2.1)$$

Next, define

$$C(z) = \text{pr}(T \leq z \leq Y | T \leq Y) = \alpha^{-1} G(z) \bar{W}(z-) = \alpha^{-1} G(z) \bar{H}(z-) \bar{F}(z-). \quad (2.2)$$

From (2.1) and (2.2), it follows that

$$dW_1^*(z)/C(z) = dF(z)/\bar{F}(z-) = \lambda(z) dz, \quad (2.3)$$

and the cumulative hazard function Λ is therefore

$$\Lambda(z) = \int_0^z \lambda(t) dt = \int_0^z \frac{dW_1^*(t)}{C(t)}. \quad (2.4)$$

Let W_{1n}^* and C_n be the empirical estimators of W_1^* and C respectively. Then

$$W_{1n}^*(y) = n^{-1} \sum_{i=1}^n I(Y_i \leq y, \delta_i = 1), \quad (2.5)$$

$$C_n(z) = n^{-1} \sum_{i=1}^n I(T_i \leq z \leq Y_i). \quad (2.6)$$

Note here that $nC_n(Y_i) \geq 1$ for $i = 1, \dots, n$. Replacing W_1^* and C in (2.4) by their empirical counterparts one can estimate $\Lambda(z)$ by

$$\hat{\Lambda}_n(z) = \sum_{i=1}^n I(Y_i \leq y, \delta_i = 1) / \{nC_n(Y_i)\}. \quad (2.7)$$

This estimator (2.7) is the cumulative hazard function of the product limit estimator of F given by Tsai, Jewell & Wang (1987).

Since one of our goals is to derive a locally adaptive estimator for $\lambda(z)$ and this involves estimating higher derivatives of $\lambda(z)$, we will present the results for the r th derivative $\lambda^{(r)}(z)$ of $\lambda(z)$, for $r \geq 0$. We consider the following kernel estimator for $\lambda^{(r)}(x)$ by convolving $K_{r,b}(x) = b^{-(r+1)}K_r(x/b)$ with $\hat{\Lambda}_n$ in (2.7):

$$\hat{\lambda}^{(r)}(z) = \int K_{r,b}(z-u) d\hat{\Lambda}_n(u) = \sum_{i=1}^n \delta_i K_{r,b}(z - Y_i) / \{nC_n(Y_i)\}, \quad (2.8)$$

where K_r is the kernel and $b = b_n$ is the bandwidth. The subscript n is obviously suppressed in $\hat{\lambda}^{(r)}$ for simplicity.

We will use a kernel function with the following properties.

Property 1. We have that K_r is in $L^2([-1, 1])$ with support $[-1, 1]$ and is of bounded variation.

Property 2. For some $p \geq r$,

$$\int x^j K_r(x) dx = \begin{cases} (-1)^r r! & (j = r), \\ 0 & (0 \leq j < p, j \neq r), \\ \text{finite but nonzero} & (j = p). \end{cases}$$

Note that, for $p > 2$, Property 2 requires kernels with negative values and that K_r depends on p implicitly. Such a kernel is said to be of order (r, p) .

For the bandwidth we require the usual conditions,

$$b \rightarrow 0, \quad nb^{2r+1} \rightarrow \infty. \quad (2.9)$$

An alternative estimator for $\lambda(z)$ can be obtained via (2.3) by estimating the subdensity $w_1^*(z) = dW_1^*(z)/dz$ and $C(z)$ separately. Specifically, let $K = K_0$ be a kernel function satisfying Property 1 and Property 2, and $b = b_n$ be the corresponding bandwidth sequence. Then a kernel estimator of $w_1^*(z)$ can be obtained by convolving $K_b(x) = K_{0,b}(x) = b^{-1}K(x/b)$ with W_{1n}^* in (2.5) for a bandwidth b satisfying (2.9) with $r = 0$:

$$\hat{w}_1^*(z) = \int K_b(z-u) dW_{1n}^*(u) = n^{-1} \sum_{i=1}^n \delta_i K_b(z - Y_i). \quad (2.10)$$

To estimate $C(z)$ the empirical function $C_n(z)$ in (2.6) needs to be modified since it takes zero value if $z < T_{(1)}$ or $z > Y_{(n)}$ and there is a positive probability that it may be zero even when $T_{(1)} < z < Y_{(n)}$, where $T_{(1)}$ is the smallest order statistic of the T 's and $Y_{(n)}$ is the largest order statistic of the Y 's. There are several ways to modify C_n , for example one can replace $C_n(z)$ by $C_n(Y_i)$ for $Y_{(i)} \leq z < Y_{(i+1)}$, or one can adjust the estimator $C_n^*(z)$ of Woodroffe (1985, p. 168) to fit left truncated and right censored data. We choose the following estimator in (2.11) due to its simplicity in evaluating the associated expectations later on:

$$\hat{C}_n(z) = \begin{cases} C_n(z) & \text{if } C_n(z) > 0, \\ k_n & \text{if } C_n(z) = 0, \end{cases} \quad (2.11)$$

where $k_n \rightarrow 0$ as $n \rightarrow \infty$.

An alternative estimator for $\lambda(z)$ is thus available by combining (2.10), (2.11) and this yields

$$\tilde{\lambda}(z) = \hat{w}_1^*(z) / \hat{C}_n(z). \quad (2.12)$$

Unlike (2.8) there is no need to estimate the r th derivative of $\lambda(z)$ for locally adaptive bandwidth choices for $\tilde{\lambda}$ since the optimal bandwidth requires estimation of $w_1^{*(2)}(z)$ but not that of $\lambda^{(2)}(z)$; see (3.7).

Properties of the estimators $\tilde{\lambda}(z)$ and $\hat{\lambda}^{(r)}(z)$ will be studied in §§ 3 and 4 respectively for $a_G < z < b_W$.

3. PROPERTIES OF $\tilde{\lambda}(z)$ AND LOCALLY ADAPTIVE BANDWIDTH

We show in this section properties, including consistency and asymptotic normality of $\tilde{\lambda}(z)$ in (2.12), for z such that $a_G < z < b_W$. Notice that $0 < C(z) < 1$ on this range. These properties are then applied to obtain locally adaptive bandwidths. An estimator similar to $\tilde{\lambda}$ was studied by Blum & Susarla (1980) for its global behaviour under the right censoring model. We are mainly concerned with the local behaviour in this paper and our approach is different from theirs.

We first concentrate on estimating the r th derivative $w_1^{*(r)}$ of w_1^* . To this goal we need to assume that, for some $p \geq r$, w_1^* is p times continuously differentiable at z . Let K_r be a kernel of order (r, p) satisfying Properties 1 and 2, and b be the bandwidth sequence satisfying (2.9). Denote

$$B_{r,p} = (-1)^p (p!)^{-1} \int x^p K_r(x) dx, \quad V_{r,p} = \int K_r^2(x) dx. \quad (3.1)$$

Recall that the kernel estimator for $w_1^{*(r)}$ is

$$\hat{w}_1^{*(r)}(z) = \int K_{r,b}(z-y) dW_{1n}^*(y) = n^{-1} \sum_{i=1}^n \delta_i K_{r,b}(z - Y_i). \quad (3.2)$$

It follows from standard results on kernel density estimators based on independent and identically distributed observations that,

$$\text{bias} \{ \hat{w}_1^{*(r)}(z) \} = b^{p-r} w_1^{*(p)}(z) B_{r,p} + o(b^{p-r}), \quad p > r, \quad (3.3)$$

$$\text{var} \{ \hat{w}_1^{*(r)}(z) \} = (nb^{2r+1})^{-1} w_1^{*(r)}(z) V_{r,p} + o\{(nb^{2r+1})^{-1}\}. \quad (3.4)$$

Hence, letting

$$d = \lim_{n \rightarrow \infty} nb^{2p+1} < \infty, \quad (3.5)$$

it follows that $(nb^{2r+1})^{\frac{1}{2}} \{ \hat{w}_1^{*(r)}(z) - w_1^{*(r)}(z) \}$ converges in distribution to

$$N(d^{\frac{1}{2}} w_1^{*(p)} B_{r,p}, w_1^{*(r)}(z) V_{r,p}).$$

Next notice that $\text{pr}(C_n(z) = 0) = \{1 - C(z)\}^n = \bar{C}^n(z)$. Hence (2.11) implies that $\hat{C}_n(z) - C_n(z) = o_p(n^{-\frac{1}{2}})$ and thus $\hat{C}_n(z) - C(z) = O_p(n^{-\frac{1}{2}})$. The consistency and asymptotic normality of $\tilde{\lambda}(z)$ thus follow by setting $r = 0$.

THEOREM 3.1. (a) We have that $\tilde{\lambda}(z)$ is a consistent estimator of $\lambda(z)$.

(b) We also have that $(nb)^{\frac{1}{2}} \{ \tilde{\lambda}(z) - \lambda(z) \}$ converges in distribution to

$$N(d^{\frac{1}{2}} \{ w_1^{*(p)}(z) / C(z) \} B_{0,p}, \{ \lambda(z) / C(z) \} V_{0,p}).$$

The actual bias and variance expressions of $\tilde{\lambda}$ require some calculations which are postponed to the Appendix.

THEOREM 3.2. For $p \geq 1$, the mean squared error of $\tilde{\lambda}(z)$ is

$$\text{MSE} \{ \tilde{\lambda}(z) \} = \frac{1}{nb} \frac{\lambda(z)}{C(z)} V_{0,p} + b^{2p} \left\{ \frac{w_1^{*(p)}(z)}{C(z)} B_{0,p} \right\}^2 + o \left(\frac{1}{nb} + b^{2p} \right), \quad (3.6)$$

where the first and second terms are respectively, the leading term of the variance and the square of the bias of $\tilde{\lambda}(z)$. The optimal bandwidth which minimizes the leading terms in $\text{MSE} \{ \tilde{\lambda}(z) \}$ above is

$$b'(z) = n^{-1/(2p+1)} \left[\frac{w_1^*(z) V_{0,p}}{2p \{ w_1^{*(p)}(z) B_{0,p} \}^2} \right]^{1/(2p+1)} \equiv n^{-1/(2p+1)} \beta'(z). \quad (3.7)$$

The mean squared error expression in (3.6) is a stronger result than the usual asymptotic mean squared error dealt with in the literature for a ratio estimator. The proof of Theorem 3.2 in the Appendix can be adopted almost verbatim to improve some results in the literature. The optimal rate of the bandwidth $n^{-1/(2p+1)}$ is the same as the usual kernel density estimator for independent and identically distributed observations. Note that $b'(z)$ in (3.7) is also the optimal bandwidth for the kernel estimator \hat{w}_1^* of w_1^* and it depends on unknown quantities $w_1^*(z)$ and $w_1^{*(p)}(z)$. The next theorem shows that any consistent estimator $\hat{\beta}(z)$ of $\beta'(z)$ yields a locally adaptive bandwidth.

Let $\hat{w}_1^*(z, \beta) = \hat{w}_1^*\{z; \beta(z)\}$ and

$$\tilde{\lambda}(z; \beta) = \tilde{\lambda}\{z; \beta(z)\} \quad (3.8)$$

denote respectively the estimator (2.10) and (2.12) with bandwidth $b = n^{-1/(2p+1)} \beta(z)$. Choose β_a and β_b such that $0 < \beta_a < \beta'(z) < \beta_b < \infty$ and define the following bandwidth process for $\beta_a \leq \beta \leq \beta_b$:

$$\zeta_n(\beta) = n^{p/(2p+1)} \{ \hat{w}_1^*(z; \beta) - w_1^*(z) \}.$$

Standard arguments for locally adaptive bandwidth choice based on independent and identically distributed observations (Abramson, 1982; Krieger & Pickands, 1981) show that the process $\{\zeta_n(\beta)\}$ converges weakly on $C[\beta_a, \beta_b]$ to a Gaussian process. This then implies that

$$\zeta_n(\hat{\beta}(z)) - \zeta_n\{\beta'(z)\} = n^{p/(2p+1)} \hat{C}_n(z) [\tilde{\lambda}\{z; \hat{\beta}(z)\} - \tilde{\lambda}\{z; \beta'(z)\}]$$

tends to zero in probability provided $\hat{\beta}(z)$ tends to $\beta'(z)$ in probability. Since $\hat{C}_n \rightarrow C(z)$ in probability we obtain that

$$n^{p/(2p+1)} [\tilde{\lambda}\{z; \hat{\beta}(z)\} - \tilde{\lambda}\{z; \beta'(z)\}]$$

tends to zero in probability. Slutsky's theorem and Theorem 3.1 thus imply the following.

THEOREM 3.3. Assume $p > r$ and that the kernel K_r is Lipschitz continuous of order $\alpha > \frac{1}{2}$ on the real line. Then for any consistent estimator $\hat{\beta}(z)$ of $\beta'(z)$ the limiting distribution of

$$n^{p/(2p+1)} \{ \tilde{\lambda}(z; \hat{\beta}) - \lambda(z) \}, \quad n^{p/(2p+1)} \{ \tilde{\lambda}(z; \beta') - \lambda(z) \}$$

are the same and is the normal distribution with mean $\{\beta'(z)\}^p \{w_1^{*(p)}(z)/C(z)\} B_{0,p}$ and variance $[\beta'(z)]^{-1} [\lambda(z)/C(z)] V_{0,p}$.

Remark 1. A consistent estimator of $\beta'(z)$ can be obtained by estimating w_1^* and $w_1^{*(p)}$ consistently at z . For $r = 0$ and p , the kernel estimators in (3.2), denoted respectively by $\hat{w}_{1,0}^*(z)$ and $\hat{w}_{1,0}^{*(p)}(z)$ to indicate that they are initial estimators, are indeed consistent. It should be noted that the initial kernels for $\hat{w}_{1,0}^*$ and $\hat{w}_{1,0}^{*(p)}$ are K_0 and K_p respectively.

The initial bandwidth bd_p for $\hat{w}_{1,0}^{*(p)}$ should be larger than the initial bandwidth bd_0 for $\hat{w}_{1,0}^*(z)$, with $n bd_0 \rightarrow \infty$ and $n bd_p^{2p+1} \rightarrow \infty$.

A candidate for $\hat{\beta}(z)$ in Theorem 3.3 is thus

$$\tilde{\beta}(z) = \left[\frac{\hat{w}_{1,0}^*(z) V_{0,p}}{2p \{ \hat{w}_{1,0}^{*(p)}(z) B_{0,p} \}^2} \right]^{1/(2p+1)} \quad (3.9)$$

4. PROPERTIES OF $\hat{\lambda}^{(r)}(z)$ AND LOCALLY ADAPTIVE BANDWIDTH

In this section we present similar results as in § 3 for the estimator $\hat{\lambda}^{(r)}(z)$ in (2.8). To proceed, we first draw the attention of the readers to the analogy between the estimator $\hat{\lambda}^{(r)}$ in (2.8) and an estimator in an unpublished report of the authors for left truncated data only. Both estimators aim at $\lambda^{(r)}(z)$ and bear similar form except for different estimators of C and the indicator δ appearing in (2.8) to account for the censoring effect. This similarity allows the results under left truncation to be adapted to our estimator (2.8). Details of the proofs are therefore omitted in this section and referred to the above article.

We assume, for $p \geq r$, that λ is p times continuously differentiable at z and restrict our attention to $a_G < z < b_W$ as in § 3.

THEOREM 4.1. (a) For $p > r$,

$$\text{bias } \{\hat{\lambda}^{(r)}(z)\} = b^{p-r} \lambda^{(p)}(z) B_{r,p} + o(b^{p-r}). \quad (4.1)$$

(b) If C is continuous at z , then

$$nb^{2r+1} \text{var } \{\hat{\lambda}^{(r)}(z)\} = \{\lambda(z)/C(z)\} V_{r,p} + o(1). \quad (4.2)$$

(c) Suppose C is continuous at z , then $\hat{\lambda}^{(r)}(z)$ is a consistent estimator of $\lambda^{(r)}(z)$.

The asymptotic normality of $\hat{\lambda}^{(r)}(z)$ is established via its Hajek projection.

THEOREM 4.2. Suppose C is continuous at z and $d = \lim nb^{2p+1} < \infty$. Then

$$(nb^{2r+1})^{\frac{1}{2}} \{\hat{\lambda}^{(r)}(z) - \lambda^{(r)}(z)\} \rightarrow N(d^{\frac{1}{2}} \lambda^{(p)}(z) B_{r,p}, \{\lambda(z)/C(z)\} V_{r,p}). \quad (4.3)$$

We will now consider the local bandwidth choice. It follows from (4.1) and (4.2) that the optimal bandwidth $b^*(z)$ which minimizes the leading term of $\text{MSE } \{\hat{\lambda}^{(r)}(z)\}$ is

$$\begin{aligned} b^*(z) &= n^{1/(2p+1)} \left[\frac{2r+1}{2(p-r)} \frac{\lambda(z)}{C(z)} \frac{V_{r,p}}{\{\lambda^{(p)}(z) B_{r,p}\}^2} \right]^{1/(2p+1)} \\ &\equiv n^{-1/(2p+1)} \beta^*(z). \end{aligned} \quad (4.4)$$

The optimal bandwidth depends on the unknown quantities $\lambda(z)$, $C(z)$ and $\lambda^{(p)}(z)$. Theorem 4.3 states that any consistent estimator $\hat{\beta}(z)$ of $\beta^*(z)$ will yield a locally adaptive bandwidth. The proof is omitted. Let

$$\hat{\lambda}^{(r)}(z; \beta) = \hat{\lambda}^{(r)}\{z; \beta(z)\} \quad (4.5)$$

be the kernel estimator (2.8) with bandwidth $b = n^{-1/(2p+1)} \beta(z)$, and

$$U_n(\beta) = n^{(p-r)/(2p+1)} \{\hat{\lambda}^{(r)}(z; \beta) - \lambda^{(r)}(z)\}.$$

THEOREM 4.3. Suppose C is continuous at z , $p > r$, and the kernel K_r is Lipschitz of order $\alpha > \frac{1}{2}$ on the real line. Then for any consistent estimator $\hat{\beta}(z)$ of $\beta^*(z)$ the limiting distributions of $U_n(\hat{\beta}(z))$ and $U_n(\beta^*(z))$ are both normal with mean $\{\beta^*(z)\}^{p-r} \lambda^{(p)}(z) B_{r,p}$ and variance $\{\beta^*(z)\}^{-(2r+1)} \{\lambda(z)/C(z)\} V_{r,p}$.

Let $\hat{\lambda}_0(z)$ and $\hat{\lambda}_0^{(p)}(z)$ denote the initial kernel estimators according to (2.8) for $r=0$ and p respectively, and let $\hat{C}_n(z)$ be as given in (2.11). A candidate for $\hat{\beta}(z)$ in Theorem 4.3 is

$$\hat{\beta}(z) = \left[\frac{2r+1}{2(p-r)} \frac{\hat{\lambda}_0(z)}{\hat{C}_n(z)} \frac{V_{r,p}}{\{\hat{\lambda}_0^{(p)}(z)B_{r,p}\}^2} \right]^{1/(2p+1)} \quad (4.6)$$

5. COMPARISONS OF THE TWO TYPES OF ESTIMATORS

Let $\hat{\lambda}(z)$ denote the estimator $\hat{\lambda}^{(r)}(z)$ in (2.8) with $r=0$. We compare it in this section with the other estimator $\tilde{\lambda}(z)$ in (2.12). The comparison will first be made on the theoretical framework as laid out in §§ 3 and 4. In § 6 they will be compared on practical performance based on simulated samples. Comments similar to those in this section are also made independently by Patil, Wells & Marron (1991) for independent and identically distributed samples.

First, the variances of the two estimators are asymptotically equivalent but they differ in the biases; see (3.6) and (4.2). Since $C(z)$ tends to zero as z approaches a_G or b_w the variances of both estimators tend to infinity as z tends to a_G or b_w .

Secondly, when a fixed bandwidth is employed, the assumptions on F , G and H for properties like consistency and asymptotic normality, centred at the expected value of the estimator, are similar for both estimators. For the adaptive bandwidths in (4.6) with $p=2$, $\hat{\lambda}(z, \hat{\beta})$ requires that $\lambda^{(2)}$ be continuous at z while $\tilde{\lambda}(z, \tilde{\beta})$ with $\tilde{\beta}$ given by (3.9) requires that $w_1^{*(2)}$ be continuous at z which requires continuous second derivatives of the censoring and truncation distributions H and G .

Thirdly, the estimator $\tilde{\lambda}$ usually yields a curve more erratic than that of $\hat{\lambda}$; see, for example, Fig. 1. This is due to the fact that the denominator \hat{C}_n in (2.12) is unsmooth and this can be corrected somewhat by replacing it by an appropriate smooth version of it. Theorem 3.1 remains true for the smooth ratio estimator although the bias and variance expressions of the smooth ratio estimator are difficult to handle.

Fourthly, when adaptive bandwidths are employed, the initial bandwidths have to be chosen with some care. In reality the initial bandwidths can be determined by a global procedure if such a procedure is available or subjectively by inspecting the range and spread of the data and the location of the point z . Once a bandwidth is chosen one can

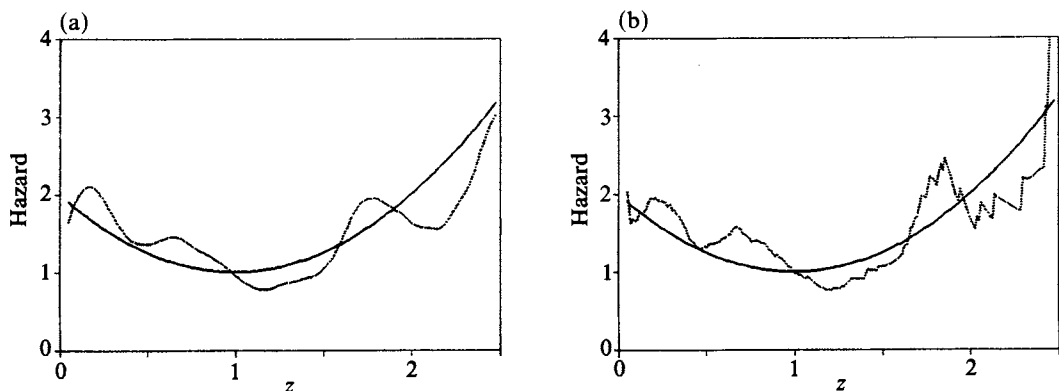


Fig. 1. (a) Dotted line, estimator $\hat{\lambda}(z)$ in (2.8) with fixed bandwidth $b=0.3$; solid line, true curve. (b) Dotted line, estimator $\tilde{\lambda}(z)$ in (2.12) with fixed bandwidth $b=0.3$; solid line, true curve.

either use it as the initial bandwidth for the adaptive procedure or iterate the adaptive procedures a few times, starting with this bandwidth, to produce another initial bandwidth. As mentioned earlier, the initial bandwidths for $\lambda^{(2)}$ and $w_1^{*(2)}$ should be larger than those for λ and w_1^* .

Fifthly, the locally adaptive estimators when applied over a range of z 's usually yield spikes at several points; see Fig. 2(a). This can be corrected as suggested by Müller (1988, p. 120) and Schucany (1989). Further discussion is given in the next section.

6. A SIMULATED EXAMPLE

We demonstrate in this section how to apply the adaptive procedures of §§ 3 and 4 in practice and compare their global performance through simulated samples.

The lifetime variable is simulated from a distribution F with hazard rate $\lambda(z) = (z-1)^2 + 1$. This failure rate is chosen so that it has a bowl shape which would resemble most practical life distributions. The constant 1 is added to avoid λ being zero at $z=1$ which distorts the optimal bandwidth expression for $\hat{\lambda}(z)$. Note that the bowl shape of this hazard function λ makes it harder to estimate near both the left and right boundaries as opposed to a hazard function which tails off to zero at either boundary.

Both the censoring and truncation distributions are simulated from exponential distributions with means 4 and 0.1 respectively. A random sample (X_i, C_i, T_i) of size 200 is simulated independently. This resulted in a censoring proportion of 14.8% for the Y data and a truncation proportion of 17% for the (Y, T) data. The observed sample size n is equal to 166. For estimating $\lambda(z)$ and $w_1^*(z)$, the kernel

$$K_0(x) = \frac{15}{16}(1 - 2x^2 + x^4),$$

of order $r=0$, $p=2$ is used. To estimate $\lambda^{(2)}(z)$ and $w_1^{*(2)}(z)$ the kernel,

$$K_2(x) = \frac{315}{32}(-1 + 9x^2 - 15x^4 + 7x^6),$$

of order $r=2$, $p=4$ is employed. Both kernels are chosen with reference to Müller (1988, p. 68) as they have certain optimality properties. Note also that both K_0 and K_2 above are Lipschitz continuous of order one on $(-\infty, \infty)$, hence are appropriate for the adaptive bandwidth choice.

A global fixed bandwidth $bd_0 = 0.3$ is chosen for both $\hat{\lambda}(z)$ and $\hat{\lambda}^{(2)}(z)$ for all $0 < z < 2.5$. Note that $\text{pr}(X \geq 2.5) = 0.003$. The resulting estimators are given in Fig. 1(a) for $\hat{\lambda}$ and in Fig. 1(b) for $\hat{\lambda}^{(2)}$. Notice that $\hat{\lambda}$ is much smoother than $\hat{\lambda}^{(2)}$ as discussed in § 5. There is a slight left boundary effect as the estimated curve drops off near zero.

For the adaptive bandwidth estimator we use $\hat{\lambda}(z, \hat{\beta})$ with $\hat{\beta}$ given in (4.6). Global initial bandwidths $bd_0 = 0.3$ and $bd_2 = 0.7$ are used for $\hat{\lambda}(z)$ and $\hat{\lambda}^{(2)}(z)$ respectively. The resulting adaptive estimator and adaptive bandwidth b are shown in Fig. 2(a) and Fig. 2(b) separately. Note that the spikes in Fig. 2(a) are basically due to the corresponding spikes of the adaptive bandwidth in Fig. 2(b). To correct this undesirable feature we truncate the $\hat{\beta}$ below at $b_* = 0.5 bd_0 = 0.15$ and above at $b^* = 2 bd_0 = 0.6$; that is $\hat{\beta}$ is replaced by the closer one of b_* or b^* should it fall outside the region (b_*, b^*) . The resulting truncated adaptive bandwidth estimator is plotted in Fig. 3(a). This corrects the spikes but not the left boundary problem. Further research on boundary modification is needed; see, for example, Rice (1984). The theoretical optimal bandwidth estimator is plotted in Fig. 3(b) for the purpose of comparison. The optimal kernel estimator has

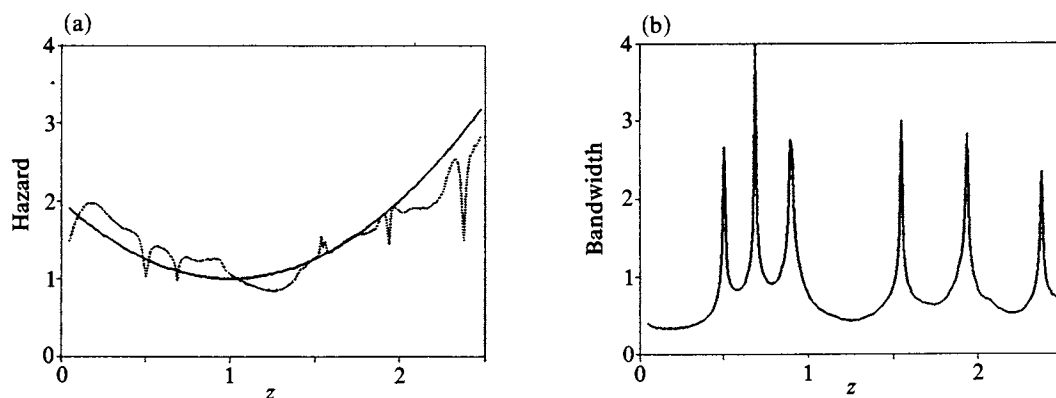


Fig. 2. (a) Dotted line, adaptive estimator $\hat{\lambda}(z, \hat{\beta})$ in (4.5) with $\hat{\beta}$ in (4.6); initial bandwidths, $bd_0 = 0.3$, $bd_2 = 0.7$; solid line, true curve. (b) Adaptive bandwidth \hat{b} corresponding to Fig. 2(a).

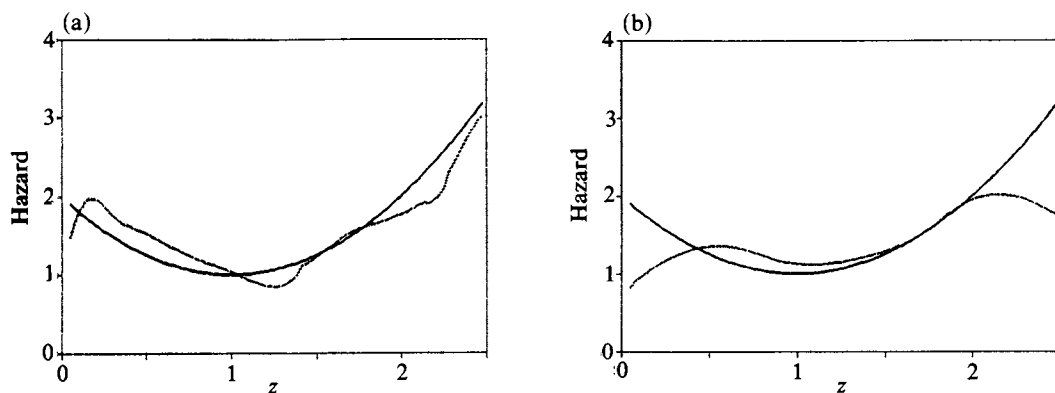


Fig. 3. (a) Dotted line, adaptive estimator in Fig. 2(a) with adaptive bandwidth truncated below at 0.15 and above at 0.6; solid line, true curve. (b) Dotted line, theoretical optimal bandwidth estimator with $b^*(z)$ in (4.4); solid line, true curve.

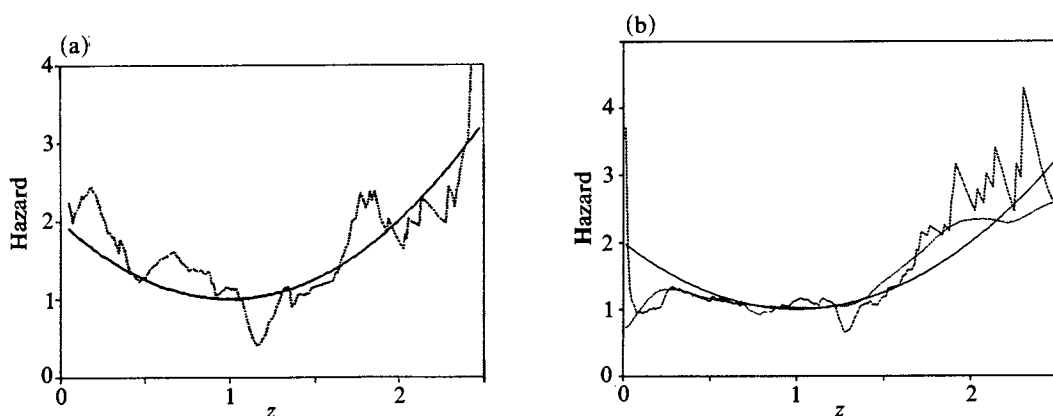


Fig. 4. (a) Dotted line, adaptive estimator $\hat{\lambda}(z, \tilde{\beta})$ in (3.8) with $\tilde{\beta}$ in (3.9); initial bandwidths, $bd_0 = 0.3$, $bd_2 = 0.7$, with adaptive bandwidth truncated below at 0.15 and above at 0.6; solid line, true curve. (b) A second simulated example of $\hat{\lambda}(z, \hat{\beta})$ as in Fig. 3(a) and $\tilde{\lambda}(z, \tilde{\beta})$ as in Fig. 4(a); solid line, true curve; light dotted line, $\hat{\lambda}$; dashed line, $\tilde{\lambda}$.

a serious boundary problem on both boundaries due to the large optimal bandwidths near both tails. Reexamining Figs 1(a), 3(a) and 3(b) there seems to be some benefit in applying the modified adaptive procedures over the fixed bandwidth procedure and they compare favourably to the theoretical optimal procedure.

The above procedures are also applied to the alternative estimator $\tilde{\lambda}(z, \tilde{\beta})$ where $\tilde{\beta}$ is given in (3.9) and the adaptive bandwidth \tilde{b} is truncated at $b_* = 0.15$ and $b^* = 0.6$. The resulting estimator is shown in Fig. 4(a). It is noteworthy that while the local bandwidth selection improves $\hat{\lambda}$ appreciably as compared to a global one, it does not seem to benefit $\tilde{\lambda}$. This phenomenon is also reflected in a second simulated sample which resulted in 16% censoring and 22% truncation. We report, in Fig. 4(b), only the locally adaptive procedures for $\hat{\lambda}(z, \hat{\beta})$ and $\tilde{\lambda}(z, \tilde{\beta})$ for this sample.

To summarize, the estimator $\hat{\lambda}$ in (2.8) seems to perform better than the alternative estimator $\tilde{\lambda}$ for this particular model. Referring to the first comment in § 5, this may be attributed to the differences in the bias terms for this particular model.

7. AN EXAMPLE

In this section the procedures in §§ 2–4 are applied to the Channing House data of Hyde (1977). The observations consist of survival data for male members of the Channing House retirement community in Palo Alto, California. The variable of interest, X , is the lifetime in months of the male retired people. The individuals are observed only if they survive long enough to participate in the community. Once they come under observation, the participants are also subject to censoring. The left truncation and right censoring model therefore applies. Here the truncation time T is age in months at entry into study plus one, denoted by $\nu + 1$ by Hyde (1977), and $Y = X \wedge C$ is the age in months when last seen in the study, denoted by λ^* there. There are a total of 97 observations. Of them, 46 have died, 46 survived to the closing date of the study and 5 withdrew from the community.

The usual identifiability assumption that $a_G \leq a_W$ may not be reasonable here but, as mentioned earlier, this does not matter for hazard rate estimation. Due to the boundary problems and since there are only two Y 's, 777 and 781, below 840 and three of them above 1125 we report only the results on the interval $[840, 1125]$ although all the data were used in the estimation procedures. The kernels K_0 and K_2 are the same as in § 5. A fixed bandwidth $b = 50$ is chosen for both $\hat{\lambda}(z)$ in (2.8) and $\tilde{\lambda}(z)$ in (2.12) for all z in $[840, 1125]$. The results are given in Fig. 5(a). There is an increasing trend in the hazard function which is expected. This data set has also been analyzed in the past as randomly censored data only. To see how the estimates would differ had the truncation factor been ignored, the same sample is analyzed by letting the truncation variable $T \equiv 0$ and the resulting $\hat{\lambda}(z)$ is displayed in Fig. 5(a) with the thinner solid line. It is clear that the hazard rate is underestimated uniformly when the truncation effect is ignored and the discrepancy is larger for smaller z 's due to the left truncation scheme. This emphasizes the importance of accounting for the truncation effect. Otherwise the result can be optimistically misleading. As for the adaptive bandwidth estimators $\hat{\lambda}(z, \hat{\beta})$ with $\hat{\beta}$ in (4.6) and $\tilde{\lambda}(z, \tilde{\beta})$ with $\tilde{\beta}$ in (3.9), initial global bandwidths $bd_0 = 50$, $bd_2 = 100$ are used for $\hat{\lambda}(z)$ and $\hat{\lambda}^{(2)}(z)$ with adaptive bandwidths truncated to 70 if it exceeds 70. The resulting truncated adaptive bandwidth estimators are displayed in Fig. 5(b), where the increasing trend in hazard rate is even more visible now than in Fig. 5(a). The hazard rate increases rapidly in the early ages below 900 months, i.e. 75 years old, then stabilizes

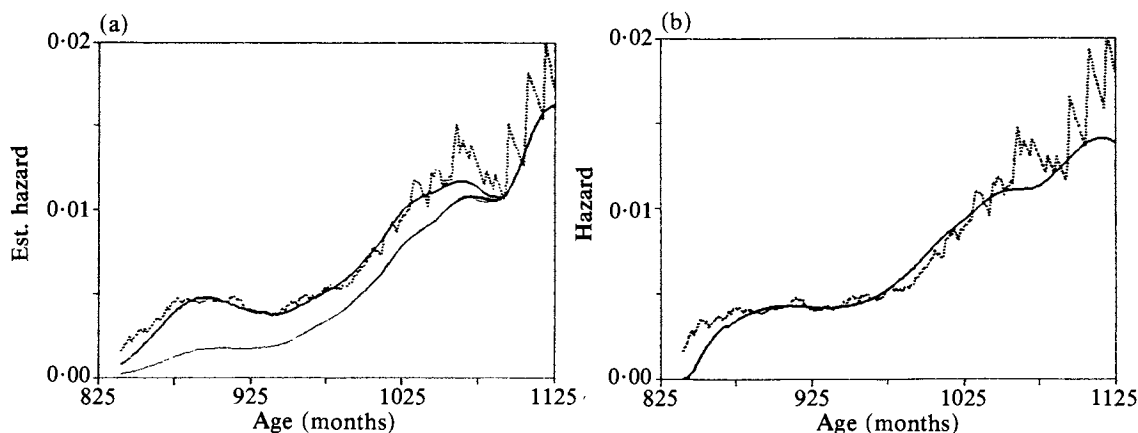


Fig. 5. (a) Estimated hazard rate for Channing House data; fixed bandwidth $b = 50$; darker solid line, $\hat{\lambda}$; thinner solid line, $\hat{\lambda}$ with the truncation effect neglected; dotted line, $\tilde{\lambda}$. (b) Adaptive estimators for Channing House data; initial bandwidth $bd_0 = 50$, $bd_2 = 100$ with adaptive bandwidths truncated above at 70; solid line, $\hat{\lambda}$; dotted line, $\tilde{\lambda}$.

between the months 900–960, i.e. ages 75–80, and starts to rise steadily again after age 80. The slight drop near 1120 months, or 93 years, is probably due to boundary effects.

ACKNOWLEDGEMENTS

Research supported in part by an U.S. Air Force Grant. The work of Ülkü Uzunoğullari was done while she was a Ph.D. student at the Department of Statistics, the Wharton School of the University of Pennsylvania. The authors would like to thank the referees for valuable suggestions.

APPENDIX

Proof of Theorem 3.2

To establish the bias and variance expression in Theorem 3.2 we need the following lemma.

LEMMA A.1. *We have:*

- (a) $E\{\hat{C}_n(z)\} = C(z) + k_n \bar{C}^n(z)$;
- (b) $\text{var}\{\hat{C}_n(z)\} = n^{-1} C(z) \bar{C}(z) + k_n^2 \bar{C}^n(z) \{1 - \bar{C}^n(z) - 2k_n^{-1} C(z)\}$;
- (c) $E\{\hat{C}_n(z)\}^{-k} = O(1)$, for any integer $k \geq 1$;
- (d) $E\{\{\bar{C}_n(z) - C(z)\}^4\} = O(n^{-2})$;
- (e) $E\{\{\hat{C}_n(z) - C(z)\}^8\} = O(n^{-4})$.

Proof. Express $n\hat{C}_n(z) = BI(B > 0) + nk_n I(B = 0)$, where $B \sim \text{Binomial}(n, C(z))$. Parts (a) and (b) follow from direct calculations. Part (c) follows from $\bar{C}(z) < 1$ and

$$E\{\{\hat{C}_n(z)\}^{-k}\} \leq (n+1)^k \bar{C}^n(z) + \{C(z)\}^{-k} (k+1)!.$$

Note that (2.11) implies that

$$E\{\{\hat{C}_n(z) - C_n(z)\}^k\} = O\{\bar{C}^n(z)\},$$

for any integer $k \geq 1$. Therefore it suffices to prove

$$E\{\{C_n(z) - C(z)\}^4\} = O(n^{-2})$$

in part (d). Notice that $C_n(z) - C(z) = n^{-1} \sum \eta_i$, where η_i are independent and identically distributed mean zero random variables with values between 0 and 1. Therefore when taking expectation of $(\sum \eta_i)^4$ only the terms $E(\eta_i^2 \eta_j^2)$ and $E(\eta_i^4)$ are nonzero. Thus we have

$$E[\{C_n(z) - C(z)\}^4] = n^{-4} \left[nE(\eta_1^4) + \binom{n}{2} \{E(\eta_1^2)\}^2 \right] = O(n^{-2}).$$

Part (e) follows by a similar combinatorial argument as in part (d) that

$$E[\{C_n(z) - C(z)\}^8] = O(n^{-4}). \quad \square$$

Proof of Theorem 3.2. Consider

$$\begin{aligned} \tilde{\lambda}(z) - \lambda(z) &= \hat{w}_1^*(z) \{ \hat{C}_n(z) \}^{-1} - \{ C(z) \}^{-1} + \{ C(z) \}^{-1} [\hat{w}_1^*(z) - E\{ \hat{w}_1^*(z) \}] \\ &\quad + \{ C(z) \}^{-1} [E\{ \hat{w}_1^*(z) \} - w_1^*(z)] \\ &= \text{I} + \text{II} + \text{III}. \end{aligned} \quad (\text{A} \cdot 1)$$

By the Cauchy-Schwarz inequality, we have

$$\{E(\text{I})\}^2 \leq E[\{w_1^*(z)\}^2] E \left[\left\{ \frac{\hat{C}_n(z) - C(z)}{\hat{C}_n(z) C(z)} \right\}^2 \right] = A \times B. \quad (\text{A} \cdot 2)$$

Let $r=0$ and use the bias and variance expressions for $\hat{w}_1^{*(r)}(z)$ in (3.3) and (3.4) to show that $A \leq M$ for some constant M . Applying the Cauchy-Schwarz inequality to the second term B in (A.2) we obtain, for Lemma A.1(c), (d), that $B = O(n^{-1})$. We have thus shown that $E(\text{I}) = O(n^{-1/2})$. The bias part in (3.6) now follows from (3.3) and (A.1).

Using similar arguments as above and Lemma A.1(e) with some calculations it can be shown that $E(\text{I}^2) = O(n^{-1})$. Hence $\text{var}(\text{I}) = O(n^{-1}) = o((nb)^{-1})$. The variance part in (3.6) now follows from Lemma A.1(c) and (3.4), applying the Cauchy-Schwarz inequality to $\text{cov}(\text{I}, \text{II})$ and replacing w_1^*/C by λ . \square

REFERENCES

- ABRAMSON, I. (1982). Arbitrariness of the pilot estimator in adaptive kernel methods. *J. Mult. Anal.* **12**, 562-7.
- BLUM, J. R. & SUSARLA, V. (1980). Maximal deviation theory of density and failure rate function estimates based on censored data. In *Multivariate Analysis—V*, Ed. P. R. Krishnaiah, pp. 213-22. Amsterdam: North-Holland.
- HYDE, J. (1977). Testing survival under right censoring and left truncation. *Biometrika* **64**, 225-30.
- KRIEGER, A. M. & PICKANDS, J. III (1981). Weak convergence and efficient density estimation at a point. *Ann. Statist.* **9**, 1066-78.
- MÜLLER, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data, Lecture Notes in Statistics*, **46**. Berlin: Springer-Verlag.
- MÜLLER, H. G. & WANG, J. L. (1990). Locally adaptive hazard smoothing. *Prob. Theory Rel. Fields* **85**, 523-38.
- PATIL, P. N., WELLS, M. T. & MARRON, J. S. (1992). Kernel based estimation of ratio function. *J. Nonpara. Statist.* To appear.
- RAMLAU-HANSEN, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11**, 452-66.
- RICE, J. (1984). Boundary modification for kernel regression. *Comm. Statist. A* **13**, 893-900.
- RICE, J. & ROSENBLATT, M. (1976). Estimation of the log survivor function and hazard function. *Sankhyā A* **38**, 60-78.
- SCHUCANY, W. R. (1989). Locally optimal window widths for kernel density estimation with large samples. *Statist. Prob. Letters* **7**, 401-5.
- STRUTHERS, C. A. & FAREWELL, V. T. (1989). A mixture model for times to AIDS data with left truncation and an uncertain origin. *Biometrika* **76**, 814-7.
- TANNER, M. A. & WONG, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* **11**, 989-93.

- TSAI, W. Y., JEWELL, N. P. & WANG, M. C. (1987). A note on the product limit estimator under right censoring and left truncation. *Biometrika* **74**, 883–6.
- WANG, M. C. (1987). Product limit estimates: A generalized maximum likelihood study. *Comm. Statist. A* **16**, 3117–32.
- WATSON, G. S. & LEADBETTER, M. R. (1964). Hazard analysis, I. *Biometrika* **51**, 175–84.
- WOODROOFE, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163–77.
- YANDELL, S. B. (1983). Nonparametric inference for rates with censored survival data. *Ann. Statist.* **11**, 1119–35.

[Received March 1991. Revised August 1991]