The Annals of Statistics
1995, Vol. 23, No. 5, 1735–1748

KERNEL ESTIMATION IN A NONPARAMETRIC MARKER DEPENDENT HAZARD MODEL[1]

BY JENS P. NIELSEN AND OLIVER B. LINTON

PFA Pension and Yale University

We introduce a new kernel hazard estimator in a nonparametric model where the stochastic hazard depends on the current value of time and on the current value of a time dependent covariate or marker. We establish the pointwise and global convergence of our estimator.

1. Introduction. Let $Z(t)$ be a d-dimensional time dependent covariate or marker process, and let $\lambda(t)$ be the stochastic hazard for an individual with history $\{Z(s);\ s \leq t\}$. Jewell and Nielsen (1993) discuss the distinction between covariate and marker. For our purposes this distinction is unimportant. We examine the following model:

$$\lambda(t) = \alpha\{Z(t), t\}Y(t), \tag{1}$$

where $Y(t)$ is an indicator of survival at time t. Submodels of (1) have a long tradition in survival analysis, for example: the Cox regression model [see Cox (1972) and Andersen and Gill (1982)]; the multiplicative model proposed by Thomas (1983) and studied by O'Sullivan (1993); and Aalen's additive risk model [see Aalen (1989) and McKeague and Utikal (1991)]. An important special case is where the intensity depends only on the marker process, that is,

$$\lambda(t) = \alpha\{Z(t)\}Y(t). \tag{2}$$

We call (2) the marker-only model. This occurs in medical applications where the exposure time is not known precisely [see Fusaro, Nielsen and Scheike (1993) for an example]. The special case of (1) where there are no covariates,

$$\lambda(t) = h(t)Y(t), \tag{3}$$

has been well treated [see Ramlau-Hansen (1983), for results on pointwise and global convergence of nonparametric kernel estimation, and Nielsen (1990) for an analysis of the plug-in and the cross-validation method of bandwidth selection in this setting].

Beran (1981) provided the first results on nonparametric inference in (1): he suggested a class of estimators for the cumulative hazard function (conditional on Z) in the special case where Z is independent of time. Dabrowska (1987) established weak convergence results for Beran's estimators employing a conditional version of the classical approach by Breslow and Crowley

Received July 1991; revised February 1995.
[1]Supported by the NSF.
AMS 1991 *subject classifications*. 62G05, 62M09.
Key words and phrases. Counting process, hazard function, kernel estimation, nonparametric estimation, survival analysis.

(1974). McKeague and Utikal (1990) (henceforth MU) analyze the more general situation in which the covariate Z is allowed to be time dependent. They also introduce two estimators of the underlying hazard α based on smoothing of the conditional cumulative hazard estimator. Sufficient conditions for global convergence for one estimator are presented.

We make several contributions. First, we introduce an alternative kernel estimator of α in (1) within the general counting process framework of MU. Our procedure is analogous to the Nadaraya–Watson regression estimator in construction. It also works in the special case (2) where time is not observed. Second, we obtain expressions for the asymptotic bias of our estimator which MU did not do; thus we obtain the optimal rate of convergence. Third, our conditions for establishing global convergence are strictly weaker than MU's. This is a consequence of the difference in proof technique, rather than estimator. McKeague and Utikal (1990) use the technique developed by Ramlau-Hansen (1983): the upper bound for the global error obtained using this technique is rather crude and increases rapidly with d; MU only consider the case $d = 1$, so the effect of the crude approximation is not too serious in their case. We consider general d and establish global convergence by verifying the conditions of Bickel and Wichura (1971). To do this we introduce a simple sufficient condition for tightness that can be of use elsewhere. Finally, our estimator is well motivated: when a uniform kernel is used, it reduces to occurrence over exposure. It is easy to compute and was employed by Fusaro, Nielsen and Scheike (1993) to estimate the risk of AIDS given current marker status based on the San Francisco Men's Health Study.

We recognize that the performance of nonparametric estimators is poor in high dimensions due to the slow rate of convergence; nevertheless these procedures can be used as inputs to some more general modelling process, for example, structured nonparametric models [see Hastie and Tibshirani (1990)] and semiparametric models [see Andersen, Borgan, Gill and Keiding (1992)], for which faster convergence rates are possible. For these reasons it is important to have a general theory.

In Section 2 we formulate the sampling scheme, while in Section 3 we define our estimator. Section 4.1 contains the pointwise properties, and Section 4.2 has the global convergence result.

We use $|\cdot|$ to denote the Euclidean norm of a vector; we use $\to_{\mathcal{P}}$ to denote convergence in probability; and we use $\Rightarrow$ to signify weak convergence. For a square integrable martingale M, let $\langle M \rangle$ denote its quadratic variation process. Throughout, I_A denotes the indicator function of the event A. Inference will be conducted on the unit cube and, unless otherwise stated, all integrations are over this set and are omitted in the sequel along with surplus superscripts.

2. A counting process formulation of the model. We observe n individuals $i = 1, \ldots, n$. Let $N_i^{(n)}$ count observed failures for the ith individual in the time interval $[0, 1]$. We assume that $\mathbf{N}^{(n)} = (N_1^{(n)}, \ldots, N_n^{(n)})$ is an n-dimensional counting process with respect to an increasing, right continu-

ous, complete filtration $\mathcal{F}_t^{(n)}$, $t \in [0, 1]$, that is, one that obeys *les conditions habituelles* [see Andersen, Borgen, Gill and Keiding (1992), page 60]. We model the random intensity process $\lambda^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_n^{(n)})$ of $\mathbf{N}^{(n)}$ as depending on marker values:

$$\lambda_i^{(n)}(t) = \alpha\{Z_i^{(n)}(t), t\}Y_i^{(n)}(t), \tag{4}$$

but do not restrict the functional form of $\alpha(\cdot)$. Here, Y_i is a predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the ith individual is under risk, while $Z_i = (Z_{i1}, \ldots, Z_{id})$ is a d-dimensional, predictable, CADLAG, covariate or marker process. Let $F(z, s) = \Pr(Z_i(s) \leq z \mid Y_i(s) = 1)$ be the conditional distribution function of the covariate process at time s and $f(z, s)$ the corresponding density with respect to the d-dimensional Lebesgue measure. We will assume that the covariate process is supported on the unit cube and that $E\{Y_i(s)\} = y(s)$, where $y(\cdot)$ is continuous. The marker $Z_i(s)$ is only observed for those s such that $Y_i(s) = 1$. Let

$$Z_i^*(s) = \begin{cases} Z_i(s), & \text{when } Y_i(s) = 1, \\ -\infty, & \text{when } Y_i(s) = 0. \end{cases}$$

We call Z_i^* the observed marker process. We assume that the stochastic processes $(N_1, Z_1^*, Y_1), \ldots, (N_n, Z_n^*, Y_n)$ are iid for the n individuals, and take $\mathcal{F}_t = \sigma(\mathbf{N}(s), \mathbf{Z}(s), \mathbf{Y}(s);\ s \leq t)$, where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$. In the sequel, all martingales and predictable processes are defined with respect to this filtration. With these definitions, λ is predictable and the processes $M_i(t) = N_i(t) - \Lambda_i(t)$, $i = 1, \ldots, n$, with compensators $\Lambda_i(t) = \int_0^t \lambda_i(s)\, ds$, are square integrable local martingales on the time interval $[0, 1]$.

3. Definition of the estimators. In this section we define estimators for α in (4); a corresponding estimator in the marker-only model (2) follows by analogy. We relate our estimator to the competing kernel estimators for h in the model (3) suggested by Ramlau-Hansen (1983) and Hjort (1994).

Let k be a one-dimensional probability density function, and, for $b \neq 0$, let $k_b(\cdot) = b^{-1}k(\cdot/b)$. Ramlau-Hansen (1983) estimated $h(t)$ in model (3) by

$$\tilde{h}(t) = \sum_{i=1}^{n} \int k_b(t - s) \frac{1}{Y(s)}\, dN_i(s), \tag{5}$$

where $Y(s) = \sum_{i=1}^n Y_i(s)$. In fact, (5) results from kernel smoothing of Aalen's estimator for the cumulative hazard [see Aalen (1978)]. Working from the local likelihood principle applied to a constant hazard function, Hjort (1994) suggested the following estimator for h:

$$\hat{h}(t) = \frac{\sum_{i=1}^n \int k_b(t - s)\, dN_i(s)}{\sum_{i=1}^n \int k_b(t - s) Y_i(s)\, ds}. \tag{6}$$

When a uniform kernel is used, Hjort's estimator reduces to number of failures divided by exposure time in a neighborhood of the considered time

value. We note here the analogy with regression. Estimator (5) is an internal estimator in the language of Jones, Davies and Park (1994), while (6) is an external estimator.

Return now to the general model (4). We use, for simplicity, product kernels and a single bandwidth throughout, although in practice one should take account of the differences in scale for each direction. Let $K(u) = \prod_{j=1}^{d} k(u_j)$, where $u = (u_1, \ldots, u_d)$, and write $K_b(\cdot) = \prod_{j=1}^{d} k_b(\cdot)$. We estimate α by the external estimator

$$\hat{\alpha}(z, t) = \frac{\sum_{i=1}^{n} \int K_b\{z - Z_i(s)\} k_b(t - s)\, dN_i(s)}{\sum_{i=1}^{n} \int K_b\{z - Z_i(s)\} k_b(t - s) Y_i(s)\, ds}. \tag{7}$$

We believe that $\hat{\alpha}(z, t)$ is the natural extension of Hjort's estimator to model (4). An alternative is the internal estimator

$$\tilde{\alpha}(z, t) = \sum_{i=1}^{n} \int \frac{K_b\{z - Z_i(s)\} k_b(t - s)}{\sum_{i=1}^{n} K_b\{z - Z_i(s)\} Y_i(s)}\, dN_i(s), \tag{8}$$

which appears to be the natural extension of (5). In fact, both Keiding, Holst and Green (1989) and MU's estimators are of this form, except that MU's $\tilde{\alpha}$, for example, uses different bandwidths in numerator and denominator and also only a uniform kernel in the denominator. We do not present any theory for $\tilde{\alpha}(z, t)$.

Finally, in (2) we would estimate α by

$$\hat{\alpha}(z) = \frac{\sum_{i=1}^{n} \int K_b\{z - Z_i(s)\}\, dN_i(s)}{\sum_{i=1}^{n} \int K_b\{z - Z_i(s)\} Y_i(s)\, ds}. \tag{9}$$

It is important to note here that (9) can be defined even when exposure time itself is not observed.

4. Asymptotic properties of the estimators. We first state two useful results. Let $g_1^{(n)}, \ldots, g_n^{(n)}$ be predictable stochastic processes. The following version of the central limit theorem for martingales follows as an immediate extension of Proposition 4.2.1 of Ramlau-Hansen (1983).

PROPOSITION 1. *Suppose that, as* $n \to \infty$, *the following hold*:

(P1) $\sum_{i=1}^{n} \int \{g_i^{(n)}(s)\}^2\, d\langle M_i \rangle(s) \to_{\mathscr{P}} \sigma^2$;

(P2) $\forall \varepsilon > 0$, $\sum_{i=1}^{n} \int \{g_i^{(n)}(s)\}^2 I_{\{|g_i^{(n)}(s)| > \varepsilon\}}\, d\langle M_i \rangle(s) \to_{\mathscr{P}} 0$.

Then

$$\sum_{i=1}^{n} \int g_i^{(n)}(s)\, dM_i(s) \Rightarrow N(0, \sigma^2).$$

In the proof of Theorem 1 we also use the following fact. Let $g_n(x)$ be a sequence of real-valued functions, where $x \in \mathbb{R}^{d+1}$, and let $X_{ni} = \int g_n\{Z_i(s), s\} Y_i(s)\, ds$. Then

$$\operatorname{Var}\left[n^{-1} \sum_{i=1}^{n} X_{ni} \right] \le n^{-1} \iint g_n^2(w, s) f(w, s) y(s)\, dw\, ds. \tag{10}$$

For $i \in \{1, \ldots, n\}$,

$$\left[\int g_n\{Z_i(s), s\}Y_i(s)\,ds\right]^2 \leq \int g_n^2\{Z_i(s), s\}Y_i(s)\,ds,$$

by the Cauchy–Schwarz inequality. Furthermore,

$$E\left[\int g_n^2\{Z_i(s), s\}Y_i(s)\,ds\right] = \iint g_n^2(w, s)f(w, s)y(s)\,dw\,ds,$$

by Tonelli's theorem, and (10) follows.

4.1. *Pointwise theory for $\hat{\alpha}$ at interior points.* Let $x = (z, t)$ be an interior point of $[0,1]^{d+1}$, that is, there is a neighborhood $\mathcal{N} = [x - \varepsilon, x + \varepsilon] \subset (0, 1)^{d+1}$, where ε is a $(d + 1)$-vector with each component positive. Let

$$\alpha^*(x) = \frac{\sum_{i=1}^n \int K_b\{z - Z_i(s)\}k_b(t - s)\,\alpha\{Z_i(s), s\}Y_i(s)\,ds}{\sum_{i=1}^n \int K_b\{z - Z_i(s)\}k_b(t - s)Y_i(s)\,ds},$$

and write

$$(\hat{\alpha} - \alpha)(x) = (\hat{\alpha} - \alpha^*)(x) + (\alpha^* - \alpha)(x) = \frac{\mathscr{V}_x + \mathscr{B}_x}{\mathscr{E}_x},$$

where

$$\mathscr{E}_x = n^{-1}\sum_{i=1}^n \int K_b\{z - Z_i(s)\}k_b(t - s)Y_i(s)\,ds$$

$$\mathscr{V}_x = n^{-1}\sum_{i=1}^n \int K_b\{z - Z_i(s)\}k_b(t - s)\,dM_i(s)$$

$$\mathscr{B}_x = n^{-1}\sum_{i=1}^n \int K_b\{z - Z_i(s)\}k_b(t - s)\big[\alpha\{Z_i(s), s\} - \alpha(z, t)\big]Y_i(s)\,ds.$$

We call $(\hat{\alpha} - \alpha^*)(x)$ the variable and $(\alpha^* - \alpha)(x)$ the stable terms. We now show that the variable part behaves asymptotically like the variance term in $d + 1$-dimensional kernel density or regression estimation, and that the stable part behaves asymptotically like the bias term in $d + 1$-dimensional kernel density or regression estimation.

THEOREM 1 (Pointwise convergence). *Let $\varphi(x) = f(z, t)y(t)$, and assume that $\varphi(x) > 0$ on $\mathcal{N}$. We assume* (S): *α is twice and ϕ is once continuously differentiable on $\mathcal{N}$. We also assume* (K): *the kernel k has support $[-1, 1]$, is symmetric about zero and is continuous. Define the kernel moments $\kappa_1 = \int_{-1}^1 v^2k(v)\,dv$ and $\kappa_2 = \int_{-1}^1 k(v)^2\,dv$. Finally, we suppose* (B): *$nb^{d+1} \to \infty$ and $b \to 0$. Then the following hold:*

(a) $$n^{1/2}b^{(d+1)/2}\{\hat{\alpha}(x) - \alpha^*(x)\} \Rightarrow N\left[0, \kappa_2^{d+1}\frac{\alpha(x)}{\varphi(x)}\right];$$

(b) $$b^{-2}\{\alpha^*(x) - \alpha(x)\} \to_{\mathscr{P}} \kappa_1 \sum_{j=1}^{d+1} \left\{ \frac{\big(\partial\alpha(x)/\partial x_j\big)\big(\partial\varphi(x)/\partial x_j\big)}{\varphi(x)} + \frac{\partial^2\alpha(x)/\partial x_j^2}{2} \right\};$$

(c) $$\hat{\sigma}_x^2 = \mathscr{E}_x^{-2} n^{-1} b^{d+1} \sum_{i=1}^{n} \int K_b^2\{z - Z_i(s)\} k_b^2(t-s)\, dN_i(s) \to_{\mathscr{P}} \sigma_x^2 \equiv \kappa_2^{d+1} \frac{\alpha(x)}{\varphi(x)}.$$

REMARK 1. The theory for the marker-only estimator (9) under the marker-only model (2) is essentially the same. In this case, one must replace $\varphi(z, t)$ by $\Phi(z) = \int \varphi(z, s)\, ds$ (and only sum from 1 to d) in (a) and (b).

REMARK 2. The asymptotic variance is as given in MU's Theorem 3, apart from the kernel constants. Theorem 1 allows the construction of pointwise confidence bands for $\alpha^*(x)$ that have asymptotically correct coverage. When either the estimator is undersmoothed, that is, $bn^{1/(d+5)} \to 0$, or

$$\sum_{j=1}^{d+1} \left\{ \frac{\big(\partial\alpha(x)/\partial x_j\big)\big(\partial\varphi(x)/\partial x_j\big)}{\varphi(x)} + \frac{\partial^2\alpha(x)/\partial x_j^2}{2} \right\} = 0,$$

then the bias is negligible, and these intervals are in fact centered on $\alpha(x)$.

PROOF OF THEOREM 1. First we prove (a). We first show that $\mathscr{E}_x \to_{\mathscr{P}} \varphi(x)$. By changing variables to $u = (z - w)/b$ and $r = (t - s)/b$, we have

$$E(\mathscr{E}_x) = \int K_b(z-w) k_b(t-s) f(w, s) y(s)\, dw\, ds = \int_{-t/b}^{(1-t)/b} \int_{-z/b}^{(1-z)/b} K(u) k(r) \varphi(x - bq)\, du\, dr,$$

where $q = (u, r)$. Since x is an interior point, the range of integration in the second integral eventually contains the support of the product kernel. In the sequel we will ignore this boundary issue. Now, $\int_{[-1,1]^{d+1}} K(u)k(r)\varphi(x - bq)\, du\, dr \to \varphi(x)$ by continuity and Lebesgue's dominated convergence theorem. Furthermore,

$$\operatorname{Var}(\mathscr{E}_x) \le n^{-1} \int\int K_b^2(z-w) k_b^2(t-s) f(w, s) y(s)\, dw\, ds,$$

by (10). Changing variables and employing dominated convergence, we get $\operatorname{Var}(\mathscr{E}_x) \le O(n^{-1}b^{-(d+1)})$. It therefore remains to show that $n^{1/2}b^{(d+1)/2}\mathscr{Y}_x \Rightarrow N[0, \kappa_2^{d+1}\alpha(x)\varphi(x)]$. To apply Proposition 1, it suffices to show (i) and (ii):

(i) $\sum_{i=1}^n \int H_{ni}^2(z, t, s)\alpha\{Z_i(s), s\} Y_i(s)\, ds \to_{\mathscr{P}} \kappa_2^{d+1}\alpha(x)\varphi(x)$;

(ii) $\forall \varepsilon > 0$: $\sum_{i=1}^n \int H_{ni}^2(z, t, s) \mathrm{I}_{\{|H_{ni}(z,t,s)| > \varepsilon\}} \alpha\{Z_i(s), s\} Y_i(s)\, ds \to_{\mathscr{P}} 0$,

where $H_{ni}(z, t, s) = n^{-1/2}b^{(d+1)/2}K_b\{z - Z_i(s)\}k_b(t-s)$. (i) By (10), we can

replace the left-hand side of (i) by its expected value

$$(11) \qquad b^{d+1}\int\int K_b^2(z-w)k_b^2(t-s)\alpha(w,s)f(w,s)y(s)\,dw\,ds.$$

By a change of variables, (11) is approximately

$$\int_{[-1,1]^{d+1}} K^2(u)k^2(r)\alpha(x-bq)\varphi(x-bq)\,du\,dr,$$

which converges to $\kappa_2^{d+1}\alpha(x)\varphi(x)$ by dominated convergence. (ii) We have $\forall \varepsilon > 0, \exists n_0$ such that, $\forall n > n_0$,

$$\mathrm{I}_{\{n^{-1/2}b^{(d+1)/2}K_b[z-Z_i(s)]k_b(t-s)>\varepsilon\}} = 0,$$

for all s and for $i = 1,\ldots,n$. Therefore, (ii) follows and (a) is proven.

Now we prove (b). First, note that

$$E(\mathscr{B}_x) = \int_{[-1,1]^{d+1}} K(u)k(r)\{\alpha(x-bq)-\alpha(x)\}\varphi(x-bq)\,du\,dr,$$

after a change of variables. By Taylor's theorem,

$$\alpha(x+bq)-\alpha(x) = b\sum_{j=1}^{d+1}\frac{\partial\alpha(x)}{\partial x_j}q_j + b^2\sum_{j=1}^{d+1}\sum_{l=1}^{d+1}\frac{\partial^2\alpha(x^*)}{\partial x_j\,\partial x_l}\frac{q_jq_l}{2},$$

$$\varphi(x+bq) = \varphi(x) + b\sum_{j=1}^{d+1}\frac{\partial\varphi(x^{**})q_j}{\partial x_j},$$

where $|x_j^* - x_j|$, $|x_j^{**} - x_j| < b|q_j|$, for $j \in \{1,\ldots,d+1\}$. However, since $\int_{-1}^{1} k(v)v\,dv = 0$, we have

$$E(\mathscr{B}_x) = b^2\kappa_1\sum_{j=1}^{d+1}\left\{\frac{\partial\alpha(x)}{\partial x_j}\frac{\partial\varphi(x)}{\partial x_j} + \varphi(x)\frac{\partial^2\alpha(x)/\partial x_j^2}{2}\right\}\{1+o(1)\},$$

by continuity and dominated convergence. Now we verify that $\mathscr{B}_x$ can be approximated by its expected value. This follows because

$$\begin{aligned}\mathrm{Var}(\mathscr{B}_x) &\le n^{-1}\int\int K_b^2(z-w)k_b^2(t-s)\{\alpha(w,s)-\alpha(x)\}^2 \\ &\qquad \times f(w,s)y(s)\,dw\,ds \\ &= n^{-1}b^{-(d+1)}\int_{[-1,1]^{d+1}} K^2(u)k^2(r)\{\alpha(x-bq)-\alpha(x)\}^2 \\ &\qquad \times \varphi(x-bq)\,du\,dr,\end{aligned}$$

by (10) and a change of variables. This is $O(b^2n^{-1}b^{-(d+1)})$ by Taylor expansion and symmetry of k, by the continuity of α and φ and by dominated convergence. Finally, recall that $\mathscr{E}_x \to_{\mathscr{P}} \varphi(x)$, and on division we get (b).

Part (c) follows from

$$(12) \quad \mathscr{E}_x^{-2}n^{-1}b^{d+1}\sum_{i=1}^{n}\int K_b^2\{z-Z_i(s)\}k_b^2(t-s)\,d\langle M_i\rangle(s) \equiv \tilde{\sigma}_x^2 \to_{\mathscr{P}} \sigma_x^2$$

and

$$\hat{\sigma}_x^2 - \tilde{\sigma}_x^2 = \mathscr{E}_x^{-2} n^{-1} b^{d+1} \sum_{i=1}^{n} \int K_b^2\{z - Z_i(s)\} k_b^2(t-s)\, dM_i(s) = o_P(1).$$

By Lenglart's inequality [see Shorack and Wellner (1986), page 892], it suffices to bound the quadratic variation

$$\mathscr{E}_x^{-4} n^{-2} b^{2d+2} \sum_{i=1}^{n} \int K_b^4\{z - Z_i(s)\} k_b^4(t-s)\, d\langle M_i \rangle(s),$$

which is $o_P(1)$ by (10), since $\mathscr{E}_x = O_p(1)$ and $nb^{d+1} \to \infty$. □

The rate $b \sim n^{-1/(d+5)}$ is optimal as far as pointwise convergence is concerned; this agrees with the optimal rate of convergence in $(d+1)$-dimensional kernel density estimation. Of course in high dimensions the rate is very slow. Furthermore, it is hard to visualize the effects when more than one covariate is included. For these reasons, it may be desirable to impose some additional structure on α. For example, the multiplicative $[\alpha(x) = \alpha_0(t)\alpha_1(z_1) \cdots \alpha_d(z_d)]$ or additive $[\alpha(x) = \alpha_0(t) + \alpha_1(z_1) + \cdots + \alpha_d(z_d)]$ structures discussed in Andersen, Borgan, Gill and Keiding (1992). The components $\alpha_0, \ldots, \alpha_d$ can be estimated by backfitting [see Hastie and Tibshirani (1990)] or by marginal integration [see Linton and Nielsen (1995)]. The marginal integration method requires as input an initial $(d+1)$-dimensional smoother, which is provided by our method. Semiparametric models may also be useful, such as a partial Cox model $\alpha(x) = \alpha(z_2, t) \times \exp(\beta z_1)$, where $z = (z_1, z_2)$ and β is a finite dimensional parameter vector. In this case, $\alpha(z_2, t)$ can be estimated by our procedure once we have estimates of β. See Bickel, Nielsen and Linton (1993) for further discussion of estimation inside these models.

The explicit formula given in (b) can be used to examine the magnitude of bias in some canonical models. Suppose that

$$\alpha(z, t) = \beta_1 \exp(\beta_2 z), \tag{13}$$

where z is uniform on $[0, 1]$ for all t. Then our estimator is always positively biased: relative to $\alpha(z, t)$ the bias is $\kappa_1 \beta_2^2 b^2/2$ at all interior points, regardless of the censoring amount. McKeague and Utikal (1991) simulated the special case of this where $\beta_1 = 1/2$ and $\beta_2 = 2$, in which case the relative bias would be $2b^2$, when a Gaussian kernel is used. Our theorem does not deal with estimation near the boundary at which points the bias is typically the larger $O(b)$; in this region, the asymmetric kernel approach described in Andersen, Borgan, Gill and Keiding (1992) can be used to ensure boundary bias of $O(b^2)$. Standard bias reduction techniques can also be accommodated by our theorem under additional smoothness: if a kernel is chosen that satisfies $\int u^j k(u)\, du = 0$, $j = 1, 2, \ldots, r-1$, and $\int u^r k(u)\, du < \infty$, then the stable part of $\hat{\alpha}(x)$ is $O(b^r)$, provided $\alpha(x)$ is r-times continuously differentiable.

Theorem 1 can also be used to develop a plug-in method of bandwidth choice, based either on nonparametric estimates of the mean square error (MSE) or integrated mean square error (IMSE), or on parametric estimates of the same quantities derived from a reference model as in Silverman (1978). For (13), the IMSE optimal bandwidth is

$$b = \left[\frac{2\kappa_2}{\kappa_1}\right]^{1/12} \left[\beta_1^2\beta_2^2(\exp(\beta_2) + 1)\right]^{-1/6} n^{-1/6}.$$

This can be used as a rule of thumb on substituting consistent estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ for β_1 and β_2. In the McKeague and Utikal (1991) special case, the optimal bandwidth is $0.58n^{-1/6}$ for the Gaussian kernel.

We investigate our procedure using the San Francisco Men's Health Study data [see Fusaro, Nielsen and Scheike (1993)]. This was obtained from a prospective study of 1034 single men living in the 19 San Francisco census tracts with the highest cumulative AIDS incidence as of 1983. We examine the risk of developing AIDS for HIV positive subjects. This is to be explained by the serological markers: $CD4^+$ T lymphocyte cell count and $CD8^+$ T lymphocyte cell count. The time since the subject became HIV positive was frequently not observed, and so the marker-only procedure (9) is used. We used the kernel $k(u) = I_{\{|u| \le 1\}}[\cos(\pi u) + 1]/2$. Figure 1 shows the estimated hazard function (with only one marker, the CD4 count, included) for three different bandwidths. A steep increase in risk is apparent for individuals with CD4 count less than 200.

In Figure 2 we plot the estimated hazard from the bivariate procedure with both CD4 and CD8 markers. The peak in the hazard for subjects with

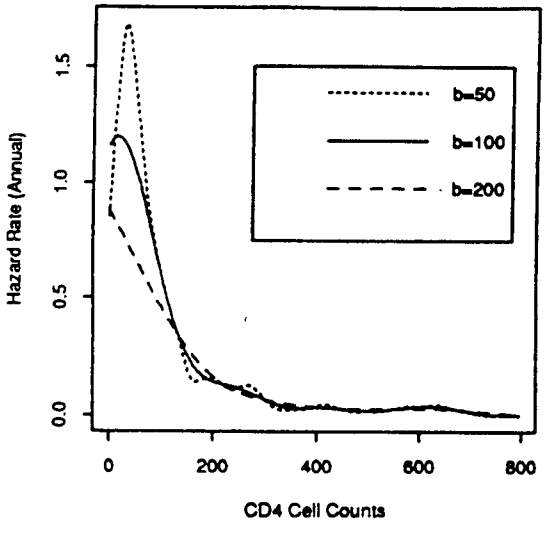

FIG. 1. *Hazard function estimate against $CD4^+$ T lymphocyte cell count for bandwidths b = 50, 100 and 200.*

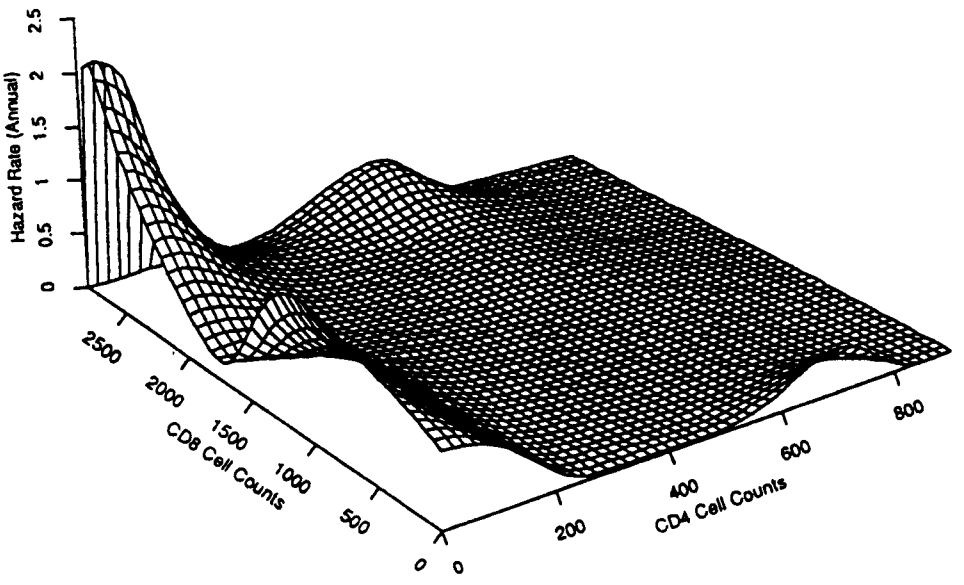

FIG. 2. *Hazard function estimate against CD4$^+$ and CD8$^+$ T lymphocyte cell count for bandwidths* (200, 650).

less than 200 CD4 and more than 1750 CD8 is quite dramatic and is consistent with the theory that elevated CD8 counts reflects immune activation in the face of advancing HIV. There is also some evidence of high risk for low CD8 count.

We also investigated a method of bandwidth choice based on cross-validation for the single marker example. By analogy with the popular integrated squared error (ISE) measure from density and, perhaps more pertinently, regression estimation [see Rudemo (1982) and Härdle (1990), Chapter 5], we chose b to minimize

$$Q(b) = \sum_{i=1}^{n} \int \hat{\alpha}_{-i}^{2}\{Z_i(s)\}Y_i(s)\,ds - 2\sum_{i=1}^{n} \int \hat{\alpha}_{-i}\{Z_i(s)\}\,dN_i(s),$$

where $\hat{\alpha}_{-i}\{Z_i(s)\}$ is a leave-one-out version of our estimator (9). This is asymptotically equivalent to minimizing the ISE performance measure $\sum_{i=1}^{n} \int [\hat{\alpha}_{-i}\{Z_i(s)\} - \alpha\{Z_i(s)\}]^2 Y_i(s)\,ds$. In an unpublished report in Danish, Ramlau-Hansen (1981) introduced a version of this method for the no-covariate special case (3). This procedure is further discussed in Nielsen (1990) and Andersen, Borgan, Gill and Keiding [(1992), page 246]. Figure 3 shows the cross-validation curve against the bandwidth. The optimal bandwidth here, $b = 65$, is quite close, in terms of the estimated hazard function (not reported), to the $b = 100$ chosen by eyeball in Fusaro, Nielsen and Schieke (1993).

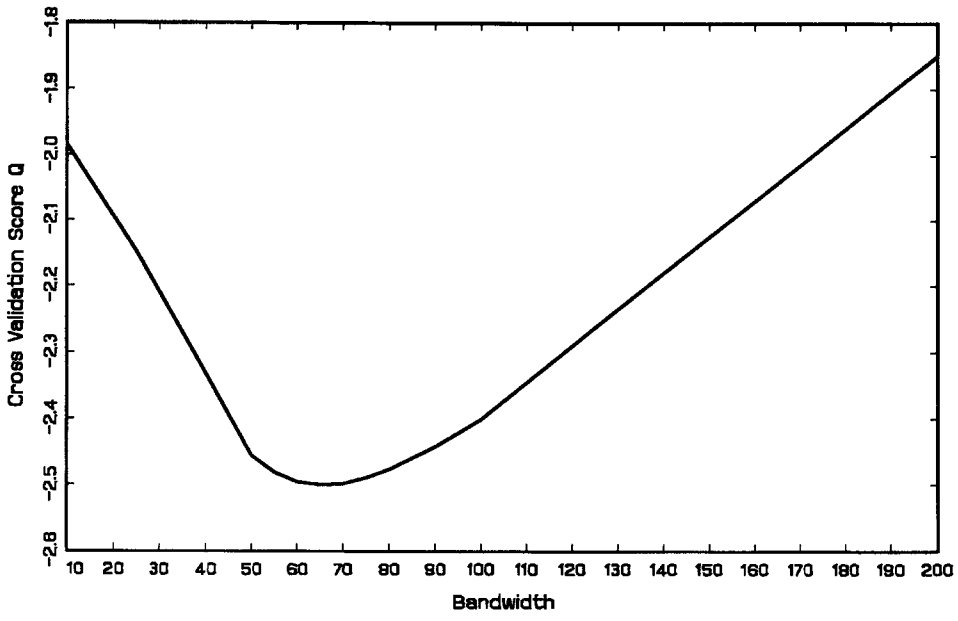

FIG. 3. *Cross-validation curve versus bandwidth.*

4.2. *Global convergence of* $\hat{\alpha}$. Our result on global convergence is established by verifying the conditions of the following lemma. This is a special case of a result by Bickel and Wichura (1971) that generalizes the tightness criterion by Billingsley [(1968), 12.11] to stochastic processes with multidimensional index set. We do not explicitly state the convergence rate; see Silverman (1978) and Nielsen (1990) for rates in, respectively, density estimation and intensity estimation inside Aalen's multiplicative intensity model.

LEMMA 1. *Let* $X(t)$ *be a stochastic process with* $t = (t_1, \ldots, t_d) \in [0, 1]^d$. *For any* $t \in [0, 1]^d$ *and* $v \in [0, 1]$, *let* $t_{j,v} = (t_1, \ldots, t_{j-1}, v, t_{j+1}, \ldots, t_d)$. *If for some* $C > 0$:

(L1) $X(t) \to_{\mathscr{P}} 0$ *for all* $t \in [0, 1]^d$;

(L2) $E\{X(t) - X(t_{j,u})\}^2 \leq C|t_j - u|^2$ *for all* $t \in [0, 1]^d$, $u \in [0, 1]$ *and* $j \in \{1, \ldots, d\}$;

then

$$\sup_{t \in [0,1]^d} |X(t)| \to_{\mathscr{P}} 0.$$

PROOF. We take our notation from Bickel and Wichura (1971). Let $B = (s, t]$ be a block, where $s, t \in [0, 1]^d$, and let

$$X(B) = \sum_{\varepsilon_1 = 0,1} \cdots \sum_{\varepsilon_q = 0,1} (-1)^{q - \sum_p \varepsilon_p} X\{s_1 + \varepsilon_1(t_1 - s_1), \ldots, s_q + \varepsilon_q(t_q - s_q)\}$$

be the increment of X around B. It is easy to verify that the moment

condition (3) in Bickel and Wichura (1971) is fulfilled if

$$E\{X(B)^2\} \le \mu(B)^2,$$

for some finite nonnegative measure μ on $[0,1]^d$. However, it follows from assumption (L2) and the triangle inequality that, for some $C > 0$,

$$E\{X(B)^2\} \le C \sum_{i=1}^{d} (t_i - s_i)^2 \le C\mu(B)^2,$$

where μ is the Euclidean measure. The result therefore follows from Theorem 3 in Bickel and Wichura (1971). □

THEOREM 2 (Global convergence). *Let $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_{d+1}$ be a $(d+1)$-dimensional subset of $[0,1]^{d+1}$, where each $\mathscr{X}_i$ is a compact interval. Assume that $\inf_{x \in \mathscr{X}} \varphi(x) > 0$. We make assumption* (K) *from Theorem* 1, *but in addition assume that k is Lipschitz continuous, that is, there exists $C > 0$, such that*

$$|k(u) - k(v)| \le C|u - v|,$$

for all u, v. We assume only that both α and φ are continuous on $\mathscr{X}$. Finally assume (B′): $b \to 0$ *and* $nb^{d+3} \to \infty$. *Then*

$$\sup_{x \in \mathscr{X}} |(\hat{\alpha} - \alpha)(x)| \to_{\mathscr{P}} 0.$$

PROOF. We use the notation from Theorem 1. It suffices to show the following:

(a) $\sup_{x \in \mathscr{X}} |\mathscr{V}_x| = o_P(1)$;
(b) $\sup_{x \in \mathscr{X}} |\mathscr{B}_x - E(\mathscr{B}_x)| = o_P(1)$;
(c) $\sup_{x \in \mathscr{X}} |E(\mathscr{B}_x)| = o(1)$;
(d) $\sup_{x \in \mathscr{X}} |\mathscr{E}_x^{-1}| = O_P(1)$.

Pointwise convergence has already been established so we concentrate on (L2). For any vector $(z_1, z_2, \ldots, z_d, t) = x \in \mathscr{X}$, let $x^* = (z_1^*, z_2, \ldots, z_d, t)$ and let $z^* = (z_1^*, z_2, \ldots, z_d)$ denote the corresponding subvector. The following equality holds because the stochastic processes representing the n individuals are assumed to be iid:

$$\begin{aligned} E[\mathscr{V}_x - \mathscr{V}_{x^*}]^2 &= n^{-2} \sum_{i=1}^{n} E\int [K_b\{z^* - Z_i(s)\} - K_b\{z - Z_i(s)\}]^2 k_b^2(t-s)\, d\langle M_i\rangle(s) \\ &= n^{-2} \sum_{i=1}^{n} E\int [K_b\{z^* - Z_i(s)\} - K_b\{z - Z_i(s)\}]^2 \\ &\qquad \times k_b^2(t-s)\alpha\{Z_i(s), s\}Y_i(s)\, ds. \end{aligned}$$

Then, changing variables we obtain, approximately,

$$n^{-1}b^{-(d+1)} \int_{[-1,1]^{d+1}} \left\{K\left(u + \frac{z^* - z}{b}\right) - K(u)\right\}^2 \times k^2(r)\alpha(x - bq)\varphi(x - bq)\, du\, dr.$$

However, $\{K(u + (z^* - z)/b) - K(u)\}^2 \le Cb^{-2}(z_1^* - z_1)^2$ by the Lipschitz continuity of k. Therefore,

$$E[\mathscr{V}_x - \mathscr{V}_{x^*}]^2 \le Cn^{-1}b^{-(d+3)}(z_1^* - z_1)^2 \to 0,$$

for any z_1^*, z_1. Similarly,

$$E\left[\mathscr{B}_{x^*} - E(\mathscr{B}_{x^*}) - \{\mathscr{B}_x - E(\mathscr{B}_x)\}\right]^2$$
$$\le n^{-1}b^{-(d+1)}\int_{[-1,1]^{d+1}} \left\{K\left(u + \frac{z^* - z}{b}\right) - K(u)\right\}^2 k^2(r)$$
$$\times\{\alpha(x - bq) - \alpha(x)\}^2\varphi(x - bq)\,du\,dr$$

and the result (b) follows from the Lipschitz continuity and boundedness of k.

In (c) the expression derived in Theorem 1 for $E(\mathscr{B}_x)$ can be employed directly. Since by assumption $\inf_{x\in\mathscr{X}}\varphi(x) > 0$, it suffices in (d) to show that $\sup_{x\in\mathscr{X}}|\mathscr{E}_x - \varphi(x)| \to_{\mathscr{P}} 0$, but this follows by the same arguments given above. □

Note that our bandwidth condition for $d = 1$ is strictly weaker than the corresponding condition from Theorem 4 in MU, which is as follows, on imposing $b = \tilde{b}$: $b \to 0$, $nw_nb^4 \to \infty$ and $w_n = o(b^4)$, where w_n is a third smoothing parameter they needed. McKeague and Utikal (1990) did not treat general d.

Acknowledgments. We would like to thank Peter Bickel, Nick Jewell, Chris Jones, Søren Johansen, Niels Keiding, Thomas Scheike and Rob Fusaro for helpful discussions. The helpful comments of two referees are gratefully acknowledged. We also thank Warren Winkelstein, Jr., and the San Francisco Men's Health Study for allowing us to use their data.

REFERENCES

AALEN, O. O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

AALEN, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8** 907–925.

ANDERSEN, P. K. and BORGAN, Ø. (1985). Counting process models for life history data: a review. *Scand. J. Statist.* **12** 97–158.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1992). *Statistical Models Based on Counting Processes.* Springer, New York.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

BERAN, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Dept. Statistics, Univ. California, Berkeley.

BICKEL, P. J., NIELSEN, J. P. and LINTON, O. B. (1993). A semiparametric hazard model with parametric time and nonparametric covariate dependency. Cowles Foundation, Yale Univ. Unpublished manuscript.

BICKEL, P. J. and WICHURA, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42** 1656–1670.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

BRESLOW, N. and CROWLEY, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.* **2** 437–453.

COX, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

DABROWSKA, D. M. (1987). Nonparametric regression with censored survival time data. *Scand. J. Statist.* **14** 181–197.

FUSARO, R., NIELSEN, J. P. and SCHEIKE, T. (1993). Marker dependent hazard estimation. An application to AIDS. *Statistics in Medicine* **12** 843–865.

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method I. *Scand. J. Statist.* **16** 97–128.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

HJORT, N. L. (1994). Dynamic likelihood hazard rate estimation. *Biometrika.* To appear.

JEWELL, N. P. and NIELSEN, J. P. (1993). A framework for consistent prediction rules based on markers. *Biometrika* **80** 153–164.

JONES, M. C., DAVIES, S. J. and PARK, B. U. (1994). Versions of kernel-type regression estimators. *J. Amer. Statist. Assoc.* **89** 825–832.

KEIDING, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *J. Roy. Statist. Soc. Ser. A* **154** 371–412.

KEIDING, N., HOLST, C. and GREEN, A. (1989). Retrospective calculation of diabetes incidence from information in a current prevalent population and historical mortality. *American Journal of Epidemiology* **130** 588–600.

LINTON, O. B. and NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric based on marginal integration. *Biometrika.* **82** 93–100.

MCKEAGUE, I. W. and UTIKAL, K. J. (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.* **18** 1172–1187.

MCKEAGUE, I. W. and UTIKAL, K. J. (1991). Goodness-of-fit tests for additive hazards and proportional hazards models. *Scand. J. Statist.* **18** 177–195.

NIELSEN, J. P. (1990). Kernel estimation of densities and hazards: a counting process approach. Ph.D. dissertation, Dept. Biostatistics, Univ. California, Berkeley.

O'SULLIVAN, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124–145.

RAMLAU-HANSEN, H. (1981). Udglatning med kernefunktioner i forbindelse med tælleprocessor. Working Paper 41, Laboratory of Actuarial Science, Univ. Copenhagen.

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

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives. *Ann. Statist.* **6** 177–184.

THOMAS, D. C. (1983). Non-parametric estimation and tests of fit for dose response relations. *Biometrics* **39** 263–268.

PFA PENSION
MARINA PARK
SUNDKROGSGADE 4
COPENHAGEN 2100KBH
DENMARK

DEPARTMENT OF ECONOMICS
YALE UNIVERSITY
P.O. BOX 208281
NEW HAVEN, CONNECTICUT 06520-8281