The Annals of Statistics
2003, Vol. 31, No. 2, 464–492

ESTIMATING MULTIPLICATIVE AND ADDITIVE HAZARD FUNCTIONS BY KERNEL METHODS

BY OLIVER B. LINTON,[1] JENS PERCH NIELSEN[2] AND SARA VAN DE GEER

London School of Economics, Codan and University of Leiden

We propose new procedures for estimating the component functions in both additive and multiplicative nonparametric marker-dependent hazard models. We work with a full counting process framework that allows for left truncation and right censoring and time-varying covariates. Our procedures are based on kernel hazard estimation as developed by Nielsen and Linton and on the idea of marginal integration. We provide a central limit theorem for the marginal integration estimator. We then define estimators based on finite-step backfitting in both additive and multiplicative cases and prove that these estimators are asymptotically normal and have smaller variance than the marginal integration method.

1. Introduction. Suppose that the conditional hazard function

$$\lambda(t|Z_i) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} P\big(T_i \le t + \varepsilon | T_i > t; \big(Z_i(s), s \le t\big)\big)$$

for the survival time T_i of an individual i with the covariate or marker process $Z_i = (Z_i(t))$ has the form

$$\lambda(t|Z_i) = \alpha\big(t, Z_i(t)\big), \tag{1}$$

where α is an unknown function of time t and the value of the covariate process of the individual at time t only. Inference for this general class of models was initiated by Beran (1981) and extended by Dabrowska (1987), McKeague and Utikal (1990) and Nielsen and Linton (1995). Nielsen and Linton (1995) established asymptotic normality and uniform convergence of their estimators of $\alpha(t, z)$ in the case where one observes the event times of a sample of mutually independent individuals along with their covariate processes, but where there has perhaps been some (noninformative) censoring and truncation. Unfortunately, the achievable rate of convergence of estimators of $\alpha(t, z)$ increases rapidly with the number of covariates, as in the regression case studied by Stone (1980). Furthermore, it is hard to visualize the model in high dimensions.

This motivates the study of separable structures, in particular, additive and multiplicative models. These models can be used in their own right or as an aid to

Received February 2001; revised April 2002.
[1]Supported in part by the NSF.
[2]Supported in part by Bergiafonden.
AMS 2000 subject classifications. 62G05, 62M09.
Key words and phrases. Additive model, censoring, kernel, proportional hazards, survival analysis.

further model specification. They allow for the visual display of the components of high-dimensional models and for a clean interpretation of effects. Also, the optimal rate of convergence in additive and other separable regression models has been shown to be better than in the unrestricted case; see Stone (1985, 1986). In this paper, we consider additive and multiplicative submodels of (1). Multiplicative separability of the baseline hazard from the covariate effect has played a central role in survival analysis as is evident from the enormous literature inspired by Cox (1972); see Andersen, Borgan, Gill and Keiding [(1993), Chapter 7] for a discussion of semiparametric and nonparametric hazard models, and see Lin and Ying (1995), Dabrowska (1997), Nielsen, Linton and Bickel (1998) and Huang (1999) for some recent contributions. Additive models are perhaps less common, but have been studied in Aalen (1980) and McKeague and Utikal (1991).

We propose a class of kernel-based marginal integration estimators for the component functions in nonparametric additive and multiplicative models. This methodology has been developed in Linton and Nielsen (1995) for regression. We extend this literature to counting process models in which a wide range of censoring and truncation can be allowed. The estimation idea involves integrating out a high-dimensional preliminary estimator, which we call the "pilot"; in our case this is provided by Nielsen and Linton's (1995) kernel hazard estimator. The averaging (or integration) reduces variance and hence permits faster convergence rates. We establish that marginal integration estimators converge pointwise and indeed uniformly at the same rate as a one-dimensional regression estimator would; we also give the limiting distributions.

Marginal integration estimators are known to be inefficient in general, and in particular to have higher mean squared error than a corresponding oracle estimator that could be computed were all but one of the component functions known; see Linton (1997, 2000) for discussion in regression and other models. This motivates our extension to "m-step" estimators, which in other contexts have been shown to improve efficiency [Linton (1997, 2000)]. The origin of this estimator lies in the backfitting methodology as applied to nonparametric regression in Hastie and Tibshirani (1990). The "usual" backfitting approach as implemented in regression (for counting processes we have not found a reference) is to use an iterative solution scheme to some sample equations that correspond to the population projection interpretation of the additive model, say. Starting from some initial condition, one continues until some convergence criterion is satisfied. Under some conditions this algorithm converges; see Opsomer and Ruppert (1997) and Mammen, Linton and Nielsen (1999). We shall work with certain backfitting equations but start with a consistent estimator of the target functions, and we shall just iterate a finite number (m) of times. We establish the asymptotic distribution of the m-step method; under some conditions, it achieves an oracle efficiency bound. Specifically, the asymptotic variance of the m-step estimator is the same as that of the estimator one would use when the other components are known; this is true for any m, and, in particular, for $m = 1$. In the additive regression case, Linton (1997) proved a similar result. We define the corresponding concepts for hazard estimation in both additive and multiplicative cases. One-step and m-step approximations to maximum likelihood estimators in parametric models have been widely studied, following Bickel (1975). The application of this idea in nonparametric estimation has only come about fairly recently; see Fan and Chen (1999).

We provide a new result on uniform convergence of kernel hazard estimators in the counting process framework. This result is fundamental to the proofs of limiting properties of many nonparametric and semiparametric procedures, including our own. The result contained herein greatly improves and extends the result contained in Nielsen and Linton (1995) and gives the optimal rate. Our proof makes use of the recently derived exponential inequality for martingales obtained in van de Geer (1995). This paper is an abbreviated version of Linton, Nielsen and van de Geer (2001), which contains more details and references to applications.

2. The marker-dependent hazard model.

2.1. *The observable data.* Let T be the survival time and let $\widetilde{T} = \min\{T, C\}$, where C is the censoring time. Suppose that T and C are conditionally independent given the left-continuous covariate process Z and suppose that the conditional hazard of T at time t given $\{Z(s),\ s \le t\}$ is $\alpha(t, Z(t))$. For each of n independent copies (T_i, C_i, Z_i), $i = 1, \dots, n$, of (T, C, Z), we observe $\widetilde{T}_i, \delta_i = \mathbb{1}(T_i < C_i)$ and $Z_i(t)$ for $t \le T_i$. Define also $Y_i(t) = \mathbb{1}(\widetilde{T}_i \le t)$, the indicator that the individual is observed to be at risk at time t, and $N_i(t) = \mathbb{1}(\widetilde{T}_i > t, \delta_i = 1)$. Then $\mathbf{N}(t) = (N_1(t), \dots, N_n(t))$ is a multivariate counting process, and N_i has intensity $\lambda_i(t) = \alpha(t, Z_i(t))Y_i(t)$, as we discuss below. See Linton, Nielsen and van de Geer (2001) for more discussion.

2.2. *The counting process formulation.* We next embed the above model inside the counting process framework laid down in Aalen (1978). This framework is very general and can be shown to accommodate a wide variety of censoring mechanisms, including that of the previous section. Let $\mathbf{N}^{(n)}(t) = (N_1(t), \dots, N_n(t))$ be an n-dimensional counting process with respect to an increasing, right-continuous, complete filtration $\mathcal{F}_t^{(n)}$, $t \in [0, T]$; that is, $\mathbf{N}^{(n)}$ is adapted to the filtration and has components N_i, which are right-continuous step functions, zero at time zero, with jumps of size one such that no two components jump simultaneously. Here, $N_i(t)$ records the number of observed failures for the ith individual during the time interval $[0, t]$ and is defined over the whole period (taken to be $[0, T]$, where T is finite). Suppose that N_i has intensity

$$\lambda_i(t) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} P\big(N_i(t+\varepsilon) - N_i(t) = 1 | \mathcal{F}_t^{(n)}\big) = \alpha\big(t, Z_i(t)\big)Y_i(t), \tag{2}$$

where Y_i is a predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the ith individual is observed to be at risk, while Z_i is a d-dimensional predictable covariate process with support in some compact set $\mathcal{Z} \subseteq \mathbb{R}^d$. The function $\alpha(t, z)$ represents the failure rate for an individual at risk at time t with covariate $Z_i(t) = z$.

We assume that the stochastic processes $(N_1, Z_1, Y_1), \ldots, (N_n, Z_n, Y_n)$ are independent and identically distributed (i.i.d.) for the n individuals. In the rest of the paper we therefore drop the n superscript for convenience. This simplifying assumption has been adopted in a number of leading papers in this field, for example, Andersen and Gill [(1982), Section 4] and McKeague and Utikal [(1990), Section 4]. Let $\mathcal{F}_{t,i} = \sigma\{N_i(u), Z_i(u), Y_i(u); u \le t\}$ and $\mathcal{F}_t = \bigvee_{i=1}^n \mathcal{F}_{t,i}$. With these definitions, λ_i is predictable with respect to $\mathcal{F}_{t,i}$ and hence $\mathcal{F}_t$, and the processes $M_i(t) = N_i(t) - \Lambda_i(t)$, $i = 1, \ldots, n$, with compensators $\Lambda_i(t) = \int_0^t \lambda_i(u)\, du$, are square integrable local martingales with respect to $\mathcal{F}_{t,i}$ on the time interval $[0, T]$. Hence, $\Lambda_i(t)$ is the compensator of $N_i(t)$ with respect to both the filtration $\mathcal{F}_{t,i}$ and the filtration $\mathcal{F}_t$. In fact, rather than observing the whole covariate process Z_i, it is sufficient to observe Z_i at times when the individual is at risk, that is, when $Y_i(s) = 1$.

2.3. *Separable models and estimands.* For notational convenience we combine time and the covariates into one vector, that is, we write $x = (t, z)$ and $X_i(t) = (t, Z_i(t))$, and label the components of x as $0, 1, \ldots, d$, with $x_0 = t$. Let $x_{-j} = (x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$ be the $d \times 1$ vector of x excluding x_j and likewise for $X_{-ji}(s)$.

The main object of interest is the hazard function $\alpha(\cdot)$ and functionals computed from it. Consider the case that α is separable either additively or multiplicatively: the multiplicative model is that

$$\alpha(x) = c_M \prod_{j=0}^{d} h_j(x_j) \tag{3}$$

for some constant c_M and functions h_j, $j = 0, 1, \ldots, d$; the additive model is

$$\alpha(x) = c_A + \sum_{j=0}^{d} g_j(x_j) \tag{4}$$

for some constant c_A and functions g_j, $j = 0, 1, \ldots, d$. The constants and functions must be such that the hazard function itself is nonnegative everywhere. Also, the functions $h_j(\cdot)$ and $g_j(\cdot)$ and constants c_A and c_M are not separately identified, and we need to make a further restriction in both cases to obtain uniqueness. Let Q be a given absolutely continuous c.d.f. and define the marginals $Q_j(x_j) = Q(\infty, \ldots, \infty, x_j, \infty, \ldots, \infty)$ and $Q_{-j}(x_{-j}) = Q(x_0, \ldots, x_{j-1}, \infty, x_{j+1}, \ldots, x_d)$, $j = 0, 1, \ldots, d$. For simplicity of notation we shall suppose that $Q = Q_0 \otimes Q_1 \otimes \cdots \otimes Q_d$, although this is not necessary for the main results. We identify the models (3) and (4) through these probability measures. Specifically, we suppose that in the additive case $\int g_j(x_j)\,dQ_j(x_j) = 0$, while in the multiplicative case $\int h_j(x_j)\,dQ_j(x_j) = 1$ for each $j = 0, \dots, d$. These restrictions ensure that the model components $(c_A, g_0, \dots, g_d)$ and $(c_M, h_0, \dots, h_d)$, respectively, are well defined and imply that $c_A = c_M = c = \int \alpha(x)\,dQ(x)$.

Now consider the following contrasts:

$$\alpha_{Q_{-j}}(x_j) = \int \alpha(x)\,dQ_{-j}(x_{-j}),$$

$$\alpha^A_{Q_{-j}}(x_j) = \alpha_{Q_{-j}}(x_j) - \int \alpha_{Q_{-j}}(x_j)\,dQ_j(x_j) = \alpha_{Q_{-j}}(x_j) - c,$$

$$\alpha^M_{Q_{-j}}(x_j) = \frac{\alpha_{Q_{-j}}(x_j)}{\int \alpha_{Q_{-j}}(x_j)\,dQ_j(x_j)} = \frac{\alpha_{Q_{-j}}(x_j)}{c}, \tag{5}$$

$$\alpha_A(x) = \sum_{j=0}^{d} \alpha^A_{Q_{-j}}(x_j) + c, \qquad \alpha_M(x) = c\prod_{j=0}^{d} \alpha^M_{Q_{-j}}(x_j),$$

$j = 0, \dots, d$. In the additive model, $\alpha_{Q_{-j}}(x_j) = g_j(x_j) + c$ so that the recentered quantity $\alpha^A_{Q_{-j}}(x_j) = g_j(x_j)$, while in the multiplicative model, $\alpha_{Q_{-j}}(x_j) = h_j(x_j)c$ and the rescaled quantity $\alpha^M_{Q_{-j}}(x_j) = h_j(x_j)$. It follows that $\alpha_{Q_{-j}}(\cdot)$ is, up to a constant factor, the univariate component of interest in both additive and multiplicative structures. What happens when neither (3) nor (4) is true but only (2) holds? In this case, the quantity $\alpha_{Q_{-j}}(\cdot)$ still has an interpretation as an average of the higher dimensional surface with respect to Q_{-j}. In addition, one can also interpret $\sum_j \alpha_{Q_{-j}}(\cdot)$ as a projection: $\sum_j \alpha_{Q_{-j}}(\cdot)$ is the closest additive function to $\alpha(x)$ when distance is computed using a product measure; see Nielsen and Linton (1998).

3. Estimation. We first define a class of estimators of the unrestricted conditional hazard function $\alpha(x)$. Defining the bandwidth parameter b and product kernel function $K_b(u_0, \dots, u_d) = \prod_{j=0}^{d} k_b(u_j)$, where $k(\cdot)$ is a one-dimensional kernel with $k_b(u_j) = b^{-1}k(u_j/b)$, we let

$$\widehat{\alpha}(x) = \frac{(1/n)\sum_{i=1}^{n}\int_0^T K_b(x - X_i(s))\,dN_i(s)}{(1/n)\sum_{i=1}^{n}\int_0^T K_b(x - X_i(s))Y_i(s)\,ds} \equiv \frac{\widehat{o}(x)}{\widehat{e}(x)} \tag{6}$$

be our estimator of $\alpha(x)$, a ratio of local occurrence $\widehat{o}(x)$ to local exposure $\widehat{e}(x)$. The estimator $\widehat{\alpha}(x)$ was introduced in Nielsen and Linton (1995) who gave some statistical properties of (6) for general d. When the bandwidth sequence is chosen of order $n^{-1/(2r+d+1)}$, the random variable $\widehat{\alpha}(x) - \alpha(x)$ is asymptotically normal with rate of convergence $n^{-r/(2r+d+1)}$, where r is an index of smoothness of $\alpha(x)$. This is the optimal rate for the corresponding regression problem without separability restrictions; see Stone (1980). We shall be using $\widehat{\alpha}(x)$ as an input into our procedures and will refer to it as the "pilot" estimator. Although $\widehat{\alpha}(x)$ is not guaranteed to be positive everywhere when the kernel K takes on negative values, the probability of a negative value decreases to 0 very rapidly.

We now define a method of estimating the components in (3) and (4) based on the principle of marginal integration. We estimate the quantities $\alpha_{Q_{-j}}(x_j)$, c, $g_j(x_j)$, $h_j(x_j)$, $\alpha_A(x)$ and $\alpha_M(x)$ by replacing the unknown quantities by estimators; thus,

$$\widehat{\alpha}_{Q_{-j}}(x_j) = \int \widehat{\alpha}(x)\, d\widehat{Q}_{-j}(x_{-j}), \qquad \widehat{c} = \int \widehat{\alpha}(x)\, d\widehat{Q}(x), \tag{7}$$

$$\widehat{\alpha}^A_{Q_{-j}}(x_j) = \widehat{\alpha}_{Q_{-j}}(x_j) - \widehat{c}, \qquad \widehat{\alpha}^M_{Q_{-j}}(x_j) = \frac{\widehat{\alpha}_{Q_{-j}}(x_j)}{\widehat{c}}, \tag{8}$$

$$\widehat{\alpha}_A(x) = \sum_{j=0}^{d} \widehat{\alpha}^A_{Q_{-j}}(x_j) + \widehat{c}, \qquad \widehat{\alpha}_M(x) = \widehat{c} \prod_{j=0}^{d} \widehat{\alpha}^M_{Q_{-j}}(x_j), \tag{9}$$

where $\widehat{\alpha}(x)$ is the unrestricted estimator (6). Here, $\widehat{Q}$ is a probability measure that converges in probability to the distribution Q, while $\widehat{Q}_j$ and $\widehat{Q}_{-j}$, $j = 0, \dots, d$, are the corresponding marginals. We assume that $\widehat{Q}$ and its marginals are continuous except at a finite number of points, which implies that the integrals in (7)–(9) are well defined because $\widehat{\alpha}(\cdot)$ is continuous when K is.

The quantities $\widehat{\alpha}_A(x)$ and $\widehat{\alpha}_M(x)$ estimate consistently $\alpha_A(x)$ and $\alpha_M(x)$, respectively, which are both equal to $\alpha(x)$ in the corresponding submodel. For added flexibility, we suggest using a different bandwidth sequence in the estimator $\widehat{c}$; we can expect to estimate the constants at rate root n because the target quantities are integrals over the entire covariate vector.

The distribution $\widehat{Q}$ can essentially be arbitrary, although its support should be contained within the support of the covariates. The most obvious choices of Q seem to be Lebesgue measure on some compact set I or an empirical measure similarly restricted. There has been some investigation of the choice of weighting in regression; see, for example, Linton and Nielsen (1995), Fan, Härdle and Mammen (1998) and Cai and Fan (2000). Finally, the marginal integration procedures we have proposed are based on high-dimensional smoothers and can suffer some small-sample problems if the dimensions are high. See Sperlich, Linton and Härdle (1999) for numerical investigation.

4. Asymptotic properties. We derive the asymptotic distribution of the marginal integration estimators $\widehat{\alpha}_{Q_{-j}}$ at interior points under the general sampling scheme (2); that is, we do not assume either of the separable structures holds. However, when either the additive or multiplicative submodel (3) or (4) is true, our results are about the corresponding univariate components. We are assuming an i.i.d. setup throughout. We could weaken this along the lines of McKeague and Utikal [(1990), condition A] but at the cost of quite complicated notation. We shall assume that the support of $Z_i(s)$ does not depend on s and is rectangular. This is just to avoid a more complicated notation. We also assume that the estimation region is a strict rectangular subset of the covariate support, and so ignore boundary effects.

For any vectors $x = (x_1, \dots, x_p)$ and $a = (a_1, \dots, a_p)$ of common length p, we let $x^a = x_1^{a_1} \cdots x_p^{a_p}$ and $|a| = \sum_{j=1}^p a_j$. Finally, for any function $g : \mathbb{R}^p \to \mathbb{R}$, let $D^a g(x) = (\partial^{|a|}/\partial x_1^{a_1} \cdots \partial x_p^{a_p}) g(x)$. For functions $g : \mathbb{R}^p \mapsto \mathbb{R}$, define the Sobolev norm of order s, $\|g\|^2_{p,s,\mathcal{I}} = \sum_{a:|a| \le s} \int_{\mathcal{I}} \{D^a g(z)\}^2 \, dz$, where $\mathcal{I} \subseteq \mathbb{R}^p$ is a compact set, and let $\mathcal{G}_{p,s}(\mathcal{I})$ be the class of all functions with domain $\mathcal{I}$ with Sobolev norm of order s bounded by some constant C. An important step in our argument is to replace $\widehat{Q}$ by Q; we shall use empirical process arguments to show that this can be done without affecting the results. We make the following assumptions:

(A1) The covariate process is supported on the compact set $\mathcal{X} = [0, T] \times \mathcal{Z}$, where $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_d$. For each $t \in [0, T]$, define the conditional [given $Y_i(s) = 1$] distribution function of the observed covariate process $F_t(z) = \Pr(Z_i(t) \le z | Y_i(t) = 1)$ and let $f_t(z)$ be the corresponding density with respect to Lebesgue measure. For each $x = (t, z) \in \mathcal{X}$, define the exposure $e(x) = f_t(z) y(t)$, where $y(t) = E[Y_i(t)]$. The functions $t \mapsto y(t)$ and $t \mapsto f_t(z)$ are continuous on $[0, T]$ for all $z \in \mathcal{Z}$.

(A2) The probability measure $Q = Q = Q_0 \otimes Q_1 \otimes \cdots \otimes Q_d$ is absolutely continuous with respect to Lebesgue measure and has density function $q = q_0 \otimes q_1 \otimes \cdots \otimes q_d$. It has support on the compact interval $I = I_0 \times \cdots \times I_d$, which is strictly contained in $\mathcal{Z}$. Furthermore, $0 < \inf_{x_j \in I_j} q_j(x_j)$ for all j.

(A3) The functions $\alpha(\cdot)$ and $e(\cdot)$ are r-times continuously differentiable on $\mathcal{X}$ and satisfy $\inf_{x \in \mathcal{X}} e(x) > 0$ and $\inf_{x \in \mathcal{X}} \alpha(x) > 0$. The integer r satisfies $(2r + 1)/3 > (d + 1)$.

(A4) The kernel k has support $[-1, 1]$, is symmetric about 0 and is of order r, that is, $\int_{-1}^1 k(u) u^j \, du = 0$, $j = 1, \dots, r - 1$, and $\mu_r(k) = \int_{-1}^1 k(u) u^r \, du \in (0, \infty)$, where $r \ge 2$ is an even integer. The kernel is also $r - 1$ times continuously differentiable on $[-1, 1]$ with Lipschitz remainder; that is, there exists a finite constant k_{lip} such that $|k^{(r-1)}(u) - k^{(r-1)}(u')| \le k_{\mathrm{lip}} |u - u'|$ for all u, u'. Finally, $k^{(j)}(\pm 1) = 0$ for $j = 0, \dots, r - 1$.

(A5) The probability measure $\widehat{Q}$ has support on I and satisfies $\sup_{x \in I} |\widehat{Q}(x) - Q(x)| = O_p(n^{-1/2})$. Furthermore, for some s with $r \ge s > d/2$, the empirical process $\{\nu_n(\cdot) : n \ge 1\}$ with $\nu_n(g) = \sqrt{n}\{\int_{I_{-j}} g(z) \, d\widehat{Q}_{-j}(z) - \int_{I_{-j}} g(z) \, dQ_{-j}(z)\}$ for any $g \in \mathcal{G}_{d+1,s}(I_{-j})$, where the set $I_{-j} = \prod_{\ell \ne j} I_\ell$, is stochastically equicontinuous on $\mathcal{G}_{d+1,s}(I_{-j})$ at $g_0(\cdot) = \alpha(x_j, \cdot)$; that is, for all $\varepsilon, \eta > 0$ there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbf{P}^*\left[\sup_{g\in\mathcal{G}_{d+1,s}(I_{-j}),\|g-g_0\|_{d+1,s,I_{-j}}\le\delta} |\nu_n(g) - \nu_n(g_0)| > \eta\right] < \varepsilon, \tag{10}$$

where $\mathbf{P}^*$ denotes outer probability.

The smoothness and boundedness conditions in (A1), (A3) and (A4) are fairly standard in local constant kernel estimation. For simplicity, these conditions are assumed to hold on the entire support of the covariate process, whereas some of our results below can be established when these conditions hold only on I. Our assumptions are strictly stronger than those of McKeague and Utikal (1990) and indeed imply the conditions of their Proposition 1. In particular, we assume smoothness of e with respect to all its arguments rather than just continuity. We use this assumption in our proof of the limiting distribution of our estimator. If instead a local polynomial pilot estimator were used [see Fan and Gijbels (1996) and Nielsen (1998)], we would most likely require only continuity of the exposure e. Nevertheless, these conditions are likely to hold for a large class of covariate processes. Certainly, time-invariant covariates can be expected to satisfy this condition. When Z is the time since a certain event, such as onset of disability, we can model the stochastic process $Z_i(t)$ as $Z_i(t) = t - Z_{0i}$ for some random variable Z_{0i} that represents the age at which disability occurred. This is essentially as in McKeague and Utikal [(1990), Example 5, page 1180], especially, and under smoothness conditions on their α_{jk} we obtain the smoothness of (in our notation) the corresponding exposure $e(x)$. The restriction on (r, d) is used to ensure that the remainder terms in the expansion of $\widehat{\alpha} - \alpha$ are of smaller order in probability than the leading terms; the remainder terms are of order $n^{-1}b^{-(d+1)}\log n + b^{2r}$, so we must have $r > d$. We require slightly stronger restrictions in order to deal with the passage from $\widehat{Q}$ to Q. The stochastic equicontinuity condition in (A5) is satisfied under conditions on the entropy of the class of functions; see van de Geer (2000).

Our main theorem gives the pointwise distribution of the marginal integration estimator $\widehat{\alpha}_{Q_{-j}}(x_j)$ and the corresponding additive and multiplicative reconstructions $\widehat{\alpha}_A(x), \widehat{\alpha}_M(x)$. As discussed earlier, we do not maintain either separability hypothesis in this theorem, and so the result is about the functionals of the underlying function $\alpha(x)$.

THEOREM 1. *Suppose that* (A1)–(A5) *hold and that* $n^{1/(2r+1)}b \to \gamma$ *for some* γ *with* $0 < \gamma < \infty$. *Then there exist functions* $m_j(\cdot), \upsilon_j(\cdot)$ *that are bounded and continuous on* I_j *such that, for any* $x_j \in I_j$,

$$n^{r/(2r+1)}\big(\widehat{\alpha}_{Q_{-j}} - \alpha_{Q_{-j}}\big)(x_j) \Rightarrow N\left[m_j(x_j), \upsilon_j(x_j)\right], \tag{11}$$

where, with $\|k\|_2^2 = \int_{-1}^{1} k(u)^2\, du$,

$$\upsilon_j(x_j) = \gamma^{-1}\|k\|_2^2 \int_{I_{-j}} \frac{\alpha(x)q_{-j}^2(x_{-j})}{e(x)}\, dx_{-j}.$$

Suppose also that $\widehat{c} - c = O_P(n^{-1/2})$. *Then* $\widehat{\alpha}_{Q_{-j}}^{A}(x_j)$ *has the same asymptotic distribution as* $\widehat{\alpha}_{Q_{-j}}(x_j)$, *while* $\widehat{\alpha}_{Q_{-j}}^{M}(x_j)$ *has asymptotic mean* $m_j(x_j)/c$ *and asymptotic variance* $\upsilon_j(x_j)/c^2$. *Finally*,

$$n^{r/(2r+1)}(\widehat{\alpha}_A - \alpha_A)(x) \Rightarrow N[m_A(x), \upsilon_A(x)], \tag{12}$$

$$n^{r/(2r+1)}(\widehat{\alpha}_M - \alpha_M)(x) \Rightarrow N[m_M(x), \upsilon_M(x)], \tag{13}$$

where $m_A(x) = \sum_{j=0}^{d} m_j(x_j)$ *and* $\upsilon_A(x) = \sum_{j=0}^{d} \upsilon_j(x_j)$, *while* $m_M(x) = \sum_{j=0}^{d} m_j(x_j)s_j(x_{-j})$ *and* $\upsilon_M(x) = \sum_{j=0}^{d} \upsilon_j(x_j)s_j^2(x_{-j})$, *where* $s_j(x_{-j}) = \prod_{k\neq j} \alpha_{Q_{-k}}(x_k)/c^d$.

The bandwidth rate $b \sim n^{-1/(2r+1)}$ gives an optimal (pointwise mean squared error) rate of convergence for $\widehat{\alpha}_{Q_{-j}}(x_j)$, $\widehat{\alpha}_A(x)$ and $\widehat{\alpha}_M(x)$ [i.e., this is the same rate as the optimal rate of convergence in one-dimensional kernel regression estimation; see Stone (1980)]. The bias function $m_j(\cdot)$ is just proportional to the integrated bias of the pilot estimator, in our case the Nadaraya–Watson estimator. If instead we were to use a local polynomial pilot estimator [see Nielsen (1998) for the definition of the local linear estimator in hazard estimation], we would obtain a simpler expression for the bias and indeed an estimator that has better properties [see Fan and Gijbels (1996)]. Also, by undersmoothing in the direction not of interest (we have used the same bandwidth for all directions), we obtain a different bias expression that corresponds to the bias of the corresponding one-dimensional oracle smoother; see below. See Linton and Nielsen (1995) for discussion. Finally, the estimator $\widehat{c}$ is root-n consistent under slightly different bandwidth conditions: specifically, a standard proof along the lines of Nielsen, Linton and Bickel (1998) would require that $\sqrt{n}b^r \to 0$, which requires undersmoothing in all directions.

5. *m*-step backfitting. The marginal integration estimators defined above are inefficient. We suggest an alternative estimation method that is more efficient. We shall assume throughout this section that the corresponding submodel (additive or multiplicative) is true and that the associated normalization is made. We first outline an infeasible procedure that sets performance bounds against which to measure the feasible procedures that we have introduced.

5.1. *Oracle estimation.* Suppose that an oracle has told us what c and $g_l(\cdot)$, $l \neq j$, are in the additive model and equivalently in the multiplicative model what c and $h_l(\cdot)$, $l \neq j$, are. The question is, how would we use this information to obtain a better estimator of the remaining component? We pursue a local likelihood approach to this question; this, it turns out, leads to a procedure with smaller variance than the marginal integration estimators. This approach has been discussed in Linton (2000) in the context of generalized additive regression models. Fan and Gijbels (1996) discuss the application of local partial likelihood to estimation of nonparametric proportional hazard models where the data are i.i.d. and the covariates are one dimensional. Hastie and Tibshirani (1990) discuss quasi-backfitting methods for estimating nonparametric proportional hazard models where the data are i.i.d. and the covariates are multidimensional and the covariate effect is multiplicative. Our situation is more general, and we shall not rely on the partial likelihood idea because that only works in the multiplicative case and even then it will only solve part of the problem; that is, we are also interested in the baseline hazard.

The (conditional on Y and X) log-likelihood for a counting process is $\sum_{i=1}^{n}\int_0^T \ln\lambda_i(s)\,dN_i(s) - \sum_{i=1}^{n}\int_0^T \lambda_i(s)\,ds$, where $\lambda_i(s) = \alpha(X_i(s))Y_i(s)$. Suppose that the additive model is true and that an oracle has told us what c and $g_l(\cdot)$, $l \neq j$, are. Then define the normalized local log-likelihood function

$$l_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr) \times \bigl[\ln\alpha\bigl(\theta, X_{-ji}(s)\bigr)\,dN_i(s) - \alpha\bigl(\theta, X_{-ji}(s)\bigr)Y_i(s)\,ds\bigr], \tag{14}$$

where $\alpha(\theta, x_{-j}) = \theta + c + \sum_{l\neq j}^{d} g_l(x_l)$ as before. Let $\widehat{\theta}$ maximize $l_{nj}(\theta)$ with respect to $\theta \in \Theta$, where Θ is some compact subset of $\mathbb{R}$ that contains $\theta_0 = g_j(x_j)$ and that satisfies $\inf_{\theta\in\Theta}\inf_{x_{-j}}\alpha(\theta, x_{-j}) > 0$, and let $\widetilde{g}_j^o(x_j) = \widehat{\theta}$. This estimator is not explicitly defined and is, in general, nonlinear. In the multiplicative case, we use (14) but with $\alpha(\theta, x_{-j}) = \theta\cdot c\cdot\prod_{l\neq j} h_l(x_l)$; in this case, the local log-likelihood estimator is explicitly defined; indeed, it is

$$\widetilde{h}_j^o(x_j) = \widehat{\theta} = \frac{\sum_{i=1}^{n}\int_0^T k_b(x_j - X_{ji}(s))\,dN_i(s)}{c\sum_{i=1}^{n}\int_0^T k_b(x_j - X_{ji}(s))\prod_{k\neq j} h_k(X_{ki}(s))Y_i(s)\,ds}. \tag{15}$$

Define also $\widetilde{\alpha}_A^o(x) = \sum_{j=0}^{d}\widetilde{g}_j^o(x_j) + c$ and $\widetilde{\alpha}_M^o(x) = c\prod_{j=0}^{d}\widetilde{h}_j^o(x_j)$.

The estimators $\widetilde{g}_j^o(x_j)$ and $\widetilde{h}_j^o(x_j)$ are basically one-dimensional conditional hazard smooths on the covariate process $X_j(\cdot)$, and their properties are easy to derive from existing theory like Nielsen and Linton (1995).

THEOREM 2. *Suppose that* (A1), (A3) *and* (A4) *hold and that* $n^{1/(2r+1)}b \to \gamma$ *for some* $0 < \gamma < \infty$. *Then, when the corresponding additive/multiplicative model* *is true, there exist functions* $m_j^{oA}(\cdot), \upsilon_j^{oA}(\cdot), m_j^{oM}(\cdot), \upsilon_j^{oM}(\cdot)$ *that are bounded and continuous on* I_j *such that, for any* $x_j \in I_j$,

$$n^{r/(2r+1)}\big(\widetilde{g}_j^o(x_j) - g_j(x_j)\big) \Rightarrow N\big[m_j^{oA}(x_j), \upsilon_j^{oA}(x_j)\big], \tag{16}$$

$$n^{r/(2r+1)}\big(\widetilde{h}_j^o(x_j) - h_j(x_j)\big) \Rightarrow N\big[m_j^{oM}(x_j), \upsilon_j^{oM}(x_j)\big], \tag{17}$$

$$n^{r/(2r+1)}(\widetilde{\alpha}_A^o - \alpha_A)(x) \Rightarrow N\big[m^{oA}(x), v^{oA}(x)\big], \tag{18}$$

$$n^{r/(2r+1)}(\widetilde{\alpha}_M^o - \alpha_M)(x) \Rightarrow N\big[m^{oM}(x), v^{oM}(x)\big], \tag{19}$$

where $m^{oA}(x) = \sum_{j=0}^d m_j^{oA}(x_j)$ *and* $v^{oA}(x) = \sum_{j=0}^d \upsilon_j^{oA}(x_j)$, *while* $m^{oM}(x) = \alpha(x)\sum_{j=0}^d m_j^{oM}(x_j)/h_j(x_j)$ *and* $v^{oM}(x) = \alpha^2(x)\sum_{j=0}^d \upsilon_j^{oM}(x_j)/h_j^2(x_j)$, *where*

$$\begin{aligned} \upsilon_j^{oA}(x_j) &= \gamma^{-1}\|k\|_2^2 \frac{1}{\int_{I_{-j}} (e(x)/\alpha(x))\, dx_{-j}}, \\ \upsilon_j^{oM}(x_j) &= \gamma^{-1}\|k\|_2^2 \frac{h_j^2(x_j)}{\int_{I_{-j}} \alpha(x)e(x)\, dx_{-j}}. \end{aligned} \tag{20}$$

We suppose that the variances in (20) set the standard for the two models. It follows that $\upsilon_j^{oA}(x_j) \le \upsilon_j(x_j)$ and $\upsilon_j^{oM}(x_j) \le \upsilon_j(x_j)/c^2$ by the Cauchy–Schwarz inequality. Therefore, the marginal integration procedure is inefficient relative to the oracle estimator.

5.2. *Feasible estimation.* In this section we define a feasible version of the above oracle estimators and derive their asymptotic distribution. We first define the starting point of our algorithms, which are initial consistent estimators of $g_j(x_j)$ and $h_j(x_j)$, specifically, renormalized versions of the marginal integration estimators. Thus, we take, for $j = 0, 1, \dots, d$, $\widetilde{g}_j^{[0]}(x_j) = \widehat{\alpha}_{Q_{-j}}(x_j) - \widehat{c}$, $\widetilde{h}_j^{[0]}(x_j) = \widehat{\alpha}_{Q_{-j}}(x_j)/\widehat{c}$ and $\widehat{c} = \int \widehat{\alpha}(x)\, d\widehat{Q}(x)$. We have shown that these are consistent estimates of $g_j(x_j)$, $h_j(x_j)$ and c, respectively, for any $x_j \in I_j$. Although $\widehat{\alpha}_{Q_{-j}}(x_j), \widehat{c}$ are not guaranteed to be positive everywhere, the probability of negative values decreases to 0 very rapidly. For our procedure below we should compute these quantities on the entire covariate support $\mathcal{X}_j$ except that this will cause problems because of the well-known boundary bias of local constant-type kernel smoothers. For each j and n, let $\mathcal{X}_{j,n}^{\text{in}}$ denote the interior region, so, for example, $\mathcal{X}_{0,n}^{\text{in}} = [b, T - b]$. Then define the boundary region $\mathcal{X}_{j,n}^{\text{out}}$ as the complement of $\mathcal{X}_{j,n}^{\text{in}}$ in $\mathcal{X}_j$. We trim out the boundary region and average over interior points only; specifically, we define $\widetilde{g}_j^{[0]}(x_j), \widetilde{h}_j^{[0]}(x_j)$ as above for any $x_j \in \mathcal{X}_{j,n}^{\text{in}}$ but $\widetilde{g}_j^{[0]}(x_j), \widetilde{h}_j^{[0]}(x_j) = 0$ for any $x_j \in \mathcal{X}_{j,n}^{\text{out}}$. The results reported in Theorem 1 continue to hold when $I_j = \mathcal{X}_{j,n}^{\text{in}}$.

In the additive case, for each $it = 0, 1, \dots,$ define the estimated normalized local likelihood function

$$\widetilde{l}_{nj}^{[it+1]}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\big(x_j - X_{ji}(s)\big) \times \big[\ln \widetilde{\alpha}^{[it]}\big(\theta, X_{-ji}(s)\big)\,dN_i(s) - \widetilde{\alpha}^{[it]}\big(\theta, X_{-ji}(s)\big)Y_i(s)\,ds\big],$$

where $\widetilde{\alpha}^{[it]}(\theta, x_{-j}) = \theta + c + \sum_{l\neq j}^{d} \widetilde{g}_l^{[it]}(x_l)$. For each $it = 0, 1, \dots,$ let $\widetilde{g}_j^{[it+1]}(x_j) = \widehat{\theta}$ maximize $\widetilde{l}_n^{[it+1]}(\theta)$ with respect to $\theta \in \Theta$. In the multiplicative model, define for each j and x_j the following updated estimator:

$$\widetilde{h}_j^{[it+1]}(x_j) = \frac{\sum_{i=1}^{n}\int_0^T k_b(x_j - X_{ji}(s))\,dN_i(s)}{\widehat{c}\sum_{i=1}^{n}\int_0^T k_b(x_j - X_{ji}(s))\prod_{k\neq j}\widetilde{h}_k^{[it]}(X_{ki}(s))Y_i(s)\,ds}, \tag{21}$$

where $it = 0, 1, \dots.$ We have the following result.

THEOREM 3. *Suppose that all the conditions of Theorem* 1 *apply. Then there exist bounded continuous functions* $b_{Ak}^{[m]}(\cdot)$ *and* $b_{Mk}^{[m]}(\cdot)$, $k = 0, 1, \dots, d$, *such that*

$$n^{r/(2r+1)}\{\widetilde{g}_j^{[m]}(x_j) - \widetilde{g}_j^{o}(x_j)\} \to {}_p b_{Aj}^{[m]}(x_j) \qquad \textit{when } (3) \textit{ is true},$$

$$n^{r/(2r+1)}\{\widetilde{h}_j^{[m]}(x_j) - \widetilde{h}_j^{o}(x_j)\} \to {}_p b_{Mj}^{[m]}(x_j) \qquad \textit{when } (4) \textit{ is true}.$$

This theorem says that the m-step estimator has the same asymptotic variance as the oracle estimator, although the biases are different. This is true for any $m \geq 1$. The number of iterations only affects the bias of the estimator and perhaps the quality of the asymptotic approximation. Thus, from a statistical point of view, one iteration from $\widetilde{g}_j^{[0]}(x_j)$ and $\widetilde{h}_j^{[0]}(x_j)$ seems to be all that is needed. This result is similar to what is known in the parametric case, that is, that one step from an initial root-n consistent estimator is asymptotically equivalent to the full maximum likelihood (or more generally optimization) estimator; see Bickel (1975).

APPENDIX

For two random variables X_n, Y_n, we say that $X_n \simeq Y_n$ whenever $X_n = Y_n(1 + o_p(1))$.

Preliminary results. We first establish an exponential inequality, which is a version of Bernstein's inequality for sums of independent martingales. This is used in establishing the uniform convergence of $\widehat{\alpha}$, which is the third result of this section.

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability triple and let $\{\mathcal{F}_t\}_{t\geq 0}$ be a filtration satisfying the *usual conditions*. Consider n independent martingales $M_1, \dots, M_n$. Let $V_{2,i}$ be the predictable variation of M_i and let $V_{m,i}$ be the mth-order variation process of M_i, $i = 1, \dots, n$, $m = 3, 4, \dots.$

LEMMA 1. *Fix $0 < T \leq \infty$ and suppose that for some $\mathcal{F}_T$-measurable random variable $R_n^2(T)$ and some constant $0 < K < \infty$, one has $\sum_{i=1}^n V_{m,i}(T) \leq (m!/2)K^{m-2}R_n^2(T)$. Then, for all $a > 0$, $b > 0$,*

$$\Pr\left(\sum_{i=1}^n M_i(T) \geq c \text{ and } R_n^2(T) \leq d^2\right) \leq \exp\left[-\frac{c^2}{2(cK+d^2)}\right]. \tag{22}$$

PROOF. Define for $0 < \lambda < 1/K$, $i = 1, \dots, n$, $Z_i(t) = \lambda M_i(t) - S_i(t)$, $t \geq 0$, where S_i is the compensator of

$$W_i = \tfrac{1}{2}\lambda^2\langle M_i^c, M_i^c\rangle + \sum_{s\leq\cdot}\bigl(\exp[\lambda|\Delta M_i(s)|] - 1 - \lambda|\Delta M_i(s)|\bigr).$$

Then $\exp Z_i$ is a supermartingale, $i = 1, \dots, n$ [see the proof of Lemma 2.2 in van de Geer (1995)]. So $E\exp Z_i(T) \leq 1$, $i = 1, \dots, n$. But then also $E\exp[\sum_{i=1}^n Z_i(T)] \leq 1$. One easily verifies that $\sum_{i=1}^n S_i(T) \leq \lambda^2 R_n^2(T)/2(1-\lambda K)$. So on the set $A = \{\sum_{i=1}^n M_i(T) \geq c$ and $R_n^2(T) \leq d^2\}$, one has $\exp[\sum_{i=1}^n Z_i(T)] \geq \exp[\lambda c - \lambda^2 d^2/2(1-\lambda K)]$. Therefore, $\Pr(A) \leq \exp[-\lambda c + \lambda^2 d^2/2(1-\lambda K)]$. The result follows by choosing $\lambda = c/(d^2 + Kc)$. □

This result is formulated for fixed T, and K may depend on T and n. If the conditions of Lemma 1 hold for all T, n, then it can be extended to stopping times [see Section 8.2 in van de Geer (2000) for related results].

In the next lemma, we assume as in the main text that T is fixed and finite and write $\int = \int_0^T$. We also assume that the $\Lambda_i^n(t)$ exist and are bounded by a (nonrandom) constant $\overline{\Lambda}$ for all $1 \leq i \leq n$ and $0 \leq t \leq T$.

LEMMA 2. *Let Θ be a bounded subset of $\mathbb{R}^{d+1}$ and, for each $\theta \in \Theta$, consider independent predictable functions $g_{1,\theta}, \dots, g_{n,\theta}$. Suppose that for some constants L_n, K_n, and $\rho_n \geq 1$, we have*

$$\begin{aligned} |g_{i,\theta}(t) - g_{i,\tilde\theta}(t)| &\leq L_n|\theta - \tilde\theta| && \text{for all } \theta, \tilde\theta \in \Theta \text{ and all } i \geq 1 \text{ and } t \geq 0,\\ |g_{i,\theta}(t)| &\leq K_n && \text{for all } \theta \in \Theta \text{ and all } i \geq 1 \text{ and } t \geq 0, \end{aligned} \tag{23}$$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \int |g_{i,\theta}(t)|^2\,dt &\leq \rho_n^2 && \text{for all } \theta \in \Theta \text{ and all } n > 1,\\ L_n &\leq n^\nu && \text{for all } n > 1 \text{ and some } \nu < \infty \end{aligned} \tag{24}$$

and

$$K_n \leq \sqrt{\frac{n}{\log n}}\rho_n \qquad \text{for all } n > 1. \tag{25}$$

Then, for some constant c_0, *we have, for all* $C \geq c_0$ *and* $n > 1$,

$$\Pr\left(\sup_{\theta\in\Theta} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\int g_{i,\theta}\, d(N_i^{(n)} - \Lambda_i^{(n)})\right| \geq C\rho_n\sqrt{\log n}\right) \leq c_0 \exp\left[-\frac{C\log n}{c_0}\right].$$

PROOF. From Lemma 1, we know that, for each $\theta \in \Theta$, $a > 0$ and $R > 0$,

$$\begin{aligned}&\Pr\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\int g_{i,\theta}\, d(N_i^{(n)} - \Lambda_i^{(n)})\right| \geq a \text{ and } \frac{1}{n}\sum_{i=1}^{n}\int g_{i,\theta}^2\, d\Lambda_i^{(n)} \leq R^2\right)\\ &\quad\leq 2\exp\left[-\frac{a^2}{2(aK_n n^{-1/2} + R^2)}\right].\end{aligned} \tag{26}$$

Let $\varepsilon > 0$ to be chosen later and let $\{\theta_1, \ldots, \theta_N\} \subset \Theta$ be such that, for each $\theta \in \Theta$, there is a $j(\theta) \in 1, \ldots, N$, such that $|\theta - \theta_{j(\theta)}| \leq \varepsilon$. Then, by the Lipschitz condition (23), one has $(1/\sqrt{n})|\sum_{i=1}^{n}\int(g_{i,\theta} - g_{i,\theta_{j(\theta)}})\, d(N_i^{(n)} - \Lambda_i^{(n)})| \leq \sqrt{n}L_n\varepsilon(1 + \bar{\Lambda})$, where $\bar{\Lambda}$ is an upper bound for $\Lambda_i^{(n)}(t)$, $1 \leq i \leq n$, $n \geq 1$, $t \geq 0$.

Now, in (26), take $a = C\rho_n\sqrt{\log n}/2$ and $R_n^2 = \rho_n^2\bar{\lambda}$, with $\bar{\lambda}$ an upper bound for $\lambda_i^{(n)}(t)$, $1 \leq i \leq n$, $n \geq 1$, $t \geq 0$. Moreover, take $\varepsilon = a/(\sqrt{n}L_n(1 + \bar{\Lambda}))$. With these values, we find

$$\begin{aligned}&\Pr\left(\sup_{\theta\in\Theta} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\int g_{i,\theta}\, d(N_i^{(n)} - \Lambda_i^{(n)})\right| \geq C\rho_n\sqrt{\log n}\right)\\ &\quad= \Pr\left(\sup_{\theta\in\Theta} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\int g_{i,\theta}\, d(N_i^{(n)} - \Lambda_i^{(n)})\right| \geq 2a\right)\\ &\quad\leq \Pr\left(\max_{j=1,\ldots,N} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\int g_{i,\theta_j}\, d(N_i^{(n)} - \Lambda_i^{(n)})\right| \geq a\right)\\ &\quad\leq 2\exp\left[\log N - \frac{a^2}{2(aK_n n^{-1/2} + \rho_n^2\bar{\lambda})}\right].\end{aligned}$$

Because Θ is a bounded, finite-dimensional set, we know that, for some constant c_1, $\log N \leq c_1\log(1/\varepsilon)$. By our choice, $\varepsilon = C\rho_n\sqrt{\log n}/(2\sqrt{n}L(1 + \bar{\Lambda}))$, and using condition (24), we see that, for $C \geq 1$ (say) and some constant c_2, $\log N \leq c_2 \log n$. Invoking, moreover, condition (25), we arrive at

$$\begin{aligned}&\Pr\left(\sup_{\theta\in\Theta} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\int g_{i,\theta}\, d(N_i^{(n)} - \Lambda_i^{(n)})\right| \geq C\rho_n\sqrt{\log n}\right)\\ &\quad\leq 2\exp\left[c_2\log n - \frac{C^2\rho_n^2\log n}{8(C\rho_n\sqrt{\log n}K_n n^{-1/2}/2 + \rho_n^2\bar{\lambda})}\right]\end{aligned}$$

$$
\begin{aligned}
&\leq 2\exp\left[c_2 \log n - \frac{C^2 \log n}{8(C/2+\bar{\lambda})}\right]\\
&\leq 2\exp\left[c_2 \log n - \frac{C \log n}{8}\right]\\
&\leq 2\exp\left[-\frac{C \log n}{16}\right],
\end{aligned}
$$

where in the last two steps we take $C \geq 2\bar{\lambda}$ and $C \geq 16c_2$. $\square$

Note that by the continuity of (23) and the boundedness of Θ, the statement of Lemma 2 does not give rise to measurability problems. Note, moreover, that (23)–(25) imply that K_n, ρ_n, and L_n cannot be chosen in an arbitrary manner. Most important here is that the sup-norm should not grow too fast as compared to the L_2 norm.

LEMMA 3. *Suppose that the assumptions stated in Theorem* 1 *hold. Then, for any $a = (a_0, \ldots, a_d)$ with $|a| \leq r-1$, we have*:

(a) $$\sup_{x\in I}\left|D^a\widehat{e}(x) - D^a e(x)\right| = O_P(b^{r-|a|}) + O_P\left\{\left(\frac{\log n}{nb^{d+1+2|a|}}\right)^{1/2}\right\},$$

(b) $$\sup_{x\in I}\left|D^a\widehat{\alpha}(x) - D^a \alpha(x)\right| = O_P(b^{r-|a|}) + O_P\left\{\left(\frac{\log n}{nb^{d+1+2|a|}}\right)^{1/2}\right\}.$$

PROOF. We write $D^a\widehat{e}(x) - D^a e(x) = D^a\widehat{e}(x) - ED^a\widehat{e}(x) + ED^a\widehat{e}(x) - eD^a(x)$, a decomposition into a "stochastic" part $D^a\widehat{e}(x) - ED^a\widehat{e}(x)$ and a "bias" part $ED^a\widehat{e}(x) - D^a e(x)$. Nielsen and Linton (1995) showed, for the case $a = 0$, that $ED^a\widehat{e}(x) - D^a e(x) = O(b^r)$ for any interior point x. The extension to general a just uses integration by parts and the same Taylor series expansion.

We now turn to the stochastic part of $\widehat{e}(x)$. We claim that $\sup_{x\in I}|\widehat{e}(x) - E\widehat{e}(x)| = O_P\{(\log n/nb^{d+1})^{1/2}\}$. The pointwise result (without the logarithmic factor) is given in Nielsen and Linton (1995). The uniformity (at the cost of the logarithmic factor) follows by standard arguments, the key component of which is the application of an exponential inequality like that obtained in Lemma 2. We write $\widehat{e}(x) - E\widehat{e}(x) = \sum_{i=1}^n \zeta_{n,i}^c(x)$, where $\zeta_{n,i}^c(x) = \zeta_{n,i}(x) - E\zeta_{n,i}(x)$ with $\zeta_{n,i}(x) = n^{-1}\int_0^T K_b(x - X_i(s))Y_i(s)\,ds$. Note that $\zeta_{n,i}^c(x)$ are independent and mean-zero random variables with $m_n = \sup_{x,i}|\zeta_{n,i}^c(x)| = c_1 n^{-1}b^{-(d+1)}$ for some constant c_1; thus, m_n is uniformly bounded because $nb^{d+1} \to \infty$ by assumption. Following Nielsen and Linton (1995), we have $\sigma_{ni}^2 = \mathrm{var}[\zeta_{n,i}^c(x)] \leq c_2 n^{-1}b^{-(d+1)}$ for some constant c_2. Let $\{B(x_1, \varepsilon_L), \ldots, B(x_L, \varepsilon_L)\}$ be a cover of I, where $B(x_l, \varepsilon)$ is the ball of radius ε centered at x_l. Hence, $\varepsilon_L = c_3/L$ for some constant c_3. We have, for some constant c_4,

$$\sup_{x\in I}\left|\sum_{i=1}^n \zeta_{n,i}^c(x)\right| \le \max_{1\le l\le L}\left|\sum_{i=1}^n \zeta_{n,i}^c(x_l)\right| + \max_{1\le l\le L}\sup_{x\in B(x_l,\varepsilon)}\sum_{i=1}^n |\zeta_{n,i}^c(x_l)-\zeta_{n,i}^c(x)|$$
$$\le \max_{1\le l\le L}\left|\sum_{i=1}^n \zeta_{n,i}^c(x_l)\right| + \frac{c_4\varepsilon_L}{b^{2d+2}}$$

using the differentiability of k. Provided $\varepsilon_L\sqrt{n/(b^{3d+3}\log n)}\to 0$, we have, by the Bonferroni and Bernstein inequalities,

$$\Pr\left(\sqrt{\frac{nb^{d+1}}{\log n}}\max_{1\le l\le L}\left|\sum_{i=1}^n \zeta_{n,i}^c(x_l)\right|>\lambda\right)$$
$$\le \sum_{l=1}^L \Pr\left(\left|\sum_{i=1}^n \zeta_{n,i}^c(x_l)\right|>\lambda\sqrt{\frac{\log n}{nb^{d+1}}}\right)+o(1)$$
$$\le \sum_{l=1}^L \exp\left(-\frac{\lambda^2(\log n/(nb^{d+1}))}{2c_2(1/(nb^{d+1}))+(c_1/(nb^{d+1}))\lambda\sqrt{\log n/(nb^{d+1})}}\right)$$
$$=\sum_{l=1}^L \exp(-(\log n^{\lambda^2/2c_2})).$$

By taking λ large enough, the latter probability goes to 0 fast enough to kill $L(n)=n^\kappa$ with $\kappa=1+\eta+(3d+3)/(2r+1)$ for some $\eta>0$, and this choice of L satisfies the restriction. The result for general a follows the same pattern; differentiation to order a changes K to K^a and adds an additional bandwidth factor of order $b^{-2|a|}$.

To establish (b), we first write $\widehat{\alpha}(x)=\widehat{o}(x)/\widehat{e}(x)$ and $\alpha(x)=o(x)/e(x)$, where $o(x)=\alpha(x)e(x)$. We then apply the chain rule and Lemma 3 to obtain

$$\sup_{x\in I}\left|D^a\widehat{\alpha}(x)-D^a\alpha(x)\right|$$
$$\le \kappa\sum_{|c|\le|a|}\sup_{x\in I}\left|D^c\widehat{o}(x)-D^co(x)\right|+O_P(b^{r-|a|})+O_P\left\{\left(\frac{\log n}{nb^{d+1+2|a|}}\right)^{1/2}\right\}$$

for some positive finite constant κ, and it suffices to establish the result for the numerator statistic $D^c\widehat{o}(x)-D^co(x)$ only. Again, we shall just work out the details for the case $a=0$. The bias calculation $E\widehat{o}(x)-o(x)$ is as for $E\widehat{e}(x)-e(x)$ discussed above. Therefore, it suffices to show that $\sup_{x\in I}|\widehat{o}(x)-E\widehat{o}(x)|=\sup_{x\in I}|V_n(x)|$ is the stated magnitude, where $V_n(x)=n^{-1}\sum_{i=1}^n\int_0^T K_b(x-X_i(s))\,d(N_i(s)-\Lambda_i(s))$. We now apply Lemma 2 with $g_{i,\theta}(t)=K_b(x-X_i(t))$, $\theta=x$ and $\Theta=I$. Conditions (23)–(25) hold with probability tending to 1 for some constant γ and $K_n = \gamma \cdot b^{-(d+1)}$, $L_n = \gamma \cdot b^{-2(d+1)}$ and $\rho_n^2 = \gamma \cdot b^{-(d+1)}$ by the boundedness and differentiability of the kernel. It now follows that for some constant c_0 we have, for all $C \geq c_0$ and $n > 1$, $\Pr[\sup_{x\in I} \sqrt{nb^{d+1}/\log n}|V_n(x)| \geq C] \leq c_0 \exp(-C \log n/c_0)$ as required. $\square$

PROOF OF THEOREM 1. Standard empirical process arguments give that $\nu_n(\widehat{\alpha}(x_j,\cdot)) - \nu_n(\alpha(x_j,\cdot)) \xrightarrow{p} 0$ using (A5), Lemma 3 and the fact that $r > d/2$. Thus, it suffices to work with the stochastic integrator $\widehat{Q}_{-j}$ replaced by the deterministic Q_{-j}.

Write $(\widehat{\alpha} - \alpha)(x) = (V_n(x) + B_n(x))/\widehat{e}(x)$, where $V_n(x) = n^{-1}\sum_{i=1}^n \int_0^T K_b \times (x - X_i(s))\, dM_i(s)$ and $B_n(x) = n^{-1}\sum_{i=1}^n \int_0^T K_b(x - X_i(s))[\alpha(X_i(s)) - \alpha(x)]Y_i(s)\, ds$. Therefore,

$$(\widehat{\alpha}_{Q_{-j}} - \alpha_{Q_{-j}})(x_j) = V_{Q_{-j}}(x_j) + B_{Q_{-j}}(x_j), \tag{27}$$

where $V_{Q_{-j}}(x_j) = n^{-1} \sum_{i=1}^n \int_0^T H_i^{(n)}(x_j,s)\, dM_i(s)$ and $B_{Q_{-j}}(x_j) = \int_{I_{-j}} B_n(x)/\widehat{e}(x)\, dQ_{-j}(x_{-j})$, with $H_i^{(n)}(x_j,s) = \int_{I_{-j}} K_b(x - X_i(s))/\widehat{e}(x)\, dQ_{-j}(x_{-j})$. The proof of (11) is divided into the proofs of the following two results:

$$n^{r/(2r+1)} V_{Q_{-j}}(x_j) \Rightarrow N(0, \upsilon_j(x_j)), \tag{28}$$

$$n^{r/(2r+1)} B_{Q_{-j}}(x_j) \xrightarrow{p} m_j(x_j). \tag{29}$$

PROOF OF (28). Define

$$\widetilde{h}_i^{(n)}(x_j,s) = \int_{I_{-j}} \frac{W_{ni}(x,s)}{e(x)}\, dQ_{-j}(x_{-j}),$$

$$\widehat{h}_i^{(n)}(x_j,s) = \int_{I_{-j}} \frac{W_{ni}(x,s)}{\widehat{e}(x)}\, dQ_{-j}(x_{-j})$$

and

$$\ddot{h}_i^{(n)}(x_j,s) = \int_{I_{-j}} \frac{W_{ni}(x,s)}{\widehat{e}_{-i}(x)}\, dQ_{-j}(x_{-j}),$$

where $\widehat{e}_{-i}(x) = n^{-1}\sum_{j\neq i} \int_0^T K_b(x - X_j(s))Y_j(s)\, ds$ is the leave-one-out exposure estimator, while $W_{ni}(x,s) = (b/n)^{1/2} K_b(x - X_i(s))$. Then define

$$(nb)^{1/2}\widetilde{V}_{Q_{-j}}(x_j) = \sum_{i=1}^n \int_0^T \widetilde{h}_i^{(n)}(x_j,s)\, dM_i(s).$$

The proof of (28) is given in a series of lemmas below. We approximate $V_{Q_{-j}}(x_j)$ by $\widetilde{V}_{Q_{-j}}(x_j)$ and then apply a martingale central limit theorem to this quantity. Lemma 4 gives the CLT for $\widetilde{V}_{Q_{-j}}(x_j)$, while Lemmas 5 and 6 show that the remainder terms are of smaller order.

LEMMA 4. *$(nb)^{1/2}\widetilde{V}_{Q_{-j}}(x_j) \Rightarrow N(0, \upsilon_j^*(x_j))$, where $\upsilon_j^*(x_j) = \gamma \cdot \upsilon_j(x_j)$.*

PROOF. Since the $\widetilde{h}_i^{(n)}$ processes are predictable, asymptotic normality follows by an application of Rebolledo's central limit theorem for martingales [see Proposition 1 of Nielsen and Linton (1995)]. Specifically, we must show that, for all $\varepsilon > 0$,

$$\sum_{i=1}^{n} \int_0^T \{\widetilde{h}_i^{(n)}(x_j, s)\}^2 \mathbb{1}(|\widetilde{h}_i^{(n)}(x_j, s)| > \varepsilon)\, d\langle M_i\rangle(s) \xrightarrow{p} 0, \tag{30}$$

$$\sum_{i=1}^{n} \int_0^T \{\widetilde{h}_i^{(n)}(x_j, s)\}^2\, d\langle M_i\rangle(s) \xrightarrow{p} \upsilon_j^*(x_j), \tag{31}$$

where $\langle M\rangle$ is the quadratic variation of a process M, in our case $\langle M_i\rangle(s) = \Lambda_i(s) = \alpha(s, Z_i(s))Y_i(s)$. We make a further approximation of $\widetilde{h}_i^{(n)}(x_j, s)$ by

$$\overline{h}_i^{(n)}(x_j, s) = n^{-1/2} b^{-1/2} k\left(\frac{x_j - X_{ji}(s)}{b}\right) \frac{q_{-j}(X_{-ji}(s))}{e(x_j, X_{-ji}(s))},$$

which is valid because $\sum_{i=1}^n \int_0^T [\{\widetilde{h}_i^{(n)}(x_j, s)\}^2 - \{\overline{h}_i^{(n)}(x_j, s)\}^2]\, d\langle M_i\rangle(s) \xrightarrow{p} 0$. Then we have

$$\begin{aligned}
&\sum_{i=1}^{n} \int_0^T \{\overline{h}_i^{(n)}(x_j, s)\}^2\, d\langle M_i\rangle(s) \\
&\qquad \xrightarrow{p} E\left[\int_0^T \frac{1}{b} k^2\left(\frac{x_j - X_{ji}(s)}{b}\right) \frac{q_{-j}^2(X_{-ji}(s))}{e^2(x_j, X_{-ji}(s))} \alpha(s, Z_i(s)) Y_i(s)\, ds\right]
\end{aligned}$$

by the law of large numbers for independent random variables. The above expectation is approximately equal to $\upsilon_j(x_j)$, by Fubini's theorem, a change of variables and dominated convergence. The proof of (30) follows because $\sup_{s\in[0,T]} |\widetilde{h}_i^{(n)}(x_j, s)| \le \overline{k}/\sqrt{nb}$ for some constant $\overline{k} < \infty$. $\square$

To complete the proof of (28), we now must show that

$$(nb)^{1/2}\{\widetilde{V}_{Q_{-j}}(x_j) - V_{Q_{-j}}(x_j)\} \xrightarrow{p} 0. \tag{32}$$

By the triangle inequality,

$$\begin{aligned}
&(nb)^{1/2}|\widetilde{V}_{Q_{-j}}(x_j) - V_{Q_{-j}}(x_j)| \\
&\quad \le \left|\sum_{i=1}^{n} \int_0^T \widehat{h}_i^{(n)}(x_j, s)\, dM_i(s) - \sum_{i=1}^{n} \int_0^T \ddot{h}_i^{(n)}(x_j, s)\, dM_i(s)\right| \\
&\qquad + \left|\sum_{i=1}^{n} \int_0^T \ddot{h}_i^{(n)}(x_j, s)\, dM_i(s) - \sum_{i=1}^{n} \int_0^T \widetilde{h}_i^{(n)}(x_j, s)\, dM_i(s)\right|.
\end{aligned}$$

Therefore, it suffices to show that each of these terms goes to 0 in probability. This is shown in Lemmas 5 and 6 below.

LEMMA 5.

$$\sum_{i=1}^{n}\int_0^T \widehat{h}_i^{(n)}(x_j,s)\,dM_i(s) - \sum_{i=1}^{n}\int_0^T \ddot{h}_i^{(n)}(x_j,s)\,dM_i(s) \xrightarrow{p} 0. \tag{33}$$

PROOF. By the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\big|\widehat{h}_i^{(n)}(x_j,s) - \ddot{h}_i^{(n)}(x_j,s)\big| \\
&\quad = \left|\int_{I_{-j}} W_{ni}(x,s)\frac{\widehat{e}_i(x)-\widehat{e}(x)}{\widehat{e}(x)\widehat{e}_{-i}(x)}\,dQ_{-j}(x_{-j})\right| \\
&\quad \le \frac{[\int_{I_{-j}} W_{ni}^2(x,s)\,dQ_{-j}(x_{-j})\cdot\int_{I_{-j}}\{\widehat{e}_{-i}(x)-\widehat{e}(x)\}^2\,dQ_{-j}(x_{-j})]^{1/2}}{\inf_{x\in I}|\widehat{e}(x)\widehat{e}_{-i}(x)|},
\end{aligned}$$

where $\widehat{e}_{-i}(x)-\widehat{e}(x) = n^{-1}\int_0^T K_b\{x - X_i(t)\}Y_i(t)\,dt$. By straightforward bounding arguments and Lemma 3, we can show that $\sup_{0\le s\le T}|\int_{I_{-j}} W_{ni}^2(x,s)\times dQ_{-j}(x_{-j})| = O_P(n^{-1}b^{-(d+1)})$, $\int_{I_{-j}}\{\widehat{e}_{-i}(x)-\widehat{e}(x)\}^2\,dQ_{-j}(x_{-j}) = O_P(n^{-2}\times b^{-(d+1)})$ and $\inf_{x\in I}|\widehat{e}(x)\widehat{e}_{-i}(x)| \ge \varepsilon + o_p(1)$ for some $\varepsilon > 0$. It follows that

$$\begin{aligned}
&\left|\sum_{i=1}^{n}\int_0^T \widehat{h}_i^{(n)}(x_j,s)\,dM_i(s) - \sum_{i=1}^{n}\int_0^T \ddot{h}_i^{(n)}(x_j,s)\,dM_i(s)\right| \\
&\qquad \le n O_P\left(\frac{1}{nb^{(d+1)/2}}\right)O_P\left(\frac{1}{n^{1/2}b^{(d+1)/2}}\right),
\end{aligned}\tag{34}$$

which is $o_P(1)$ because $nb^{2(d+1)}\to\infty$. $\square$

LEMMA 6.

$$\sum_{i=1}^{n}\int_0^T \ddot{h}_i^{(n)}(x_j,s)\,dM_i(s) - \sum_{i=1}^{n}\int_0^T \widetilde{h}_i^{(n)}(x_j,s)\,dM_i(s) \xrightarrow{p} 0. \tag{35}$$

PROOF. We write

$$\begin{aligned}
\overline{M}_t &= \overline{M}_{t1} + \overline{M}_{t2} + \overline{M}_{t3} \\
&= \sum_{i=1}^{n}\int_0^T\left\{\int_{I_{-j}} W_{ni}(x,s)\frac{e(x)-E(\widehat{e}_{-i}(x))}{e^2(x)}\,dQ_{-j}(x_{-j})\right\}dM_i(s) \\
&\quad + \sum_{i=1}^{n}\int_0^T\left\{\int_{I_{-j}} W_{ni}(x,s)\frac{E(\widehat{e}_{-i}(x))-\widehat{e}_{-i}(x)}{e^2(x)}\,dQ_{-j}(x_{-j})\right\}dM_i(s) \\
&\quad + \sum_{i=1}^{n}\int_0^T\left\{\int_{I_{-j}} W_{ni}(x,s)\frac{\{e(x)-\widehat{e}_{-i}(x)\}^2}{e^2(x)\widehat{e}_{-i}(x)}\,dQ_{-j}(x_{-j})\right\}dM_i(s).
\end{aligned}\tag{36}$$

We first examine $\overline{M}_{t1}$. We have $\{E[\widehat{e}_{-i}(x)] - e(x)\}/e^2(x) = b^r \gamma_n(x)$ for some bounded continuous function γ_n, and hence

$$\begin{aligned}\int_{I_{-j}} & W_{ni}(x,s)\big(E_i[\widehat{e}_{-i}(x)] - e(x)\big)e^{-2}(x)\,dQ_{-j}(x_{-j}) \\ &= n^{-1/2}b^{-1/2}b^r k\bigg(\frac{x_j - X_{ji}(s)}{b}\bigg)\gamma_n^*\big(x_j, X_{-ji}(s)\big)\end{aligned}$$

for some bounded continuous function γ_n^*. Therefore,

$$\overline{M}_{t1} \simeq \frac{b^r}{\sqrt{nb}}\sum_{i=1}^n \int_0^T \gamma_n^*\big(x_j, X_{-ji}(s)\big)k\bigg(\frac{x_j - X_{ji}(s)}{b}\bigg)\,dM_i(s) = O_p(b^r),$$

which follows by the same arguments used in the proof of Theorem 1 of Nielsen and Linton (1995) because this term is like the normalized stochastic part of a one-dimensional kernel smoother multiplied by b^r. Therefore, $\overline{M}_{t1} = o_p(1)$.

The term $\overline{M}_{t3}$ in (36) is handled by direct methods using the uniform convergence of $\widehat{e}_{-i}(x)$, which follows from Lemma 3. Thus, $\overline{M}_{t3} = O_P(n^{-1/2}\times b^{-3(d+1)/2}) + O_P(n^{1/2}b^{2r-(d+1)/2})$.

We now deal with the stochastic term $\overline{M}_{t2}$, which is of the form $\overline{M}_{t2} = \sum_{i=1}^n \int_0^T h_i^{(n)}(u)\,dM_i(u)$, where the M_i process is a martingale, but $h_i^{(n)}(u)$ is not a predictable process according to the usual definition. Therefore, we must use the argument developed in Nielsen (1999) and Linton, Nielsen and van de Geer [(2001), Lemma 4] to solve this "predictability problem." Let

$$\begin{aligned}h_i^{(n)}(u) &= \int_{I_{-j}} W_{ni}(x,u)\frac{\widehat{e}_{-i}(x) - E[\widehat{e}_{-i}(x)]}{e^2(x)}\,dQ_{-j}(x_{-j}) \\ &= \sum_{l=1,\ l\neq i}^n \{a_{nil}(u) - E_i a_{nil}(u)\},\end{aligned} \tag{37}$$

where E_i denotes conditional expectation given $X_i(u)$, while

$$\begin{aligned}a_{nil}(u) &= (n^3 b^{-1})^{-1/2} \\ &\quad\times \int_{I_{-j}}\int_0^T Y_l(s)\frac{K_b(x - X_i(u))K_b(x - X_l(s))}{e^2(x)}\,ds\,dQ_{-j}(x_{-j}).\end{aligned}$$

Let also $h_{i,j}^{(n)}(u) = \sum_{l=1,l\neq i,j}^n \{a_{nil}(u) - E_i a_{nil}(u)\}$. We show that

$$\begin{aligned}\sum_{i=1}^n E\int_0^T \big\{h_i^{(n)}(u)\big\}^2\,d\Lambda_i(u) &\leq \frac{c'}{nb^{d+1}}\frac{1}{nb}\sum_{i=1}^n\int_0^T Ek^2\bigg(\frac{x_j - X_{ji}(u)}{b}\bigg)\,d\Lambda_i(u) \\ &= O(n^{-1}b^{-(d+1)}) = o_p(1),\end{aligned}$$

because $nb^{(d+1)} \to \infty$. Furthermore, $h_i^{(n)}(u) - h_{i,j}^{(n)}(u) = a_{nij}(u) - E_i a_{nij}(u)$, so that similar arguments show that

$$E \int_0^T \{h_i^{(n)}(u) - h_{i,j}^{(n)}(u)\}^2 \, d\Lambda_i(u) \le O(n^{-3} b^{-(d+1)}).$$

Applying Lemma 4 of Linton, Nielsen and van de Geer (2001), we have established that $E[\overline{M}_{t2}^2] = o(1)$, as required. This concludes the proof of (35). □

PROOF OF (29). We have

$$B_{Q_{-j}}(x_j) = \int_{I_{-j}} \frac{B_n(x)}{e(x)} \, dQ_{-j}(x_{-j}) + \int_{I_{-j}} B_n(x) \frac{\widehat{e}(x) - e(x)}{\widehat{e}(x) e(x)} \, dQ_{-j}(x_{-j}),$$

where, by the uniform convergence result of Lemma 3(a),

$$\begin{aligned} &\left| \int_{I_{-j}} B_n(x) \frac{\widehat{e}(x) - e(x)}{\widehat{e}(x) e(x)} \, dQ_{-j}(x_{-j}) \right| \\ &\qquad \le \frac{\sup_{x_{-j} \in I_{-j}} |B_n(x)| \sup_{x_{-j} \in I_{-j}} |\widehat{e}(x) - e(x)|}{\inf_{x_{-j} \in I_{-j}} |\widehat{e}(x) e(x)|} \\ &\qquad = O_P(b^r) O_P(b^r) = o_P(b^r). \end{aligned}$$

The term $\int_{I_{-j}} (B_n(x)/e(x)) \, dQ_{-j}(x_{-j})$ is handled by Taylor's expansion. Specifically, we show using (A4) and the fact that x is an interior point of $\mathcal{X}$ that

$$E\left[\int_{I_{-j}} \frac{B_n(x)}{e(x)} \, dQ_{-j}(x_{-j}) \right] = \frac{\mu_r(k)}{r!} b^r \sum_{j=0}^{d} \int_{I_{-j}} \beta_j^{(r)}(x) \, dQ_{-j}(x_{-j}) \{1 + o(1)\}$$

by continuity and dominated convergence. The variance of $\int_{I_{-j}} (B_n(x)/e(x)) \, dQ_{-j}(x_{-j})$ is of smaller order. This concludes the proof of (11). □

PROOF OF (12) AND (13). By Taylor's expansion,

$$\widehat{\alpha}_A(x) - \alpha_A(x) = \sum_{j=0}^{d} \{\widehat{\alpha}_{Q_{-j}}(x_j) - \alpha_{Q_{-j}}(x_j)\} + O_P(n^{-1/2}),$$

$$\widehat{\alpha}_M(x) - \alpha_M(x) = \frac{1}{c^d} \sum_{j=0}^{d} \{\widehat{\alpha}_{Q_{-j}}(x_j) - \alpha_{Q_{-j}}(x_j)\} \prod_{k \ne j} \alpha_{Q_{-k}}(x_k) + \delta_n,$$

where $\delta_n = O_P(n^{-1/2}) + O_P(\sum_{j=0}^{d} |\widehat{\alpha}_{Q_{-j}}(x_j) - \alpha_{Q_{-j}}(x_j)|^2)$. We next substitute in the expansions for $\widehat{\alpha}_{Q_{-j}}(x_j) - \alpha_{Q_{-j}}(x_j)$, which were obtained above. To show that $\widehat{\alpha}_{Q_{-j}}(x_j) - \alpha_{Q_{-j}}(x_j)$ and $\widehat{\alpha}_{Q_{-k}}(x_k) - \alpha_{Q_{-k}}(x_k)$ are uncorrelated, it suffices to show that the leading stochastic terms are so. We have

$$
\begin{aligned}
&\operatorname{cov}\Biggl(\sum_{i=1}^{n}\int_0^T \widetilde{h}_i^{(n)}(x_j,s)\,dM_i(s),\sum_{i=1}^{n}\int_0^T \widetilde{h}_i^{(n)}(x_k,s)\,dM_i(s)\Biggr)\\
&\quad = b\int_{\mathcal{X}}\Biggl[\int_{I_{-j}}\frac{k_b(x_j-w_j)k_b(x'_k-w_k)\prod_{m\neq j,k}k_b(x'_m-w_m)}{e(x_j,x'_{-j})}\,dQ_{-j}(x'_{-j})\\
&\qquad\times\int_{I_{-k}}\frac{k_b(x'_j-w_j)k_b(x_k-w_k)\prod_{m\neq j,k}k_b(x'_m-w_m)}{e(x_k,x'_{-k})}\,dQ_{-k}(x'_{-k})\Biggr]\\
&\qquad\times e(w)\,dw\,d\langle M_i(s)\rangle\\
&\quad \simeq b\int_{\mathcal{X}}k_b(x_j-w_j)k_b(x_k-w_k)\frac{q_{-j}(w_{-j})}{e(x_j,w_{-j})}\frac{q_{-k}(w_{-k})}{e(x_k,w_{-k})}e(w)\alpha(w)\,dw\,ds\\
&\quad = O(b).
\end{aligned}
\tag{38}
$$

The first equality follows by the independence of the processes, while the second equality follows by a change of variables and dominated convergence. Therefore, the covariance between the normalized component estimators is $O(b)$—so the covariance between the unnormalized estimators is $O(1/n)$. $\square$

PROOF OF THEOREM 2. We give the result for the additive case only because the multiplicative estimator is somewhat easier—it is explicitly defined and the proof follows directly from the results of Nielsen and Linton (1995). Let $S_n(\theta) = l_{nj}(\theta) - l_{nj}(\theta_0)$, where $\theta_0 = g_j(x_j)$. Then we show that

$$\sup_{\theta\in\Theta}|S_n(\theta)-S(\theta)| = o_p(1), \tag{39}$$

$$S(\theta) < 0 = S(\theta_0) \qquad \forall\,\theta\neq\theta_0, \tag{40}$$

where

$$S(\theta)=\int\Biggl[\ln\biggl\{\frac{\alpha(\theta,x_{-j})}{\alpha(x)}\biggr\}-\frac{\alpha(\theta,x_{-j})}{\alpha(x)}+1\Biggr]\alpha(x)e(x)\,dx_{-j}.$$

The result (39) follows from the same arguments as in Lemma 3. The result (40) follows because $S(\theta)$ is continuous in θ at θ_0 and because $\ln(x) - x + 1 < 0$ for all $x \neq 1$. It follows that at least one consistent solution $\widehat{\theta}$ exists to the pseudo-likelihood equation. By standard arguments, we obtain that

$$\widehat{\theta}-\theta_0 = -\biggl[\frac{\partial^2 S_n}{\partial\theta^2}(\theta_0)\biggr]^{-1}\frac{\partial S_n}{\partial\theta}(\theta_0)\times[1+o_p(1)],$$

where

$$\frac{\partial S_n}{\partial\theta}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr)\left[\frac{dN_i(s)}{\alpha(\theta_0, X_{-ji}(s))} - Y_i(s)\,ds\right],$$

$$\frac{\partial^2 S_n}{\partial\theta^2}(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr)\frac{dN_i(s)}{\alpha^2(\theta_0, X_{-ji}(s))}.$$

We have $(\partial S_n/\partial\theta)(\theta_0) = T_{n1} + T_{n2}$, where

$$T_{n1} = n^{-1}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr)\frac{dM_i(s)}{\alpha(\theta_0, X_{-ji}(s))}$$

satisfies the central limit theorem of Nielsen and Linton (1995) at rate $n^{-1/2}b^{-1/2}$, while

$$T_{n2} = n^{-1}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr)\left[\frac{\alpha(X_i(s))}{\alpha(\theta_0, X_{-ji}(s))} - 1\right]Y_i(s)\,ds$$

is a bias term that converges in probability after dividing through by b^r to a constant for each $x_j \in I_j$. We also have

$$\begin{aligned}
\frac{\partial^2 S_n}{\partial\theta^2}(\theta_0) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr)\frac{d\Lambda_i(s)}{\alpha^2(\theta_0, X_{-ji}(s))}\\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr)\frac{dM_i(s)}{\alpha^2(\theta_0, X_{-ji}(s))}\\
&= -\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr)\frac{d\Lambda_i(s)}{\alpha^2(\theta_0, X_{-ji}(s))} + o_p(1)\\
&= -\int \frac{e(x)}{\alpha(x)}\,dx_{-j} + o_p(1).
\end{aligned} \tag{41}$$

Together, these results imply (16). $\square$

PROOF OF THEOREM 3. We just show the argument for the additive case. First, we establish the result for $it = 1$. We have

$$\begin{aligned}
&\sup_{\theta\in\Theta}\bigl|\widetilde{l}_{nj}^{[1]}(\theta) - l_{nj}(\theta)\bigr|\\
&\qquad \le \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr)\ln\frac{\widetilde{\alpha}^{[0]}(\theta, X_{-ji}(s))}{\alpha(\theta, X_{-ji}(s))}\,dN_i(s)\right|
\end{aligned}$$

$$+\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j-X_{ji}(s)\bigr)\bigl[\widetilde{\alpha}^{[0]}\bigl(\theta,X_{-ji}(s)\bigr)\right.$$
$$\left.-\alpha\bigl(\theta,X_{-ji}(s)\bigr)\bigr]Y_i(s)\,ds\right|$$
$$=o_p(1),$$

provided $\sup_{\theta\in\Theta}\sup_{x_{-j}}|\widetilde{\alpha}^{[0]}(\theta,x_{-j})-\alpha(\theta,x_{-j})|=o_p(1)$. In fact, from Lemma 3,

$$\begin{aligned}\sup_{\theta\in\Theta}\sup_{x_{-j}}&\bigl|\widetilde{\alpha}^{[0]}(\theta,x_{-j})-\alpha(\theta,x_{-j})\bigr|\\&\le d\sup_x|\widehat{\alpha}(x)-\alpha(x)|+O_p(n^{-1/2})\\&=O_p\left(\sqrt{\frac{\log n}{nb^{d+1}}}\right)+O_p(b^r),\end{aligned}$$

which is $o_p(1)$ under our bandwidth conditions. It follows that $\widetilde{g}_j^{[1]}(x_j)=\widehat{\theta}$ is consistent. Indeed, it follows that $\sup_{x_j}|\widetilde{g}_j^{[1]}(x_j)-g_j(x_j)|=o_p(1)$.

By the same arguments as in the proof of Theorem 2, we have

$$\widehat{\theta}-\theta_0=-\left[\frac{\partial^2\widetilde{l}_{nj}^{[1]}}{\partial\theta^2}(\theta_0)\right]^{-1}\frac{\partial\widetilde{l}_{nj}^{[1]}}{\partial\theta}(\theta_0)\times[1+o_p(1)],\tag{42}$$

where

$$\begin{aligned}\frac{\partial\widetilde{l}_{nj}^{[1]}}{\partial\theta}(\theta_0)&=\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j-X_{ji}(s)\bigr)\left[\frac{dN_i(s)}{\widetilde{\alpha}^{[0]}(\theta_0,X_{-ji}(s))}-Y_i(s)\,ds\right],\\ \frac{\partial^2\widetilde{l}_{nj}^{[1]}}{\partial\theta^2}(\theta_0)&=-\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\bigl(x_j-X_{ji}(s)\bigr)\frac{dN_i(s)}{\widetilde{\alpha}^{[0]}(\theta_0,X_{-ji}(s))^2}.\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}&\left|\frac{\partial^2\widetilde{l}_{nj}^{[1]}}{\partial\theta^2}(\theta_0)+\int\frac{e(x)}{\alpha(x)}\,dx_{-j}\right|\\&\quad\le\left|\frac{\partial^2\widetilde{l}_{nj}^{[1]}}{\partial\theta^2}(\theta_0)-\frac{\partial^2 l_{nj}}{\partial\theta^2}(\theta_0)\right|+\left|\frac{\partial^2 l_{nj}}{\partial\theta^2}(\theta_0)+\int\frac{e(x)}{\alpha(x)}\,dx_{-j}\right|\\&\quad=o_p(1)\end{aligned}$$

by the uniform convergence of $\widetilde{\alpha}^{[0]}(\theta_0,X_{-ji}(s))$ to $\alpha(\theta_0,X_{-ji}(s))$ and (41).

Furthermore, we have

$$\begin{aligned}
&\frac{\partial \tilde{l}_{nj}^{[1]}}{\partial\theta}(\theta_0) - \frac{\partial l_{nj}}{\partial\theta}(\theta_0)\\
&\qquad = -\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\big(x_j - X_{ji}(s)n\big)\\
&\qquad\qquad\qquad \times \frac{[\tilde{\alpha}^{[0]}(\theta_0, X_{-ji}(s)) - \alpha(\theta_0, X_{-ji}(s))]\,dN_i(s)}{\alpha(\theta_0, X_{-ji}(s))^2}\\
&\qquad\quad + \frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\big(x_j - X_{ji}(s)\big)\\
&\qquad\qquad\qquad \times \frac{[\tilde{\alpha}^{[0]}(\theta_0, X_{-ji}(s)) - \alpha(\theta_0, X_{-ji}(s))]^2\,dN_i(s)}{\alpha(\theta_0, X_{-ji}(s))^2\tilde{\alpha}^{[0]}(\theta_0, X_{-ji}(s))}.
\end{aligned}$$

The second term is $O_p(\log n/nb^{d+1}) + O_p(b^{2r})$ by uniform convergence arguments and is $o_p(n^{-r/(2r+1)})$ under our bandwidth conditions.

We have $\tilde{\alpha}^{[0]}(\theta_0, X_{-ji}(s)) - \alpha(\theta_0, X_{-ji}(s)) = V^{[0]}(\theta_0, X_{-ji}(s)) + B^{[0]}(\theta_0, X_{-ji}(s)) + O_p(n^{-1/2})$, where $V^{[0]}(\theta_0, X_{-ji}(s)) = \sum_{l\neq j} V_{Q_{-l}}(X_{li}(s))$ and $B^{[0]}(\theta_0, X_{-ji}(s)) = \sum_{l\neq j} B_{Q_{-l}}(X_{li}(s))$. Then

$$\begin{aligned}
&\frac{\partial \tilde{l}_{nj}^{[1]}}{\partial\theta}(\theta_0) - \frac{\partial l_{nj}}{\partial\theta}(\theta_0)\\
&\quad = -\frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\big(x_j - X_{ji}(s)\big)\frac{V^{[0]}(\theta_0, X_{-ji}(s))}{\alpha(\theta_0, X_{-ji}(s))^2}\,dN_i(s)\\
&\qquad - \frac{1}{n}\sum_{i=1}^{n}\int_0^T k_b\big(x_j - X_{ji}(s)\big)\frac{B^{[0]}(\theta_0, X_{-ji}(s))}{\alpha(\theta_0, X_{-ji}(s))^2}\,dN_i(s) + o_p(n^{-r/(2r+1)})\\
&\quad \equiv -\big[V_j^{[1]}(x_j) + B_j^{[1]}(x_j)\big] + o_p(n^{-r/(2r+1)}).
\end{aligned}$$

The terms $V_j^{[1]}(x_j)$ and $B_j^{[1]}(x_j)$ are averages of the stochastic and bias terms of $\tilde{\alpha}^{[0]}$; therefore, $V_j^{[1]}$ is of smaller order than $\sum_{l\neq j} V_{Q-l}$, although $B_j^{[1]}$ is the same magnitude as $\sum_{l\neq j} B_{Q-l}$.

LEMMA 7. *Suppose that* $n^{r/(2r+1)}B_l^{[0]}(x_l) \xrightarrow{p} b_l^{[0]}(x_l)$ *for some bounded continuous functions* $b_l^{[0]}(x_l)$. *Then there exist bounded continuous functions* $b_j^{[1]}(x_j)$ *such that* $n^{r/(2r+1)}B_j^{[1]}(x_j) \xrightarrow{p} b_j^{[1]}(x_j)$.

PROOF. It suffices to show that for some $b_j^{[1]}(x_j)$ we have

$$\left|\widetilde{B}_j^{[1]}(x_j) - \frac{1}{n^{r/(2r+1)}} b_j^{[1]}(x_j)\right| = o_p(n^{-r/(2r+1)}), \tag{43}$$

$$\left|B_j^{[1]}(x_j) - \widetilde{B}_j^{[1]}(x_j)\right| = o_p(n^{-r/(2r+1)}), \tag{44}$$

where

$$\widetilde{B}_j^{[1]}(x_j) = \sum_{l \neq j} n^{-1} \sum_{i=1}^{n} \int_0^T k_b\big(x_j - X_{ji}(s)\big) \frac{\overline{B}^{[0]}(\theta_0, X_{-ji}(s))}{\alpha(\theta_0, X_{-ji}(s))^2}\, ds$$

in which $\overline{B}^{[0]}(\theta_0, X_{-ji}(s)) = \sum_{l \neq j} \overline{B}_{Q_{-l}}(X_{li}(s))$, where $\overline{B}_{Q_{-l}}(x_l) = \int_{I_{-l}} B_n(x)/e(x)\, dQ_{-l}(x_{-l})$.

The magnitude of $\widetilde{B}_j^{[1]}(x_j)$ is the same as the magnitude of $\overline{B}^{[0]}(\cdot)$, which has been shown earlier to be $O_P(b^r)$, so that (43) is evident. By the triangle inequality, we have

$$\begin{aligned}
&\left|B_j^{[1]}(x_j) - \widetilde{B}_j^{[1]}(x_j)\right| \\
&\quad \leq \left|\sum_{l \neq j} n^{-1} \sum_{i=1}^{n} \int_0^T k_b\big(x_j - X_{ji}(s)\big) \overline{B}^*_{Q_{-l}}(X_{li}(s))\, ds\right| + o_p(n^{-r/(2r+1)}),
\end{aligned}$$

where

$$\overline{B}^*_{Q_{-l}}(x_l) = -\int_{I_{-l}} \frac{B_n(x)}{e(x)} \frac{\widehat{e}(x) - e(x)}{\widehat{e}(x)}\, dQ_{-l}(x_{-l}).$$

By the Cauchy–Schwarz inequality the first term on the right-hand side is bounded by a constant times $b^{-1} \sup|\overline{B}_{Q_{-l}}(x_l)|$ times $\sup|\widehat{e}(x) - e(x)|$, which is $o_p(n^{-r/(2r+1)})$. $\square$

LEMMA 8. *We have* $V_j^{[1]}(x_j) = o_P(n^{-r/(2r+1)})$.

PROOF. Let

$$\widetilde{V}_j^{[1]}(x_j) = \sum_{l \neq j} \frac{1}{n} \sum_{i=1}^{n} \int_0^T k_b\big(x_j - X_{ji}(s)\big) \frac{\widetilde{V}_{Q_{-l}}(X_{li}(s))}{\alpha(\theta_0, X_{-ji}(s))^2}\, ds,$$

$$\overline{V}_j^{[1]}(x_j) = \sum_{l \neq j} \frac{1}{n} \sum_{i=1}^{n} \int_0^T k_b\big(x_j - X_{ji}(s)\big) \frac{\overline{V}_{Q_{-l}}(X_{li}(s))}{\alpha(\theta_0, X_{-ji}(s))^2}\, ds,$$

where

$$\overline{V}_{Q_{-j}}(x_j) = n^{-1} \sum_{i=1}^{n} \int_0^T k_b\bigl(x_j - X_{ji}(s)\bigr) \frac{q_{-j}(X_{-ji}(s))}{e(x_j, X_{-ji}(s))}\, dM_i(s).$$

Then, by the triangle inequality, $|V_j^{[1]}(x_j)| \le |\overline{V}_j^{[1]}(x_j)| + |\widetilde{V}_j^{[1]}(x_j) - V_j^{[1]}(x_j)| + |\overline{V}_j^{[1]}(x_j) - \widetilde{V}_j^{[1]}(x_j)|$. Interchanging summations, we have

$$\begin{aligned} \overline{V}_j^{[1]}(x_j) &= \frac{1}{n} \sum_{k=1}^{n} \int_0^T \sum_{l \neq j} \frac{q_{-l}(X_{-lk}(t))}{e(x_l, X_{-lk}(t))} \\ &\qquad \times \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_0^T \frac{k_b(x_j - X_{ji}(s)) k_b(X_{li}(s) - X_{lk}(t))\, ds}{\alpha(\theta_0, X_{-ji}(s))^2} \right\} dM_k(t) \qquad (45) \\ &\simeq \frac{1}{n} \sum_{k=1}^{n} \int_0^T \sum_{l \neq j} \frac{q_{-l}(X_{-lk}(t))}{e(x_l, X_{-lk}(t))} \int e\bigl(x_j, X_{lk}(t), x_{-j,l}\bigr)\, dx_{-j,l}\, dM_k(t), \end{aligned}$$

where $x_{-j,l}$ is the subvector of x that has excluded x_j and x_l. The approximation in (45) is valid by the same arguments given in (38)—namely, the covariance between different components is $O(1/n)$. See Linton (1997) for a similar calculation. It now follows that $\overline{V}_j^{[1]}(x_j) = O_p(n^{-1/2})$ as required. Finally, we show that $|\widetilde{V}_j^{[1]}(x_j) - V_j^{[1]}(x_j)| + |\overline{V}_j^{[1]}(x_j) - \widetilde{V}_j^{[1]}(x_j)| = o_p(n^{-r/(2r+1)})$ by using similar arguments to those used already. □

The proof for general *it* is by induction. First, we have established the expansion for $it = 1$. Now suppose that the expansion holds for iteration *it*. Then the expansion holds for iteration $it + 1$ by the same calculations given above for iteration one. The multiplicative case follows by somewhat more direct arguments. □

Acknowledgments. We thank Larry Brown, James Berger, John Marden and Martin Jacobsen, as well as several referees for helpful comments. We also thank the NSF and ESRC for financial support.

REFERENCES

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

AALEN, O. O. (1980). A model for nonparametric regression analysis of counting processes. *Mathematical Statistics and Probability Theory*. *Lecture Notes in Statist*. **2** 1–25. Springer, New York.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

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

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

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434.

CAI, Z. and FAN, J. (2000). Average regression surface for dependent data. *J. Multivariate Anal.* **75** 112–142.

COX, D. R. (1972). Regression models and life-tables (with discussion). *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.

DABROWSKA, D. M. (1997). Smoothed Cox regression. *Ann. Statist.* **25** 1510–1540.

FAN, J. and CHEN, J. (1999). One-step local quasi-likelihood estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 927–943.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

FAN, J., HÄRDLE, W. and MAMMEN, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.* **26** 943–971.

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

HUANG, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27** 1536–1563.

LIN, D. Y. and YING, Z. (1995). Semiparametric analysis of general additive–multiplicative hazard models for counting processes. *Ann. Statist.* **23** 1712–1734.

LINTON, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84** 469–473.

LINTON, O. B. (2000). Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory* **16** 502–523.

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

LINTON, O. B., NIELSEN, J. P. and VAN DE GEER, S. (2001). Estimating multiplicative and additive marker-dependent hazard functions by kernel methods. Available at econ.lse.ac.uk/staff/olinton/research/.

MAMMEN, E., LINTON, O. B. and NIELSEN, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490.

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. (1998). Marker dependent kernel hazard estimation from local linear estimation. *Scand. Actuar. J.* **2** 113–124.

NIELSEN, J. P. (1999). Super-efficient hazard estimation based on high-quality marker information. *Biometrika* **86** 227–232.

NIELSEN, J. P. and LINTON, O. B. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *Ann. Statist.* **23** 1735–1748.

NIELSEN, J. P. and LINTON, O. B. (1998). An optimisation interpretation of integration and backfitting estimators for separable nonparametric models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 217–222.

NIELSEN, J. P., LINTON, O. B. and BICKEL, P. J. (1998). On a semiparametric survival model with flexible covariate effect. *Ann. Statist.* **26** 215–241.

OPSOMER, J. D. and RUPPERT, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211.

SPERLICH, S., LINTON, O. B. and HÄRDLE, W. (1999). Integration and backfitting methods in additive models: Finite sample properties and comparison. *Test* **8** 419–458.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

VAN DE GEER, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.* **23** 1779–1801.

VAN DE GEER, S. (2000). *Applications of Empirical Process Theory*. Cambridge Univ. Press.

O. B. LINTON
DEPARTMENT OF ECONOMICS
LONDON SCHOOL OF ECONOMICS
HOUGHTON STREET
LONDON WC2A 2AE
UNITED KINGDOM
E-MAIL: lintono@lse.ac.uk

J. P. NIELSEN
CODAN
60 GAMMEL KONGEVEJ
DK-1790 COPENHAGEN V
DENMARK
E-MAIL: npj@codan.dk

S. VAN DE GEER
MATHEMATICAL INSTITUTE
UNIVERSITY OF LEIDEN
NIELS BOHRWEG 1
2300 RA LEIDEN
THE NETHERLANDS
E-MAIL: geer@math.leidenuniv.nl