Estimation in Additive Cox Models by Marginal Integration*

Toshio Honda†
Graduate School of Economics, Hitotsubashi University
2-1 Naka, Kunitachi, Tokyo 186-8601, JAPAN

September 20, 2004

Abstract

Assuming an additive model on the covariate effect in proportional hazards regression, we consider the estimation of the component functions. The estimator is based on the marginal integration method. Then we use a new kind of nonparametric estimator as the pilot estimator of the marginal integration. The pilot estimator is constructed by an analogy to the two-sample problems and by appealing to the principles of local partial likelihood and local linear fitting. We derive the asymptotic distribution of the marginal integration estimator of the component functions. The result of a simulation study is also given.

Key words and phrases: additive modeling, censoring time, failure time, local linear fitting, local partial likelihood, marginal integration method, two-sample estimator, proportional hazards models.

1 Introduction

Suppose that we observe the failure time T_i and the covariate vector process $X_i = \{X_i(t) \mid t \geq 0\}$ under censoring and that we want to investigate the effect of X_i on T_i by assuming a proportional hazards model.

A fully nonparametric procedure would be an ideal method when we cannot specify any parametric forms of the covariate effect in the hazard function. However, it is well

*Running head : Additive Cox Models

†e-mail : honda@econ.hit-u.ac.jp

known that fully nonparametric procedures may suffer from the curse of dimensionality. Then we need to impose some kinds of assumption on the forms of the covariate effect to alleviate the the curse of dimensionality of fully nonparametric regression. Additive modeling is among those assumptions. We consider additive modeling of the covariate effect in proportional hazards regression since it is a natural generalization of the Cox regression model in Cox (1972) and familiar in the literature. See Hastie and Tibshirani (1990b) for additive models.

Suppose that we have n independent and identical observations on $[0,\tau]$, $(\delta_i, X_i, T_i \wedge C_i)$, $i = 1,\ldots,n$, where C_i is the censoring time, $\delta_i = I(T_i \le C_i)$, and τ is a known constant. We assume that the censoring mechanism is independent. See (5.7) of Kalbfleisch and Prentice (2002). We assume that X_i is 2-dimensional for simplicity of presentation. Then $X_i(t) = (X_{1i}(t), X_{2i}(t))^T$. Let $(0,0)^T$ be the standard point of the covariate vector. We denote the transpose of a matrix A by A^T. We comment on the cases of more covariates in Section 3.

Defining $N_i(t)$ and $Y_i(t)$ by $N_i(t) = I(T_i \le t \wedge C_i)$ and $Y_i(t) = I(T_i \wedge C_i \ge t)$, we assume that there exists an appropriate filtration $\{\mathcal{F}_t \,|\, t \in [0,\tau]\}$ such that $\{N_i(t)\}$ is adapted and $\{X_i(t)\}$ and $\{Y_i(t)\}$ are predictable. Let the filtration satisfy the usual conditions as in Dabrowska (1997) and $\mathcal{F}_{t-}$ denote the smallest σ-algebra containing $\bigcup_{s<t} \mathcal{F}_s$.

Then our model is specified by assuming that the intensity process of $\{N_i(t)\}$ on $[0,\tau]$ is

$$\Lambda_i(dt) = E\{N_i(dt) \,|\, \mathcal{F}_{t-}\} = Y_i(t)\exp(\psi(X_i(t)))\lambda_0(t)dt, \tag{1.1}$$

where $\psi(\cdot)$ is an unknown smooth function and $\lambda_0(\cdot)$ is an unknown bounded function. In this papaer we impose the additivity assumption on $\psi(\cdot)$:

$$\psi(\xi) = \psi_1(\xi_1) + \psi_2(\xi_2), \tag{1.2}$$

where $\xi = (\xi_1, \xi_2)^T$, $\psi_1(0) = 0$, and $\psi_2(0) = 0$. Then $\lambda_0(\cdot)$ is the baseline hazard function. Because of the assumption of independent censoring, the intensity process of $\{I(T_i \leq t)\}$ is (1.1) with $Y_i(t)$ replaced by $I(T_i \geq t)$ and the effect of the covariate vector is still $\exp(\psi(X_i(t)))$. In addition, when we have (1.1), $\{M_i(t)\}$ defined by

$$M_i(t) = N_i(t) - \int_0^t Y_i(t) \exp(\psi(X_i(t)))\lambda_0(s)ds$$

is a martingale process with respect to $\{\mathcal{F}_t\}$. See Chapters 2-3 of Andersen et al. (1993) for the mathematical treatment of these subjects.

We estimate $\psi_1(x_1)$ for a fixed x_1 by marginal integration. Then we actually estimate $\psi_1(x_1) - \psi_1(0)$ because of the definition of our estimator. Since we impose no restriction on $\lambda_0(\cdot)$, functions in (1.2) are only identifiable up to a constant addition. Thus only differences such as $\psi_1(x_1) - \psi_1(0)$ make sense. If we estimate $\psi_1(x_1) + c$ by using observations with $|X_{1i}(t) - x_1| < a$, where a is a bandwidth, c depends on the estimation procedure and the estimate does not make sense by itself.

First we estimate $\psi(x_1, X_{2i}(0)) - \psi(0, X_{2i}(0))(= \psi_1(x_1)$ because of the additivity assumption) in our estimation procedure. We denote the estimates by $\hat{\psi}(x_1, X_{2i}(0)) - \hat{\psi}(0, X_{2i}(0))$. They are called the pilot estimates of the marginal integration estimator and explicitly defined in Section 2. Then we estimate $\psi_1(x_1)$ by

$$\hat{\psi}_1(x_1) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}(x_1, X_{2i}(0)) - \hat{\psi}(0, X_{2i}(0))). \tag{1.3}$$

We set $\hat{\psi}_1(0) = 0$. We can also define $\hat{\psi}_1(x_1)$ by

$$\hat{\psi}_1(x_1) = \int(\hat{\psi}(x_1, x_2) - \hat{\psi}(0, x_2))q(x_2)dx_2, \tag{1.4}$$

where $q(\cdot)$ is a density function satisfying some regularity conditions. (1.3) means we choose the density function of $X_2(0)$ as $q(\cdot)$ in (1.4).

As in other papers on marginal integration estimators, we show that $\hat{\psi}_1(x_1)$ has the

same rate of convergence as in the cases where X_i is 1-dimensional. We can use the above procedure to estimate $\psi_2(x_2)$ for a fixed x_2, too.

Here we comment on our pilot estimator. Fan et al. (1997) considered a fully nonparametric estimator of the covariate effects in proportional hazards models and the estimator is familiar in the literature. However, in this paper we do not use the estimator, which we call the integration estimator since it is obtained by integrating the estimated derivative function. We use a new kind of nonparametric estimator as the pilot estimator, which we call the two-sample estimator. Both of the estimators have similar asymptotic properties. See Honda (2004) for their asymptotic properties. Note that the integration estimator requires a more restrictive smoothness condition.

In order to describe the differences between the two estimators, we temporarily assume that X_i is 1-dimensional and time-independent. Then the integration estimator of $\psi(x)(= \psi(x) - \psi(0))$ is defined by

$$\hat{\psi}(x) = \int_0^x \hat{\psi}'(s)ds,$$

where $\hat{\psi}'(s)$ is the estimated derivative function. On the other hand, the two-sample estimator depends only on

$$\{(\delta_i, X_i, T_i \wedge C_i) \,|\, |X_i - x| \leq a\} \cup \{(\delta_i, X_i, T_i \wedge C_i) \,|\, |X_i| \leq a\},$$

where a is a bandwidth. We introduce a dummy variable Z_i such that

$$Z_i = \begin{cases} 1, & |X_i - x| \leq a \\ 0, & \text{otherwise} \end{cases}, \tag{1.5}$$

Then we estimate $\psi(x)$ as the coefficient of the dummy variable Z_i by appealing to the principles of local linear fitting and local partial likelihood.

Linton et al. (2003) also considered the estimation of the component functions in (1.2) by local constant fitting and the marginal integration method. They first estimate

the hazard function without any assumptions on the form of the hazard function. Note that time t is among the covariate vector in their pilot estimation. Then they obtain the estimators of the component functions by carrying out marginal integration. The asymptotic properties are given in Section 4 of Linton et al. (2003). They also proved that the m-step backfitting improves the marginal intergration method. When covariates are time-independent, time t does not have to be among covariates in the pilot estimation and $d+1$ in (A.3) on p.470 of Linton et al. (2003) will be replaced by d.

We describe some differences between their estimators and ours.

1. Their estimators are not guaranteed to be nonnegative. We have no possibility of negative estimates of $\exp(\psi_i(x_i))$.

2. They use kernels of 5th or higher order. See (A.4) on p.470 of Linton et al. (2003). Kernels of 3rd or higher order will be appropriate for time-independent covariates. Kernels of higher order are not robust to boundary effects. We just use symmetric kernels because we apply local polynomial fitting.

3. Their assumption (A.3) on p.470 of Linton et al. (2003) is about the smoothness of functions. When X_i is 2-dimensional, $\lambda_0(\cdot)$ and $\psi(\cdot)$ in (1.1) must be 5-times continuously differentiable in Linton et al. (2003). When X_i is 3-dimensional, $\lambda_0(\cdot)$ and $\psi(\cdot)$ must be 6-times continuously differentiable. If covariates are time-independent, $\psi(\cdot)$ will have to be 3-times and 5-times continuously differentiable, respectively. Then their estimators achieve the optimal rates for the smoothness assumptions. In this paper, we concentrate on the cases where $\psi(\cdot)$ is twice continuously differentiable and $\lambda_0(\cdot)$ is just bounded. We impose some assumptions on the sample path properties of X_i in Section 3. Then our estimators achieve the convergence rate of $n^{-2/5}$ and $n^{-2/5}\log n$ for 2-dimensional and 3-dimensional covariate vectors, respectively. See also Remark 3.1 in Section 3. When we deal

with time-independent covariates, we can do without Assumption A6 below, which is about the sample path properties of X_i.

Linton (1997), Fan et al. (1998), Linton (2000), Linton et al. (2003) and some other authors considered improving the marginal integration method, for example, by choosing weight functions or using marginal integration estimators as initial estimators of backfitting procedures. Backfitting is another estimating method for additive models. We also consider a one-step backfitting procedure in Remark 3.2 in Section 3. However, we have no rigorous result on the backfitting procedure at present. We just describe the procedure.

We refer to the literature on nonparametric estimation of hazard functions. As for proportional hazards models, O'Sullivan (1993) studied smoothing splines, Huang et al. (2000) applied regression splines to ANOVA models which include the model of this paper, Pons (2000) studied varying-coefficient models by following Fan et al. (1997). Fan et al. (1997) considered another nonparametric estimator of the covariate effect for the cases of parametric baseline hazard functions. Nielsen et al. (1998) considered a similar problem.

Some authors considered nonparametric or semiparametric estimation of hazard functions with no proportionality assumption on hazard functions. For example, Kooperberg et al. (1995) considered regression splines, Li and Doss (1995) studied kernel and nearest neighbor estimators, Nielsen and Linton (1995) considered kernel estimators. The results in Nielsen and Linton (1995) have been extended in Linton et al. (2003). Dabrowska (1997) studied a kind of partially linear models.

Finally we refer to the literature on additive models. As mentioned above, additive models are proposed to alleviate the curse of dimensionality of fully nonparametric procedures. There are several kinds of estimators for additive models. The marginal integration method was proposed by Newey (1994), Linton and Nielsen (1995), and Tjøstheim and

Auestad(1994). See Linton (2000) and Sperlich et al. (2002) for recent developments. As for backfitting, see Hastie and Tibshirani (1990b) for the algorithms. Hastie and Tibshirani (1990a) applied backfitting to the same problem as in this paper. However, they gave no theoretical analysis on the asymptotic properties. See Opsomer and Ruppert (1997), Opsomer (2000), and Mammen et al. (1999) for recent theoretical developments. A comprehensive review of regression splines to ANOVA modeling is given in Stone et al. (1997). See Gu (2002) for detailed expositions on smoothing splines. A review of the above methods is given in Schimek and Turlach (2000).

It is often the case that covariate vectors are sparsely observed over subsets of the region of interest. For example, let $0 < A_1 < A_2 < A_3$ and suppose that we are interested in $\psi_1(x_1) - \psi_1(0)$, where $x_1 \in (A_2, A_3)$. Our estimator has no difficulty even if we have almost no observation with $X_{1i}(t) \in (A_1, A_2)$. However, regression spline estimators may need some remedies since they usually carry out estimation on the whole region simultaneously.

In Section 2, we define the estimator and present the asymptotic properties in Theorem 2.1. The result of a simulation study is also given. Two remarks are given in Section 3. The proof of Theorem 2.1 are confined to Section 4.

2 Marginal integration estimator

In this section we define our estimator of $\psi_1(x_1)$ and present the asymptotic properties in Theorem 2.1 below. The result of a simulation study is also presented. The proof of Theorem 2.1 is given in Section 4.

Consider a rectangular region $(LB_1, UB_1) \times (LB_2, UB_2)$ in which $X_i(t)$, $0 \le t \le \tau$, takes the values. Then assume the standard point $(0,0)^T$ is inside the region and that $LB_1 < x_1 < UB_1$ and $x_1 \neq 0$.

We estimate $\psi_1(x_1)$ in (1.2) for a fixed x_1. First we stimate $\psi(x_1, X_{2i}(0)) - \psi(0, X_{2i}(0))$,

$i = 1, \ldots, n$, as defined later in (2.9), then obtain the estimator $\hat{\psi}_1(x_1)$ of $\psi_1(x_1)$ as in (1.3).

We introduce some assumptions and notations to define $\hat{\psi}_1(x_1)$. Assumption A1 is related to the kernel function and a usual one in the literature on nonparametric regression. Assumption A2 is about bandwidths a and b for $X_{1i}(t)$ and $X_{2i}(t)$, respectively. Assumption A3 is about the smoothness of $\psi(\cdot)$.

Assumption A1: The symmetric nonnegative kernel function $K(\cdot)$ is bounded and Lipschitz continuous. The support is contained in $[-1, 1]$. Besides $\int K(u)du = 1$.

Assumption A2: $a \to 0$, $b \to 0$, $(nab^2)^{-1}(\log n)^2 \to 0$, $na^9 \to 0$, and $nab^8 \to 0$.

Assumption A3: The component function $\psi_1(\xi_1)$ is twice continuously differentiable around 0 and x_1. The other component function $\psi_2(\xi_2)$ is twice continuously differentiable on (LB_2, UB_2). In addition $\psi_2(\xi_2)$ and all the derivatives are bounded. Note that $\lambda_0(t)$ is only assumed to be bounded.

We define β_0 for local polynomial fitting. The elements of $\beta_0 = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ are given by

$$\beta_1 = \psi_1(x_1),\ \beta_2 = a\psi_1'(x_1),\ \beta_3 = a\psi_1'(0),\ \beta_4 = b\psi_2'(x_2). \tag{2.1}$$

Note that β_0 depends on $x = (x_1, x_2)^T$ and that x_1 is fixed and x_2 varies with $x_2 = X_{2i}(0)$.

We define local linear fitting in our pilot estimator. When $|X_{1i}(t)| \le a$ and $|X_{2i}(t) - x_2| \le b$, we have by application of the Taylor series expansion that

$$\begin{aligned} \psi(X_{1i}(t), X_{2i}(t)) &= \frac{X_{1i}(t)}{a} a\psi_1'(0) + \psi_2(x_2) \\ &\quad + \frac{X_{2i}(t) - x_2}{b} b\psi_2'(x_2) + \mathrm{O}(a^2) + \mathrm{O}(b^2). \end{aligned} \tag{2.2}$$

Remember (1.2). When $|X_{1i}(t) - x_1| \le a$ and $|X_{2i}(t) - x_2| \le b$, we also have

$$\psi(X_{1i}(t), X_{2i}(t)) = \psi_1(x_1) + \frac{X_{1i}(t) - x_1}{a} a\psi_1'(x_1) + \psi_2(x_2) \tag{2.3}$$

$$+\frac{X_{2i}(t)-x_2}{b}b\psi_2'(x_2)+\mathrm{O}(a^2)+\mathrm{O}(b^2).$$

Hence our localized covariate vector $\tilde{X}_i(t)$ for $x=(x_1,x_2)^T$ is defined by

$$\tilde{X}_i(t)=\begin{cases}\big(0,0,\dfrac{X_{1i}(t)}{a},\dfrac{X_{2i}(t)-x_2}{b}\big)^T, & K_a(X_{1i}(t))K_b(X_{2i}(t)-x_2)>0\\ \big(1,\dfrac{X_{1i}(t)-x_1}{a},0,\dfrac{X_{2i}(t)-x_2}{b}\big)^T, & K_a(X_{1i}(t)-x_1)K_b(X_{2i}(t)-x_2)>0\\ (0,0,0,0)^T, & \text{otherwise}\end{cases},$$

where

$$K_a(\xi_1)=\frac{1}{a}K\big(\frac{\xi_1}{a}\big)\quad\text{and}\quad K_b(\xi_2)=\frac{1}{b}K\big(\frac{\xi_2}{b}\big).$$

The definition of $\tilde{X}_i(t)$ has no constant term for $\psi_2(x_2)$ since the constant term disappears in the local partial likelihood even if it is included. See Section 3 of Fan et al. (1997). The first element of $\tilde{X}_i(t)$ corresponds to the dummy variable Z_i in (1.5). If $|x_1|\le 2a$, we should modify the definition of $\tilde{X}_i(t)$. Hereafter we assume that n is large enough to have $|x_1|>2a$.

We denote the localized versions of $N_i(t)$ and $M_i(t)$ by $\tilde{N}_i(t)$ and $\tilde{M}_i(t)$, respectively. They are defined by

$$\tilde{N}_i(t)=\tilde{K}_{ab}(X_i(t))N_i(t)\quad\text{and}\quad\tilde{M}_i(t)=\tilde{K}_{ab}(X_i(t))M_i(t), \tag{2.4}$$

where

$$\tilde{K}_{ab}(X_i(t))=K_a(X_{1i}(t))K_b(X_{2i}(t)-x_2)+K_a(X_{1i}(t)-x_1)K_b(X_{2i}(t)-x_2).$$

By combining (2.2) and (2.3) and using the definitions of β_0 and $\tilde{X}_i(t)$, we have that

$$\psi(X_{1i}(t),X_{2i}(t))-\psi_2(x_2)=\beta_0^T\tilde{X}_i(t)+\mathrm{O}(a^2)+\mathrm{O}(b^2) \tag{2.5}$$

if $\tilde{K}_{ab}(X_i(t))>0$. The first term of the RHS of (2.5) represents the local linear fitting in the definition of our pilot estimator.

We define $C_{Kjk}(\eta_1,\eta_2)$, C_{Kjk}, and D_{Kj} by

$$C_{Kjk}(\eta_1,\eta_2)=\int u^jv^k\exp(\eta_1u+\eta_2v)K(u)K(v)dudv, \tag{2.6}$$

$$C_{Kjk}=C_{Kjk}(0,0),\quad D_{Kj}=\int u^jK^2(u)du.$$

We estimate β_0 by maximizing the local partial likelihood. To present the local partial likelihood, we define $S^{(j)}$, $j = 0, 1, 2$, as in Dabrowska (1997).

$$
\begin{aligned}
S^{(0)}(t,\eta) &= \sum_{i=1}^{n} Y_i(t) \exp(\eta^T \tilde{X}_i(t)) \tilde{K}_{ab}(X_i(t)), \\
S^{(1)}(t,\eta) &= \sum_{i=1}^{n} Y_i(t) \tilde{X}_i(t) \exp(\eta^T \tilde{X}_i(t)) \tilde{K}_{ab}(X_i(t)), \\
S^{(2)}(t,\eta) &= \sum_{i=1}^{n} Y_i(t) (\tilde{X}_i(t))^{\otimes 2} \exp(\eta^T \tilde{X}_i(t)) \tilde{K}_{ab}(X_i(t)),
\end{aligned} \tag{2.7}
$$

where $\eta \in R^4$ and $v^{\otimes 2}$ means vv^T.

Then the local partial likelihood $L(\beta)$ is given by

$$
L(\beta) = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau (\beta^T \tilde{X}_i(t) - \log S^{(0)}(t,\beta)) \tilde{N}_i(dt). \tag{2.8}
$$

See Section 3 of Fan et al. (1997) for the derivation of (2.8). We denote the estimator of β_0 by $\hat{\beta}$, which is defined by

$$
L(\hat{\beta}) = \max L(\beta), \tag{2.9}
$$

where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^T$ and $\hat{\beta}$ depends on $x = (x_1, x_2)^T$. We write $\hat{\psi}(x_1, x_2) - \hat{\psi}(0, x_2)$ for $\hat{\beta}_1$ to stress its dependence on x_2. The marginal integration estimator of $\psi_1(x_1)$ is defined as in (1.3).

We describe assumptions A4-6 before stating Theorem 2.1. Assumptions A4 and A6 are about the properties of X_i. When covariates are time-independent, Assumption A6 is unnecessary.

Assumption A4: The 2-dimensional covariate vector $X_i(t)$, $0 \le t \le \tau$, has the density function $f(\xi, t)$, where $\xi = (\xi_1, \xi_2)^T$. It is twice continuously differentiable with respect to ξ and no less than ϵ_1 on $((-\epsilon_2, \epsilon_2) \cup (x_1 - \epsilon_2, x_1 + \epsilon_2)) \times (LB_2, UB_2) \times [0, \tau]$ for some positive constants ϵ_1 and ϵ_2. All the derivatives and $f(\xi, t)$ are bounded on the region. We also assume that

$$
|f(\xi, t_1) - f(\xi, t_2)| < L|t_1 - t_2| \quad \text{and} \quad \Big|\frac{\partial f}{\partial \xi_j}(\xi, t_1) - \frac{\partial f}{\partial \xi_j}(\xi, t_2)\Big| < L|t_1 - t_2|,
$$

where $j = 1, 2$ and L is independent of ξ. Next we denote the marginal density functions by $f_1(\xi_1, t)$ and $f_2(\xi_2, t)$, respectively. The former is continuous on $((-\epsilon_2, \epsilon_2) \cup (x_1 - \epsilon_2, x_1 + \epsilon_2)) \times [0, \tau]$ and the latter is continuous and bounded on $(LB_2, UB_2) \times [0, \tau]$.

Assumption A5: We denote $\mathrm{E}(Y_i(t)|X_i(t) = \xi)$ by $g(\xi, t)$. Then $g(\xi, t)$ is twice continuously differentiable with respect to ξ and all the derivatives are bounded on $((-\epsilon_3, \epsilon_3) \cup (x_1 - \epsilon_3, x_1 + \epsilon_3)) \times (LB_2, UB_2) \times [0, \tau]$ for some positive constant ϵ_3. Besides $g(\xi, \tau) \geq \epsilon_4$ on the region for some positive constant ϵ_4. We also assume that

$$|g(\xi, t_1) - g(\xi, t_2)| < L|t_1 - t_2| \quad \text{and} \quad \Big|\frac{\partial g}{\partial \xi_j}(\xi, t_1) - \frac{\partial g}{\partial \xi_j}(\xi, t_2)\Big| < L|t_1 - t_2|,$$

where $j = 1, 2$ and L is independent of ξ.

Assumption A6: We assume that $X_i(T_i)$ has the bounded density. In addition,

$$\sup_{0 \leq s,t \leq \tau} \frac{|X_i(t) - X_i(s)|}{|t - s|^{\epsilon_1}} < W_i \text{ a.s.} \quad \text{and} \quad \mathrm{E}\{W_i^{\epsilon_2}\} < \infty$$

for some positive ϵ_1 and ϵ_2.

The former assumption of A6 deals with the terms like $\sum_{i=1}^n \tilde{K}_{ab}(X_i(T_i))/n$. The latter assumption implies a kind of Hölder continuity of $X_i(t)$ and deals with the terms like $\sum_{i=1}^n \tilde{K}_{ab}(X_i(t))/n$.

Here we state the main result of this paper. The proof is deferred to Section 4.

Theorem 2.1 *In addition to the setup around (1.1)- (1.2), we assume that assumptions A1-5 hold. Then we have*

$$(na)^{1/2}(\hat{\psi}_1(x_1) - \psi_1(x_1) - Bias) \to \mathrm{N}\Big(0, D_{K2} \int (\tilde{V}(x_1, x_2))^{-1} f_2^2(x_2, 0) dx_2\Big)$$

in law as $n \to \infty$, *where*

$$\begin{aligned} Bias &= \frac{a^2}{2} C_{K20}(\psi_1''(x_1) - \psi_1''(0)) + \mathrm{o}(a^2) + \mathrm{o}(b^2), \\ \tilde{V}(x) &= \int_0^\tau (W(x,t))^{-1} f(0, x_2, t) g(0, x_2, t) f(x, t) g(x, t) e^{\psi_1(x_1) + \psi_2(x_2)} \lambda_0(t) dt, \\ W(x,t) &= f(0, x_2, t) g(0, x_2, t) + f(x, t) g(x, t) e^{\psi_1(x_1)}. \end{aligned}$$

□

If $b = \mathrm{O}(a)$, the optimal bandwidth for X_{1i} is $a = cn^{-1/5}$, where c depends on $\psi_1''(x_1)$, $\psi_1''(0)$, and $\int (\tilde{V}(x_1, x_2))^{-1} f_2^2(x_2, 0) dx_2$. We can estimate $\psi_1''(x_1)$ and $\psi_1''(0)$ by local quadratic fitting as in Fan et al. (1997). We give an estimator of the asymptotic variance.

$$\frac{D_{K2}}{n} \sum_{j=1}^{n} \hat{f}_2(X_{2j}(0), 0) \Big[\int_0^\tau \Big\{ \Big(\frac{1}{n} \sum_{i=1}^{n} Y_i(t) K_a(X_{1i}(t)) K_b(X_{2i}(t) - x_2) \Big)^{-1} + \Big(\frac{1}{n} \sum_{i=1}^{n} Y_i(t) K_a(X_{1i}(t) - x_1) K_b(X_{2i}(t) - x_2) e^{\hat{\psi}_1(x_1)} \Big)^{-1} \Big\}^{-1} \times \frac{e^{\hat{\psi}_2(X_{2j}(0))} \sum_{k=1}^{n} N_k(dt)}{\sum_{k=1}^{n} Y_k(t) e^{\hat{\psi}_1(X_{1k}(t)) + \hat{\psi}_2(X_{2k}(t))}} \Big]^{-1},$$

where $\hat{f}_2(\cdot, 0)$ is a nonparametric estimator of $f_2(\cdot, 0)$. We have to use a rule of thumb for $\hat{\psi}_1(\cdot)$ and $\hat{\psi}_2(\cdot)$ in the above expression.

Theorem 2.1 implies that if $b = \mathrm{O}(a)$ and $a = cn^{-1/5}$, the rate of convergence of $\hat{\psi}_1(x_1)$ is optimal. We have no optimality criterion for b at present. We have to calculate $\hat{\beta}$ for each $(x_1, X_{2i}(0))$. If b is too small, it might be impossible to get $\hat{\beta}$ for some $(x_1, X_{2i}(0))$'s. Thus it might be good to avoid small values of b.

We give only an outline of the proof of Theorem 2.1 here.

The estimate $\hat{\beta}$ of β_0 in (2.1) is written as

$$\hat{\beta} - \beta_0 = \Big(- \frac{\partial^2 L}{\partial \beta \partial \beta^T}(\beta^*) \Big)^{-1} U(\beta_0), \tag{2.10}$$

where

$$\beta^* = \beta_0 + \theta_x(\hat{\beta} - \beta_0), \ 0 < \theta_x < 1 \quad \text{and} \quad U(\beta) = \frac{\partial L}{\partial \beta}(\beta).$$

All of $\hat{\beta}$, β_0, $L(\beta)$, and $U(\beta_0)$ depend on $x = (x_1, x_2)$. The pilot estimates in (1.3) are the first element of $\hat{\beta}$ with x_1 fixed and $x_2 = X_{2i}(0)$.

We decompose $U(\beta_0)$ as $U(\beta_0) = U_1(\beta_0) + U_2(\beta_0)$, where

$$U_1(\beta_0) = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \Big(\tilde{X}_i(t) - \frac{S^{(1)}(t, \beta_0)}{S^{(0)}(t, \beta_0)} \Big) \tilde{M}_i(dt), \tag{2.11}$$

$$U_2(\beta_0) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\Big(\tilde{X}_i(t) - \frac{S^{(1)}(t,\beta_0)}{S^{(0)}(t,\beta_0)}\Big) \times \tilde{K}_{ab}(X_i(t))e^{\psi_1(X_{1i}(t))+\psi_2(X_{2i}(t))}\lambda_0(t)dt. \tag{2.12}$$

The bias of $\hat{\psi}_1(x_1)$ comes from $U_2(\beta_0)$ and the convergence in law to the normal distribution comes from $U_1(\beta_0)$.

First we prove that

$$\hat{\beta} - \beta_0 = \mathrm{O_p}(a^2) + \mathrm{O_p}(b^2) + \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}), \tag{2.13}$$

uniformly in x_2. Next by exploiting (2.13), we replace β^* in

$$\Big(-\frac{\partial^2 L}{\partial\beta\partial\beta^T}(\beta^*)\Big)^{-1} \tag{2.14}$$

with β_0. Then we carry out another close examination of (2.14) by using the definition of β_0 in (2.1) and then deal with the summation in (1.3).

We carried out a small simulation study to see the performances of the estimator in Linton et al. (2003) and our estimator. We tried three models below, where the baseline hazard functions are constant functions and covariates are time-independent. The sample size is 600, we took $x_1 = -0.5$ and 0.5, and the repetition number is 100 in the simulation study. We used Splus for the simulation study. We mean by $Z \sim Ex(\lambda)$ that Z has the exponential distribution with mean λ^{-1}.

(1) $T_i \sim Ex(\exp(\psi_1(x_1) + \psi_2(x_2)))$ and $C_i \sim Ex(\exp(\psi_1(x_1) + \psi_2(x_2))/1.75)$, where $\psi_j(x_j) = 5\log(x_j + 2.5)$, $j = 1, 2$.

(2) $T_i \sim Ex(\exp(\psi_1(x_1) + \psi_2(x_2)))$ and $C_i \sim Ex(\exp(\psi_1(x_1) + \psi_2(x_2))/1.75)$, where $\psi_j(x_j) = 2\sin(1.57x_j)$, $j = 1, 2$.

(3) $T_i \sim Ex(\exp(\psi_1(x_1) + \psi_2(x_2)))$ and $C_i \sim Ex(\exp(\psi_1(x_1) + \psi_2(x_2))/1.75))$, where $\psi_j(x_j) = 2x_j$, $j = 1, 2$.

In (1)-(3), $X_j \sim Unif(-1, 1)$, $j = 1, 2$, where we denote the uniform distribution on (a, b) by $Unif(a, b)$.

First we define an adapted version of the estimator in Linton et al.(2003).

$$\hat{\psi}_1(x_1) = \log \hat{\alpha}_U(x_1) - \log \hat{\alpha}_U(0), \tag{2.15}$$

where

$$\hat{\alpha}_U(\xi_1) = \frac{2-2a}{\delta} \sum_{j=1}^{(2-2a)/\delta} \hat{\alpha}(\xi_1, -1+a+j\delta), \quad \hat{\alpha}(\xi_1,\xi_2) = \frac{\hat{o}(\xi_1,\xi_2)}{\hat{e}(\xi_1,\xi_2)},$$
$$\hat{o}(\xi_1,\xi_2) = \frac{1}{n}\sum_{i=1}^{n} K_a(\xi_1 - X_{1i}) K_a(\xi_2 - X_{2i})\delta_i,$$
$$\hat{e}(\xi_1,\xi_2) = \frac{1}{n}\sum_{i=1}^{n} K_a(\xi_1 - X_{1i}) K_a(\xi_2 - X_{2i}) T_i \wedge C_i.$$

We chose $\delta = 0.01$, $a = 0.2$, and $K(u) = (1-u^2)^2 I\{-1 < u < 1\}$ for Table 1. We chose $\delta = 0.01$, $a = 0.3$, and $K(u) = (1-u^2)^3(1.5 - 5.5u^2) I\{-1 < u < 1\}$ for Table 2. The second kernel is twice continuously differentiable and satisfies $\int u^2 K(u) du = 0$. The first kernel function does not satisfy the assumption on the kernel function in Linton et al.(2003). In Table 2, NA means the number of the cases where $\hat{\alpha}_U(x_1)$ or $\hat{\alpha}_U(0)$ is negative in the 100 repetitions. The means and variances in Table 2 are computed by excluding such cases.

The result on the estimator of this paper is presented in Table 3. We took $a = b = 0.2$ and $K(u) = I\{-1 < u < 1\}$. We used the uniform kernel to exploit the coxph function of Splus with β_0 as the initial value.

In addition we removed some X_{2i} in the summation of (1.3) to stabilize the performance of the estimator. See (a) and (b) below. If we do nothing, we will have some extreme estimates. It is because the estimator is based on imaginary two-samples and a numerical optimization procedure.

(a) We removed X_{2i} in the summation such that $X_{2i} \in [-1, -1+b/2]$ or $[1-b/2, 1]$.

(b) We removed X_{2i} in the summation such that $ab\sum_{j=1}^{n} K_a(x_1 - X_{1j}) K_b(X_{2i} - X_{2j})\delta_j \le 3$ or $ab\sum_{j=1}^{n} K_a(X_{1j}) K_b(X_{2i} - X_{2j})\delta_j \le 3$.

Table 1: The estimator in Linton et al. (2003) (the first kernel)

		$x_1 = 0.5$	$x_1 = -0.5$
(1)	true value	0.911	-1.116
	mean	0.876	-1.076
	variance	0.087	0.073
(2)	true value	1.414	-1.414
	mean	1.358	-1.336
	variance	0.103	0.093
(3)	true value	1.000	-1.000
	mean	0.956	-0.953
	variance	0.089	0.075

Table 2: The estimator in Linton et al. (2003) (the second kernel)

		$x_1 = 0.5$	$x_1 = -0.5$
(1)	true value	0.911	-1.116
	mean	0.700	-0.993
	variance	1.607	0.710
	NA	7	7
(2)	true value	1.414	-1.414
	mean	1.399	-1.343
	variance	0.411	0.655
	NA	8	7
(3)	true value	1.000	-1.000
	mean	0.849	-0.987
	variance	0.531	0.414
	NA	4	6

Table 3: The estimator proposed in this paper

		$x_1 = 0.5$	$x_1 = -0.5$
(1)	true value	0.911	-1.116
	mean	1.015	-1.222
	variance	0.053	0.058
(2)	true value	1.414	-1.414
	mean	1.527	-1.523
	variance	0.067	0.066
(3)	true value	1.000	-1.000
	mean	1.100	-1.094
	variance	0.060	0.055

Table 2 implies that we should be very careful when we use higher-order kernels. Some remedies as (a) and (b) above will be necessary to improve the practical performance of the estimator. The first kernel does not satisfy the assumption on the kernel function. However, the result in Table 1 is much better than expected. In Tables 1 and 3, the variances are much more serious than the biases.

3 Remarks on extensions

Theorem 2.1 shows that the proposed marginal integration estimator of $\psi_1(x_1) = \psi_1(x_1) - \psi_1(0)$ has the same rate of convergence as in the cases where X_i is 1-dimensional. We have assumed that X_i is 2-dimensional so far. We comment on the cases of more covariates in Remark 3.1. Remark 3.2 is about a one-step backfitting procedure.

Remark 3.1 We refer to the cases of more covariates. When $X_i(t) = (X_{1i}(t), X_{2i}(t), X_{3i}(t))^T$ and $\psi(\xi) = \psi_1(\xi_1) + \psi_2(\xi_2) + \psi_3(\xi_3)$, where $\xi = (\xi_1, \xi_2, \xi_3)^T$, let us use a common bandwidth b for both $X_{2i}(t)$ and $X_{3i}(t)$. Suppose that $\psi_j(\xi_j)$, $j = 1, 2, 3$, are twice continuously differentiable. Then only assumption A2 should be modified as follows:

$$a \to 0,\ b \to 0,\ (nab^4)^{-1}(\log n)^2 \to 0,\ na^9 \to 0,\ nab^8 \to 0. \tag{3.1}$$

However, this implies that only a suboptimal rate $O(n^{-2/5}\log n)$ is guaranteed for $\hat{\psi}_1(x_1)$ by the results of this paper.

If $\psi_2(\xi_2)$ and $\psi_3(\xi_3)$ are three times continuously differentiable, we can replace $o(b^2)$ with $O(b^3)$ in the asymptotic bias of the pilot estimators. Then the optimal rate for twice continuously differentiable $\psi_1(x_1)$ is achieved by taking $a = c_1 n^{-1/5}$ and $b = c_2 n^{-\alpha}\,(2/15 < \alpha < 1/5)$.

When $X_i(t) = (X_{1i}(t), X_{2i}(t), X_{3i}(t), X_{4i}(t))^T$ and the component functions of $\psi(\xi)$ other than $\psi_1(\xi_1)$ are three times continuously differentiable, only a suboptimal rate $O(n^{-2/5}\log n)$ is guaranteed by taking $a = c_1 n^{-1/5}(\log n)^{1/2}$ and $b = c_2 n^{-2/15}(\log n)^{1/3}$.

As is pointed out by several authors, the marginal integration method requires the calculation of pilot estimates and may not be applicable to the cases of covariate vectors of higher dimension. □

Remark 3.2 We consider a one-step backfitting procedure to estimate $\psi_1(x_1)$ in our setup. We conjecture that it has the same asymptotic properties as the oracle estimator which is defined as if we knew $\psi_2(\cdot)$. However, we could just give a heuristic argument. Thus we describe only the procedure and do not present any theoretical arguments on the asymptotic properties here.

First we define the oracle estimator of $\psi_1(x_1)$ of two-sample type.

We begin with the definitions of the localized covariate vector $\bar{X}_i(t)$ for x_1 and so on.

$$\bar{X}_i(t) = \begin{cases} (0, 0, X_{1i}(t)/h)^T, & |X_{1i}(t)| \le h \\ (1, (X_{1i}(t) - x_1)/h, 0)^T, & |X_{1i}(t) - x_1| \le h \\ (0, 0, 0)^T, & \text{otherwise} \end{cases}, \tag{3.2}$$

$$\bar{\beta}_0 = (\beta_1, \beta_2, \beta_3)^T = (\psi_1(x_1), h\psi_1'(x_1), h\psi_1'(0))^T,$$

$$\bar{K}_h(X_{1i}(t)) = K_h(X_{1i}(t) - x_1) + K_h(X_{1i}(t)),$$

where $K_h(\cdot) = K(\cdot/h)/h$ and h is the bandwidth.

We also define $\bar{S}^{(0)}(t,\eta)$, $\bar{M}_i$, and $\bar{N}_i$ by replacing $\tilde{X}_i(t)$, $\exp(\eta^T\tilde{X}_i(t))$, and $\tilde{K}_{ab}(X_i(t))$ by $\bar{X}_i(t)$, $\exp(\eta^T\bar{X}_i(t)+\psi_2(X_{2i}(t)))$, and $\bar{K}_h(X_{1i}(t))$, respectively in (2.4) and (2.7).

Then the local partial likelihood is (2.8) with $\tilde{X}_i$, $S^{(0)}$, and $\tilde{N}_i$ replaced with $\bar{X}_i$, $\bar{S}^{(0)}$, and $\bar{N}_i$, respectively. We denote the partial local likelihood by $\bar{L}(\beta)$. The oracle estimator of $\psi_1(x_1)$ of two sample type is defined as the first element of $\hat{\beta}$ such that

$$\bar{L}(\hat{\beta}) = \max_{\beta} L(\beta).$$

We denote the oracle estimator by $\bar{\psi}_1(x_1)$.

We will have a one-step backfitting estimator of $\psi_1(x_1)$ if we replace $\psi_2(\cdot)$ in the definition of $\bar{\psi}_1(x_1)$ with our marginal integration estimator of $\psi_2(\cdot)$, which is denoted by $\hat{\psi}_2(\cdot)$.

We conjecture from a heuristic argument that the one-step backfitting estimator has the same asymptotic properties as the oracle estimator if $a/h \to \infty$, not $\to 0$, and more is assumed on the smoothness of the model. However, we have no result on the uniformity of a necessary expression of $\hat{\psi}_2(\cdot)$. Besides we have adopted the counting process approach and the predictability of the integrands is crucial to the approach. The problem of predictability seems to be tough to tackle. We omit any further details here.

□

4 Proof of Theorem 2.1

We need two propositions and several lemmas to prove Theorem 2.1. First we establish (2.13) in Proposition 4.1. Then we derive an expression of (2.14) in Proposition 4.2 by using (2.13). Finally the proof of Theorem 2.1 is presented. The proofs of lemmas are confined to the end of this section.

Several kinds of uniformity of convergence play important roles in the proofs. For example, we assume just for a technical reason that we know $|\beta_0| < M$ uniformly in

$x = (x_1, x_2)^T$ for a very large M, where $|\cdot|$ stands for the Euclidean norm. Then $x_2 \in (LB_2, UB_2)$. In Lemma 4.1 below, the expressions uniformly hold on $\{(x_2, t, \eta) \,|\, x_2 \in (LB_2, UB_2),\ t \in [0, \tau],\ |\eta| < M\}$.

We give a remark on the boundary effect with respect to x_2 before we start to prove Theorem 2.1.

Remark 4.1 When x_2 is close to LB_2 or UB_2, the properties of $\hat{\psi}(x_1, x_2) - \hat{\psi}(0, x_2)$ are different from those with $x_2 - LB_2 > b$ and $UB_2 - x_2 > b$. However, the bias is still $\mathrm{O}(a^2) + \mathrm{O}(b^2)$ and the proportion of the observations with $X_{2i}(0) - LB_2 \le b$ or $UB_2 - X_{2i}(0) \le b$ tends to 0. In fact the boundary effect does not matter in the asymptotic properties of $\hat{\psi}_1(x_1)$. Thus we do not care about the boundary effect with respect to x_2 in the proofs to make the proofs more readable. □

Lemma 4.1 is employed to evaluate $S^{(j)}(t, \eta)$ with $\eta = \beta^*$ or β_0 in (2.10). See (2.6) and (2.7) for the definitions of C_{Kjk} and $S^{(j)}$.

Lemma 4.1 *We have the following expression of* $S^{(0)}$. *Uniformly in* x_2, η, *and* t, *where* $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)^T$,

$$
\begin{aligned}
&\frac{1}{n} S^{(0)}(t, \eta) \\
&= f(x,t)g(x,t)e^{\eta_1} C_{K00}(\eta_2, \eta_4) + f(0, x_2, t)g(0, x_2, t) C_{K00}(\eta_3, \eta_4) \\
&\quad + a\frac{\partial}{\partial x_1}(f(x,t)g(x,t))e^{\eta_1} C_{K10}(\eta_2, \eta_4) + b\frac{\partial}{\partial x_2}(f(x,t)g(x,t))e^{\eta_1} C_{K01}(\eta_2, \eta_4) \\
&\quad + a\frac{\partial}{\partial x_1}(f(0, x_2, t)g(0, x_2, t)) C_{K10}(\eta_3, \eta_4) + b\frac{\partial}{\partial x_2}(f(0, x_2, t)g(0, x_2, t)) C_{K01}(\eta_3, \eta_4) \\
&\quad + \mathrm{O_p}(a^2) + \mathrm{O_p}(b^2) + \mathrm{O_p}(\{(nab)^{-1} \log n\}^{1/2}).
\end{aligned}
$$

We denote the sum of the first and second terms of the RHS of the above expression by $A_0(x, t, \eta)$. *We have similar expressions for* $S^{(1)}/n$ *and* $S^{(2)}/n$. *We give only the sums of the first and second terms of the expressions and denote them by* $A_1(x, t, \eta)$ *and* $A_2(x, t, \eta)$, *respectively.*

$$
\begin{aligned}
A_1(x,t,\eta) &= f(x,t)g(x,t)e^{\eta_1}(C_{K00}(\eta_2,\eta_4), C_{K10}(\eta_2,\eta_4), 0, C_{K01}(\eta_2,\eta_4))^T \\
&\quad + f(0,x_2,t)g(0,x_2,t)(0,0,C_{K10}(\eta_3,\eta_4), C_{K01}(\eta_3,\eta_4))^T.
\end{aligned}
$$

$$
\begin{aligned}
A_2(x,t,\eta) &= f(x,t)g(x,t)e^{\eta_1}
\begin{pmatrix}
C_{K00}(\eta_2,\eta_4) & C_{K10}(\eta_2,\eta_4) & 0 & C_{K01}(\eta_2,\eta_4) \\
C_{K10}(\eta_2,\eta_4) & C_{K20}(\eta_2,\eta_4) & 0 & C_{K11}(\eta_2,\eta_4) \\
0 & 0 & 0 & 0 \\
C_{K01}(\eta_2,\eta_4) & C_{K11}(\eta_2,\eta_4) & 0 & C_{K02}(\eta_2,\eta_4)
\end{pmatrix} \\
&\quad + f(0,x_2,t)g(0,x_2,t)
\begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & C_{K20}(\eta_3,\eta_4) & C_{K11}(\eta_3,\eta_4) \\
0 & 0 & C_{K11}(\eta_3,\eta_4) & C_{K02}(\eta_3,\eta_4)
\end{pmatrix}.
\end{aligned}
$$

□

Lemma 4.2 is necessary to show that (2.14)$= \mathrm{O_p}(1)$ uniformly in x_2 in the proof of Proposition 4.1.

Lemma 4.2 *Uniformly in x_2 and η,*

$$
\begin{aligned}
-\frac{\partial^2 L}{\partial\beta\partial\beta^T}(\eta) &= \int_0^\tau W(x,t)e^{\psi_2(x_2)}(A_0(x,t,\eta))^{-1}A_2(x,t,\eta)\lambda_0(t)dt \\
&\quad - \int_0^\tau W(x,t)e^{\psi_2(x_2)}(A_0(x,t,\eta))^{-2}(A_1(x,t,\eta))^{\otimes 2}\lambda_0(t)dt + \mathrm{o_p}(1),
\end{aligned}
$$

where $W(x,t)$ is defined in Theorem 2.1. □

Lemma 4.3 *Uniformly in x_2,*

$$
\begin{aligned}
U_1(\beta_0) &= \mathrm{O_p}(a^2) + \mathrm{O_p}(b^2) + \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}), \\
U_2(\beta_0) &= \frac{a^2}{2}(\tilde{V}(x)C_{K20}(\psi_1''(x_1) - \psi_1''(0)), 0, 0, 0)^T + \mathrm{o_p}(a^2) + \mathrm{o_p}(b^2),
\end{aligned}
$$

where $\tilde{V}(x)$ is defined in Theorem 2.1. □

Proposition 4.1 *Unformly in x_2,*

$$
\hat{\beta} - \beta_0 = \mathrm{O_p}(a^2) + \mathrm{O_p}(b^2) + \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}).
$$

Proof) We should note that $\hat{\beta} - \beta_0 = \mathrm{o_p}(1)$, unformly in x_2. This is proved by showing the uniform convergence in probability of $L(\eta) - n^{-1}\log n \sum \tilde{N}_i(\tau)$ and evaluating the limit. We omit the details. Then the proposition follows from Lemmas 4.1-3 and the uniform convergence in probability of $\hat{\beta} - \beta_0$. □

Lemma 4.4 is easy to establish. The proof is omitted. Note that it is about a sum of i.i.d. random variables and that the symmetry of the kernel is used. The proof does not use the results of the other lemmas and propositions.

Lemma 4.4 *If $G(x,t)$ is a deterministic function and continuously differentiable with respect to x_2 and $G(x,t)$ and the derivative are uniformly bounded, then we have uniformly in x_2,*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau G(x,t)\tilde{N}_i(dt) \\
&= \int_0^\tau (f(x,t)g(x,t)G(x,t)e^{\psi_1(x_1)} + f(0,x_2,t)g(0,x_2,t)G(x,t))e^{\psi_2(x_2)}\lambda_0(t)dt \\
&\quad + \mathrm{O_p}(a^2) + \mathrm{O_p}(b^2) + \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}).
\end{aligned}$$

□

Explicit expressions of $\Omega_a(x)$ and $\Omega_b(x)$ in Proposition 4.2 are not necessary in the proof of Theorem 2.1.

Proposition 4.2

$$\begin{aligned}
(4.1)\quad &\Big(-\frac{\partial^2 L}{\partial\beta\partial\beta^T}(\beta^*)\Big)^{-1} \\
&= \begin{pmatrix} \tilde{V}^{-1}(x) & 0 & 0 & 0 \\ 0 & V^{-1}(x)C_{K20}^{-1} & 0 & 0 \\ 0 & 0 & V^{-1}(0,x_2)C_{K20}^{-1} & 0 \\ 0 & 0 & 0 & (V(x)+V(0,x_2))^{-1}C_{K02}^{-1} \end{pmatrix} \\
&\quad + a\Omega_a(x) + b\Omega_b(x) + \mathrm{O_p}(a^2) + \mathrm{O_p}(b^2) + \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}),
\end{aligned}$$

uniformly in x_2, *where*

$$V(\xi) = \int_0^\tau f(\xi,t)g(\xi,t)e^{\psi_1(\xi_1)+\psi_2(\xi_2)}\lambda_0(t)dt,\ \xi = x \text{ or } (0,x_2)^T,$$

$\Omega_a(x)$ *and* $\Omega_b(x)$ *are* 4×4 *-dimensional deterministic functions and continuously differentiable with respect to* x_2, *and every element and all the derivatives are bounded.*

Proof) Note that $-\dfrac{\partial^2 L}{\partial\beta\partial\beta^T}(\beta^*)$ is written as

$$\frac{1}{n}\sum_{i=1}^n \int_0^\tau \frac{S^{(2)}(t,\beta^*)}{S^{(0)}(t,\beta^*)}\tilde{N}_i(dt) - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \frac{(S^{(1)}(t,\beta^*))^{\otimes 2}}{(S^{(0)}(t,\beta^*))^2}\tilde{N}_i(dt). \tag{4.2}$$

Only the outline of the evaluation of the first term is given.

We apply Lemma 4.1 to $S^{(0)}$ and $S^{(2)}$, then the definition of β^* in (2.10) and Proposition 4.1 allows us to replace β^* with β_0 in (4.2). It is because the remainder parts of the RHS's are $\mathrm{O_p}(a^2)+\mathrm{O_p}(b^2)+\mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2})$ in Lemma 4.1 and we have only to take the first six terms of the RHS's into account.

Next substitute the definition of β_0 in (2.1) into $S^{(0)}(t,\beta_0)$ and $S^{(2)}(t,\beta_0)$, use Lemma 4.1 again, and apply the Taylor series expansion to the first six terms of the RHS's in Lemma 4.1 with respect to η at $(\psi_1(x_1),0,0,0)^T$. Finally from Lemma 4.4, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \int_0^\tau \frac{S^{(2)}(t,\beta^*)}{S^{(0)}(t,\beta^*)}\tilde{N}_i(dt) \\
&= V(x)\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & C_{K20} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{K02} \end{pmatrix} + V(0,x_2)\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & C_{K20} & 0 \\ 0 & 0 & 0 & C_{K02} \end{pmatrix} \\
&\quad + a\int_0^\tau \Omega_1(x,t)\lambda_0(t)dt + b\int_0^\tau \Omega_2(x,t)\lambda_0(t)dt \\
&\quad + \mathrm{O_p}(a^2)+\mathrm{O_p}(b^2)+\mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}),
\end{aligned} \tag{4.3}$$

where $\Omega_i(x,t), i=1,2,$ are 4×4-dimensional deterministic functions and continuously differentiable with respect to x_2 and every element and all the derivatives are uniformly bounded. Explicit expressions of $\Omega_i, i\geq 1$, are not necessary in the proof.

A similar expression can be derived for the second term of (4.2). We omit the details.

By combining (4.2), (4.3) and the expression of the second term of (4.2), the proposition is established. □

Lemmas 4.5-6 are used in the proof of Theorem 2.1.

Lemma 4.5 *If $G(x,t)$ is a deterministic function, continuously differentiable with respec to x_2, and Lipschitz continuous in t uniformly in x_2, and $G(x,t)$ and the derivative are uniformly bounded, then we have*

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{\frac{1}{n}\sum_{j=1}^{n} G(x_1, X_{2j}(0), t)\Big(\frac{X_{2i}(t)-X_{2j}(0)}{b}\Big)^k K_b(X_{2i}(t)-X_{2j}(0))\Big\}$$
$$\times\Big\{\Big(\frac{X_{1i}(t)-x_1}{a}\Big)^l K_a(X_{1i}(t)-x_1) + \Big(\frac{X_{1i}(t)}{a}\Big)^m K_a(X_{1i}(t))\Big\} M_i(dt) = \mathrm{O_p}((na)^{-1/2}).$$

□

where $k, l, m = 0$ or 1.

Lemma 4.6 *If $Z(x,t)$ is predictable and $Z(x,t) = \mathrm{O_p}(a^2)+\mathrm{O_p}(b^2)+\mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2})$ uniformly in x_2 and t, then we have*

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big(\frac{1}{n}\sum_{j=1}^{n} Z(x_1, X_{2j}(0), t) K_b(X_{2i}(t)-X_{2j}(0))\Big)$$
$$\times(K_a(X_{1i}(t)-x_1)+K_a(X_{1i}(t)))M_i(dt)$$
$$= \mathrm{O_p}((na)^{-1/2}[a^2+b^2+\{(nab)^{-1}\log n\}^{1/2}]).$$

□

Now we prove Theorem 2.1.

Proof of Theorem 2.1) The expressions in (2.10), Proposition 4.2, and Lemma 4.3 imply that the bias of $\hat{\psi}(x_1,x_2)-\hat{\psi}(0,x_2)$ is given by

$$\frac{a^2}{2}C_{K20}(\psi_1''(x_1)-\psi_1''(0)) + \mathrm{o}(a^2)+\mathrm{o}(b^2), \tag{4.4}$$

uniformly in x_2. The bias part of Theorem 2.1 easily follows from (4.4). The details are omitted.

The rest of the proof consists of the derivation of (4.5) and the proof of the convergence in law of the sum of the products of (4.5) and the expression in Proposition 4.2.

First we deal with $U_1(\beta_0)$ in (2.11). By applying Lemma 4.1 and the Taylor series expansion with respect to η at $(\psi_1(x_1), 0, 0, 0)^T$ as in the proof of Proposition 4.2, we obtain

$$
\begin{aligned}
(4.5)\quad U_1(\beta_0) \;=\; & \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\tilde{X}_i(t) - W^{-1}(x,t)f(x,t)g(x,t)e^{\psi_1(x_1)}(1,0,0,0)^T)\tilde{M}_i(dt) \\
& + \frac{a}{n}\sum_{i=1}^{n}\int_0^\tau \Omega_3(x,t)\tilde{M}_i(dt) + \frac{b}{n}\sum_{i=1}^{n}\int_0^\tau \Omega_4(x,t)\tilde{M}_i(dt) \\
& + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau R(x,t)\tilde{M}_i(dt),
\end{aligned}
$$

where $\Omega_i(x,t), i = 3, 4$, are defined as in (4.3) and Lipschitz continuous in t uniformly in x_2, and $R(x,t)$ is predictable and satisfies, uniformly in x_2 and t,

$$
R(x,t) = \mathrm{O_p}(a^2) + \mathrm{O_p}(b^2) + \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}).
$$

Then we consider the product of (4.1) in Proposition 4.2 and (4.5). Since $U_1(\beta_0) = \mathrm{O_p}(a^2) + \mathrm{O_p}(b^2) + \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2})$, uniformly in x_2, assumption A2 implies that the last three remainder terms of (4.1) do no affect the asymptotic distribution of $\hat{\psi}_1(x_1)$.

Next assumption A2 and Lemmas 4.5-6 imply that only the first term of the RHS of (4.5) affects the asymptotic distribution of $\hat{\psi}_1(x_1)$. Besides, from assumption A2 and Lemmas 4.3-5, we have only to take the first term of (4.1) into account.

Hence we have

$$
\begin{aligned}
& e_1^T \frac{1}{n}\sum_{j=1}^{n}\Big(-\frac{\partial^2 L}{\partial\beta\partial\beta^T}(\beta^*)\Big|_{x_2=X_{2j}(0)}\Big)^{-1}(U_1(\beta_0)|_{x_2=X_{2j}(0)}) \\
&= \frac{1}{n}\sum_{j=1}^{n}(\tilde{V}(x_1, X_{2j}(0)))^{-1}\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{\frac{f(0, X_{2j}(0), t)g(0, X_{2j}(0), t)}{W(x_1, X_{2j}(0), t)}K_a(X_{1i}(t) - x_1)
\end{aligned}
$$

$$-\frac{f(x_1, X_{2j}(0), t)g(x_1, X_{2j}(0), t)e^{\psi_1(x_1)}}{W(x_1, X_{2j}(0), t)}K_a(X_{1i}(t))\Big\}K_b(X_{2i}(t) - X_{2j}(0))M_i(dt)$$
$$+\mathrm{o_p}((na)^{-1/2}).$$

where $e_1 = (1,0,0,0)^T$ and $U_1(\beta_0)|_{x_2=X_{2j}(0)}$ means $U_1(\beta_0)$ for $(x_1, X_{2j}(0))^T$ and the notation applies to other terms.

The proof is complete from Lemma 4.7 below. We can treat the summation with respect to j as in the proof of Lemma 4.5 and the proof of Lemma 4.7 is omitted.

Lemma 4.7

$$\frac{(na)^{1/2}}{n}\sum_{j=1}^{n}(\tilde{V}(x_1, X_{2j}(0)))^{-1}\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{\frac{f(0, X_{2j}(0), t)g(0, X_{2j}(0), t)}{W(x_1, X_{2j}(0), t)}K_a(X_{1i}(t) - x_1)$$
$$-\frac{f(x_1, X_{2j}(0), t)g(x_1, X_{2j}(0), t)e^{\psi_1(x_1)}}{W(x_1, X_{2j}(0), t)}K_a(X_{1i}(t))\Big\}K_b(X_{2i}(t) - X_{2j}(0))M_i(dt)$$
$$\to \mathrm{N}(0, D_{K2}\sigma^2(x_1)),$$

in law as $n \to \infty$, *where*

$$\sigma^2(x_1) = \int (\tilde{V}(x_1, x_2))^{-1} f_2^2(x_2, 0)dx_2.$$

□

□

We begin to prove lemmas.

Proof of Lemma 4.1) We just outline the proof. We have uniformly in x_2, η, and t,

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i(t)e^{\eta^T\tilde{X}_i(t)}\tilde{K}_{ab}(X_i(t)) - \mathrm{E}(Y_i(t)e^{\eta^T\tilde{X}_i(t)}\tilde{K}_{ab}(X_i(t)))\}$$
$$= \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}). \tag{4.6}$$

This is proved by following the arguments in Masry (1996) and using Assumption A6, the monotonicity of $Y_i(t)$, the continuity of $\exp(\eta^T\tilde{X}_i(t))$ in x_2 and η, and the continuity of $\tilde{K}_{ab}(\tilde{X}_i(t))$ in x_2. The proof is easier since the observations are i.i.d. Then

just evaluate the expectation in the LHS of (4.6) by employing the symmetry of the kernel function. □

Proof of Lemma 4.2) An expression of $-\dfrac{\partial^2 L}{\partial\beta\partial\beta^T}(\eta)$ is (4.2) with β^* replaced by η. Just the outline of the proof is given.

We apply Lemma 4.1. Then Lemma 4.4 implies that we should consider only A_0, A_1, and A_2. Lemma 4.2 follows from another application of Lemma 4.4. □

Proof of Lemma 4.3) We consider $U_2(\beta_0)$ first. As in Fan et al. (1997), we have

$$
\begin{aligned}
(4.7)\qquad U_2(\beta_0) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\Big(\tilde{X}_i(t)-\frac{S^{(1)}(t,\beta_0)}{S^{(0)}(t,\beta_0)}\Big)\\
&\quad\times\tilde{K}_{ab}(X_i(t))(e^{\psi_1(X_{1i}(t))+\psi_2(X_{2i}(t))}-e^{\psi_2(x_2)+\beta_0^T\tilde{X}_i(t)})\lambda_0(t)dt.
\end{aligned}
$$

By the Taylor series expansion,

$$
\begin{aligned}
(4.8)\qquad &\tilde{K}_{ab}(X_i(t))(e^{\psi_1(X_{1i}(t))+\psi_2(X_{2i}(t))}-e^{\psi_2(x_2)+\beta_0^T\tilde{X}_i(t)})\\
&= K_a(X_{1i}(t)-x_1)K_b(X_{2i}(t)-x_2)e^{\psi_1(x_1)+\psi_2(x_2)}\\
&\qquad\times\Big\{\frac{a^2}{2}\psi_1''(x_1)\Big(\frac{X_{1i}(t)-x_1}{a}\Big)^2+\frac{b^2}{2}\psi_2''(x_2)\Big(\frac{X_{2i}(t)-x_2}{b}\Big)^2\Big\}\\
&\quad+K_a(X_{1i}(t))K_b(X_{2i}(t)-x_2)e^{\psi_2(x_2)}\\
&\qquad\times\Big\{\frac{a^2}{2}\psi_1''(0)\Big(\frac{X_{1i}(t)}{a}\Big)^2+\frac{b^2}{2}\psi_2''(x_2)\Big(\frac{X_{2i}(t)-x_2}{b}\Big)^2\Big\}\\
&\quad+(\mathrm{o}(a^2)+\mathrm{o}(b^2))\tilde{K}_{ab}(X_i(t)).
\end{aligned}
$$

From the definition of $\tilde{X}_i(t)$ and Lemma 4.1, we have uniformly in x_2 and t,

$$
\begin{aligned}
(4.9)\qquad &\tilde{X}_i(t)-\frac{S^{(1)}(t,\beta_0)}{S^{(0)}(t,\beta_0)}\\
&=\begin{cases}
\Big(-\dfrac{f(x,t)g(x,t)e^{\psi_1(x_1)}}{W(x,t)},0,\dfrac{X_{1i}(t)}{a},\dfrac{X_{2i}(t)-x_2}{b}\Big)^T+\mathrm{o_p}(1),\\
\qquad K_a(X_{1i}(t))K_b(X_{2i}(t)-x_2)>0\\
\Big(\dfrac{f(0,x_2,t)g(0,x_2,t)}{W(x,t)},\dfrac{X_{1i}(t)-x_1}{a},0,\dfrac{X_{2i}(t)-x_2}{b}\Big)^T+\mathrm{o_p}(1),\\
\qquad K_a(X_{1i}(t)-x_1)K_b(X_{2i}(t)-x_2)>0
\end{cases}.
\end{aligned}
$$

Hence as in Lemma 4.1, we have from (4.8) and (4.9), uniformly in x_2,

$$U_2(\beta_0) = \frac{a^2}{2}(\tilde{V}(x)C_{K20}(\psi_1''(x_1) - \psi_1''(0)), 0, 0, 0) + \mathrm{o_p}(a^2) + \mathrm{o_p}(b^2). \tag{4.10}$$

Note that the two terms of $b^2\psi_2''(x_2)$ cancels out each other.

Next we deal with $U_1(\beta_0)$, which is given in (2.11). The first term of $U_1(\beta_0)$ is reduced to

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \tilde{X}_i(t)\tilde{K}_{ab}(X_i(t))M_i(dt). \tag{4.11}$$

This is a sum of bounded independent random variables. The variances are evaluated by using the fact that $\{M_i(t)\}$ are martingales. Then the standard argument in nonparametric regression applies and we have uniformly in x_2,

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \tilde{X}_i(t)\tilde{K}_{ab}(X_i(t))M_i(dt) = \mathrm{O_p}(\{(nab)^{-1}\log n\}^{1/2}).$$

As for the second term of $U_1(\beta_0)$, apply Lemma 4.1 to $S^{(j)}(t,\eta)/n$ and notice that

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau |\tilde{M}_i|(dt) \le \frac{C}{n}\sum_{i=1}^{n}\tilde{K}_{ab}(X_i(T_i)) + \frac{C}{n}\sum_{i=1}^{n}\int_0^\tau \tilde{K}_{ab}(\tilde{X}_i(t))\lambda_0(dt) = \mathrm{O_p}(1),$$

uniformly in x_2 for some positive constant C. We mean the total variation of $\{M_i(t)\}$ by $\{|M_i|(t)\}$. Then the same argument as for the first term of $U_1(\beta_0)$ applies again. Hence the proof is complete. □

Proof of Lemma 4.5) We deal with the case where $k = l = m = 0$ for notational simplicity. In the same way as in Lemma 4.1, we have uniformly in x_2 and t,

$$\frac{1}{n}\sum_{j=1}^{n} G(x_1, X_{2j}(0), t)K_b(x_2 - X_{2j}(0)) - G(x_1, x_2, t)f_2(x_2, 0) = \mathrm{o_p}(1). \tag{4.12}$$

The predictable variation process of

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t G(x_1, X_{2i}(t), t)f_2(X_{2i}(t), 0)(K_a(X_{1i}(t) - x_1) + K_a(X_{1i}(t)))M_i(dt)$$

satisfies

$$\mathrm{O_p}\Big(\frac{1}{n^2}\sum_{i=1}^{n}\int_0^\tau \{K_a^2(X_{1i}(t) - x_1) + K_a^2(X_{1i}(t))\}\lambda_0(t)dt\Big). \tag{4.13}$$

(4.12) and (4.13) together with Lenglart's inequality imply Lemma 4.5. □

Proof of Lemma 4.6) The lemma follows from the evaluation of the predictable variation process and the application of Lenglart's inequality. The details are omitted. □

Acknowledgments

The author appreciates helpful comments of two referees very much.

References

[1] Andersen, P. K., Borgan, Ø, Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.

[2] Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society Series B*, **34**, 187-220.

[3] Dabrowska, D. M. (1997). Smoothed Cox regression, *The Annals of Statistics*, **25**, 1510-1540.

[4] Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression, *The Annals of Statistics*, **25**, 1661-1690.

[5] Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models, *The Annals of Statistics*, **26**, 943-971.

[6] Gu, C. (2002). *Smoothing spline ANOVA models*, Springer, New York.

[7] Hastie, T. J. and Tibshirani, R. J. (1990a). Exploring the nature of covariate effects in the proportional hazards models, *Biometrics*, **46**, 1005-1016.

[8] Hastie, T. J. and Tibshirani, R. J. (1990b). *Generalized Additive Models*, Chapman and Hall, London.

[9] Honda, T. (2004). Nonparametric regression in proportional hazards regression, *Journal of the Japan Statistical Society*, **34**, 1-17.

[10] Huang, J. Z., Kooperberg, C., Stone, C. J. and Truong, Y. K. (2000). Functional ANOVA modeling for proportional hazards regression, *The Annals of Statistics*, **28**, 961-999.

[11] Kalbleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, Hoboken.

[12] Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995). The L_2 rates of convergence for hazard regression, *Scandinavian Journal of Statistics*, **22**, 143-157.

[13] Li, G. and Doss, H. (1995). An approach to nonparametric regression for life history data using local linear fitting, *The Annals of Statistics*, **23**, 787-823.

[14] Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models, *Biometrika*, **84**, 469-473.

[15] Linton, O. B. (2000). Efficient estimation of generalized additive nonparametric regression models, *Econometric Theory*, **16**, 502-523.

[16] Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika*, **82**, 93-100.

[17] Linton, O. B., Nielsen, J. P. and van de Geer, S. (2003). Estimating multiplicative and additive hazard functions by kernel regression, *The Annals of Statistics*, **31**, 464-492.

[18] Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *The Annals of Statistics*, **27**, 1443-1490.

[19] Masry, E. (1996). Multivariate local polynomial regression for time series : uniform strong convergence and rates, *Journal of Time Series Analysis*, **17**, 571-599.

[20] Newey, W. K. (1994). Kernel estimation of partial means, *Econometric Theory*, **10**, 233-253.

[21] Nielsen, J. P. and Linton, O. B. (1995). Kernel estimation in a nonparametric marker dependent hazard model, *The Annals of Statistics*, **23**, 1735-1748.

[22] Nielsen, J. P., Linton, O. B. and Bickel, P. J. (1998). On a semiparametric survival model with flexible covariate effect, *The Annals of Statistics*, **26**, 215-241.

[23] Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators, *Journal of Multivariate Analysis*, **73**, 166-179.

[24] Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression, *The Annals of Statistics*, **25**, 186-211.

[25] O'Sullivan, F. (1993). Nonparametric estimation in the Cox model, *The Annals of Statistics*, **21**, 124-145.

[26] Pons, O. (2000). Nonparametric estimation in a varying-coefficient Cox model, *Mathematical Methods of Statistics*, **9**, 376-398.

[27] Schimek, M. G. and Turlach, B. A. (2000). Additive and generalized additive models, *Smoothing and Regression : Approaches, Computation, and Application*(ed. M. G. Schimek), 277-327, Wiley, New York.

[28] Sperlich, S., Tjøstheim, D. and Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models, *Econometric Theory*, **18**, 197-251.

[29] Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion), *The Annals of Statistics*, **25**, 1371-1470.

[30] Tjøstheim, D. and Auestad, B. (1994). Nonparametric identification of nonlinear time series : Projections, *Journal of the American Statistical Association*, **89**, 1398-1409.