

# Local Linear Estimation for Time-Dependent Coefficients in Cox's Regression Models

Zongwu Cai \* and Yanqing Sun  
Department of Mathematics  
University of North Carolina  
Charlotte, NC 28223

October 23, 2000

## Abstract

This article develops a local partial likelihood technique to estimate the time-dependent coefficients in the Cox's regression model. The basic idea is a simple extension of the local linear fitting technique used in scatterplot smoothing. The coefficients are estimated locally based on the partial likelihood in a window around each time point. Multiple time-dependent covariates are incorporated in the local partial likelihood procedure. The procedure is useful as a diagnostic tool and can be used in uncovering time-dependencies or departure from the proportional hazards model. The programming involved in the local partial likelihood estimation is relative simple and it can be modified with few efforts from the existing programs for the proportional hazards model. Asymptotic properties of the resulting estimator are established and a consistent estimator of the asymptotic variance is also proposed. The approach is illustrated by a real data set from the study of gastric cancer patients and a simulation study is also presented.

**Keywords:** Asymptotics, censored data, local linear fitting, local partial likelihood, proportional hazards models, time-dependent covariates, varying-coefficient models.

---

\*Zongwu Cai was supported, in part, by the National Science Foundation grant DMS-0072400 and funds provided by the University of North Carolina at Charlotte.

# 1 Introduction

In survival analysis, of interest is to explore the possible relationship between a survival time  $T$  and a covariate vector  $\mathbf{X} = (X_1, \dots, X_p)^T$ , where  $T$  denotes the transpose of a vector or matrix. In problems involving a possibly censored lifetime, the available data are of the form  $(Y_1, \mathbf{X}_1, \delta_1), \dots, (Y_n, \mathbf{X}_n, \delta_n)$ . The survival time  $Y_i$  is complete if  $\delta_i = 1$  and censored if  $\delta_i = 0$ , and  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$  denotes the usual vector of predictors for the  $i$ th individual. Let  $\lambda(t|\mathbf{x})$  be the conditional hazard function of  $T$  given  $\mathbf{X} = \mathbf{x}$ , defined as

$$\lambda(t|\mathbf{x}) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P\{t \leq T \leq t + \Delta t | T \geq t, \mathbf{X} = \mathbf{x}\}.$$

The proportional hazards model or Cox's regression model assumes the following form (Cox, 1972, 1975)

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp\left(\sum_{j=1}^p a_j x_j\right), \quad (1)$$

where  $\lambda(t|\mathbf{x})$  is the instantaneous rate of failure at time  $t$ , given a particular value  $\mathbf{x} = (x_1, \dots, x_p)^T$  of covariate  $\mathbf{X}$ . The function  $\lambda_0(t)$  is the conditional hazard function of  $T$  given  $\mathbf{X} = \mathbf{0}$ , and it is called the baseline hazard function. Under the model (1), the conditional failure rates associated with any two values of covariate  $\mathbf{X}$  are proportional. The proportional hazards model (1) has been studied extensively by many authors in the literature. The extension to time-dependent covariates are easily dealt with using the counting process and martingale approach (Andersen and Gill, 1982). The detailed parametric and nonparametric inferences can be found in the books by Cox and Oakes (1984), Andersen, Borgan, Gill, and Keiding (1993), Fan and Gijbels (1996), among others.

Various useful alternatives and extensions have been made to enhance the flexibility of the model (1). These include the additive risk models by Cox and Oakes (1984), Huffer and McKeague (1991), McKeague and Sasieni (1994), and Lin and Ying (1994). Lin and Ying (1995) and Scheike and Zhang (2000) considered additive-multiplicative intensity models. Dabrowska (1997) studied a general stratified Cox's regression model. The other models allowing for non-proportionality in the effects of risk factors include the transformed linear models by Pettitt (1982), Bennett (1983a, 1983b), Dabrowska and Doksum (1988), and Cheng, Wei, and Ying (1995) and the treatment efficacy models by Dabrowska, Doksum, Feduska, Husing, and Neville (1992).

Another alternative to make the model (1) more flexible is to allow the coefficients  $a_j$  to depend on a variable  $R_j$ , say, a disease severity score. In particular, the coefficients may change over time  $t$ . Zucker and Karr (1990), Murphy and Sen (1991), Gamerman (1991), Hastie and Tibshirani (1993), Murphy (1993), Marzec and Marzec (1997), Martinussen, Scheike, and Skovgaard (2000), among others, studied the following Cox's regression model with time-dependent coefficients, called the

varying-coefficient model in the literature (Hastie and Tibshirani, 1993),

$$\lambda(t | \mathbf{x}) = \lambda_0(t) \exp \left( \sum_{j=1}^p a_j(t) x_j \right). \quad (2)$$

Unless  $\{a_j(t)\}$  are constant, this model represents non-proportional hazards. When  $\{a_j(t)\}$  are specified as some parametric forms, the model (2) leads to the Cox's time-dependent covariate model. In the special case of a single group variable ( $X = 0$  or  $1$ ), the model (2) reduces to

$$\lambda(t | 0) = \lambda_0(t) \quad \text{and} \quad \lambda(t | 1) = \lambda_0(t) \exp(a(t)).$$

Thus  $a(t)$  measures the logarithm of relative risk between the two groups. The Cox's regression model with time-dependent coefficients reflects some situations in practice where, for example, a treatment effect may change over time. It also provides a rich family for assessing the proportional hazards assumption and can be used to investigate departures.

Instead of estimating directly the coefficient functions  $\{a_j(\cdot)\}$ , Murphy and Sen (1991) proposed a histogram sieve estimator for the cumulative coefficient functions  $B_j(t) = \int_0^t a_j(s) ds$ ,  $j = 1, \dots, p$ , in the model (2) by assuming that the coefficient functions  $\{a_j(\cdot)\}$  are step function and discussed the consistency and asymptotic normality of the resulting estimator. Based on the sieve estimator, Murphy (1993) and Marzec and Marzec (1997) discussed goodness-of-fit tests of the model (2). Gamerman (1991) described the dynamic linear model approach by assuming that  $\lambda_0(t)$  and  $\{a_j(t)\}$  are piecewise constant functions, and constant between the distinct failure times, and using the full likelihood for estimation updating the terms sequentially in time. The assumption that the baseline hazard function and the coefficient functions are step function may not be appropriate in real applications. To overcome this drawback, Zucker and Karr (1990) and Hastie and Tibshirani (1993) used smoothing spline partial likelihood method, leaving  $\lambda_0(t)$  unspecified but modeling the coefficient functions  $\{a_j(t)\}$  smoothly. Zucker and Karr (1990) studied the consistency and asymptotic normality and Hastie and Tibshirani (1993) proposed an algorithm for solving the penalized partial likelihood problem by using iterative strategy. While this idea is powerful, it is quite a task in practice to choose  $p$  smoothing parameters simultaneously and the computation can be challenging. Further, the asymptotic properties are not easily understood due to the iterative algorithm. It is not clear if the resulting method can achieve the optimal rate of convergence.

Recently, Martinussen, Scheike, and Skovgaard (2000) proposed a one-step estimation procedure for the cumulative coefficient functions  $\{B_j(t)\}$ , by using a modified likelihood approach and the one-step Newton-Raphson iterative algorithm. Since the information at any particular time point is limited, not enough for a stable estimate, they instead integrated both sides of the Newton-Raphson iterative equation to obtain an iteration procedure for the cumulative coefficient function based on an initial estimator and its kernel smoothing.

In this article, we develop a local partial likelihood estimation technique, similar to that in Fan, Gijbels, and King (1997) for the estimation of the risk factor in hazard regression, to estimate the time-dependent coefficient functions. The basic idea is a simple extension of the local fitting technique used in scatterplot smoothing coupled with partial likelihood. In a window around each time point  $t$ , approximating the coefficient function  $a_j(s)$  by a linear function using the first-order (or higher order) Taylor expansion, we obtain a partial likelihood estimate for the linear function using the observed failure times within the window. The value of the estimated linear function at  $t$  is the estimate of the smooth coefficient function at  $t$ . Multiple time-dependent covariates are incorporated in the local likelihood procedure. The application of local likelihood estimation technique to the Cox's varying-coefficient model is useful as a diagnostic tool and can be used in uncovering time-dependencies or departure from the proportional hazards model. It is worth noting that the programming involved in the local partial likelihood estimation is relative simple. It can be carried out by modifying the existing software for the proportional hazards model.

This article is organized as follows. In Section 2, we describe the estimation procedure for the model (2). The asymptotic consistency and normality are presented in Section 3 and a consistent estimator of the asymptotic variance is proposed in the same section. In Section 4, we carry out a small scale simulation study and apply the procedure to analyze a data set from the study of gastric cancer patients. Finally, the proofs are relegated to the Appendix.

## 2 Estimation Methods

### 2.1 Local partial maximum likelihood estimation

Suppose that we have an independent censoring scheme, in which the iid censoring times  $C_1, \dots, C_n$  are independent of the iid survival times  $T_1, \dots, T_n$ , given the covariates  $\mathbf{X}_1, \dots, \mathbf{X}_n$ . We observe the censored time  $Y_i = \min(T_i, C_i)$ , the censoring indicator  $\delta_i = I(T_i \leq C_i)$ , and the associated covariate  $\mathbf{X}_i$ . Let  $\{(Y_i, \mathbf{X}_i, \delta_i)\}_{i=1}^n$  be an iid sample from the population  $(Y, \mathbf{X}, \delta)$  that follows the model (2).

Define  $N_i(t) = I(Y_i \leq t, \delta_i = 1)$ , the counting process of observed failures for individual  $i$ , and  $Y_i(t) = I(Y_i \geq t)$ , the "at risk" indicator process. The logarithm of the Cox's partial likelihood is given by

$$l(\mathbf{a}) = \sum_{k=1}^n \int_0^\infty \left[ \mathbf{X}_k(s)^T \mathbf{a}(s) - \log \left\{ \sum_{i=1}^n Y_i(s) \exp(\mathbf{X}_i(s)^T \mathbf{a}(s)) \right\} \right] dN_k(s), \quad (3)$$

where  $\mathbf{a}(t) = (a_1(t), \dots, a_p(t))^T$ .

Assume throughout the paper that the baseline hazard function  $\lambda_0(t)$  is unspecified and continuous, the coefficient functions  $\{a_j(s)\}$  have continuous second derivatives at a neighborhood of

$t$ , and the covariate  $\mathbf{X}(t)$  is a locally bounded predictable process. Then, by a Taylor's expression,

$$a_j(s) \approx a_j + b_j(s - t) \quad (4)$$

for  $s$  in a neighborhood of  $t$ . Note that  $a_j$  and  $b_j$  depend on  $t$ . Let  $\boldsymbol{\beta} = (a_1, \dots, a_p, b_1, \dots, b_p)^T$  and  $\tilde{\mathbf{X}}_i(u, u - t) = \mathbf{X}_i(u) \otimes \begin{pmatrix} 1 \\ u - t \end{pmatrix}$  with  $\otimes$  being the Kronecker product. Also, let  $h = h_n > 0$  be the bandwidth parameter that controls the size of the local neighborhood and let  $K(\cdot)$  be a kernel function that smoothly weights down the contribution of remote data points. Then, under the local model (4), by incorporating the localizing weights, we obtain the *local partial likelihood function*

$$\ell(\boldsymbol{\beta}) = \sum_{k=1}^n \int_0^\infty K_h(u - t) \left[ \tilde{\mathbf{X}}_k(u, u - t)^T \boldsymbol{\beta} - \log \left\{ \sum_{i=1}^n Y_i(s) \exp \left( \tilde{\mathbf{X}}_k(u, u - t)^T \boldsymbol{\beta} \right) \right\} \right] dN_k(u), \quad (5)$$

where  $K_h(\cdot) = K(\cdot/h)/h$ . Let  $\hat{\boldsymbol{\beta}}$  maximize (5) with respect to  $\boldsymbol{\beta}$ . Then, we obtain the *local partial maximum likelihood estimate*  $\hat{\mathbf{a}}(t)$  of  $\mathbf{a}(t)$ , which is the vector consisting of the first  $p$  components of  $\hat{\boldsymbol{\beta}}$ . Note that (5) is motivated by the local partial likelihood method, proposed by Tibshirani and Hastie (1987) and studied by Fan, Gijbels, and King (1997) for estimation of the risk factor  $\psi(\mathbf{x}) = \log\{\lambda(t|\mathbf{x})/\lambda_0(t)\}$ .

## 2.2 Concavity of the local partial likelihood

As in most of the parametric likelihood theory (see, e.g., Lehmann 1983), we only know that there exists a consistent solution to the local partial likelihood equation. But if there are multiple roots, we do not know which solution is consistent. However, if  $\ell(\boldsymbol{\beta})$  is strictly concave, then the solution to (5) is unique and must be consistent.

Let  $G_i(u, \boldsymbol{\beta}) = \exp(\tilde{\mathbf{X}}_i(u, u - t)^T \boldsymbol{\beta})$  and  $G(u, \boldsymbol{\beta}) = \sum_{i=1}^n Y_i(u) G_i(u, \boldsymbol{\beta})$ . Then, (5) can be re-expressed as follows

$$\ell(\boldsymbol{\beta}) = \sum_{k=1}^n \int_0^\infty K_h(u - t) [G_k(u, \boldsymbol{\beta}) - \log \{G(u, \boldsymbol{\beta})\}] dN_k(u),$$

and the Hessian matrix of  $\ell(\boldsymbol{\beta})$  is given by

$$\begin{aligned} \ell''(\boldsymbol{\beta}) &= - \sum_{k=1}^n \int_0^\infty \frac{K_h(u - t)}{G^2(u, \boldsymbol{\beta})} \\ &\quad \left[ G(u, \boldsymbol{\beta}) \sum_{i=1}^n Y_i(u) G_i(u, \boldsymbol{\beta}) \tilde{\mathbf{X}}_i(u, u - t) \tilde{\mathbf{X}}_i(u, u - t)^T - G'(u, \boldsymbol{\beta}) (G'(u, \boldsymbol{\beta}))^T \right] dN_k(u) \\ &= - \sum_{k=1}^n \int_0^\infty \frac{K_h(u - t)}{G^2(u, \boldsymbol{\beta})} \\ &\quad \sum_{i < j} Y_i(u) Y_j(u) G_i(u, \boldsymbol{\beta}) G_j(u, \boldsymbol{\beta}) (\mathbf{X}_i(u) - \mathbf{X}_j(u))^{\otimes 2} \otimes \begin{pmatrix} 1 & u - t \\ u - t & (u - t)^2 \end{pmatrix} dN_k(u), \quad (6) \end{aligned}$$

where  $\mathbf{A}^{\otimes 2}$  denotes  $\mathbf{A} \mathbf{A}^T$  for a vector or matrix  $\mathbf{A}$ . Clearly the right-hand side of (6) is negatively definite as  $n \rightarrow \infty$ . Thus  $\ell(\boldsymbol{\beta})$  is strictly concave.

## 2.3 Estimation of the baseline hazard function

As mentioned in the introduction, the estimation of the baseline hazard function is not a primary goal of this paper. We therefore briefly outline a possible approach to this problem. To this end, we suggest the following estimate for the cumulative baseline hazard function  $\Lambda_0(t) = \int_0^t \lambda_0(u) du$

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{1}{\widehat{S}_{n,0}^*(u)} d\bar{N}(u),$$

where

$$\widehat{S}_{n,0}^*(t) = n^{-1} \sum_{i=1}^n Y_i(t) \exp(\mathbf{X}_i^T \widehat{\mathbf{a}}(t))$$

for some consistent estimator  $\widehat{\mathbf{a}}(t)$  of  $\mathbf{a}(t)$ , and  $\bar{N}(t) = \frac{1}{n} \sum_{i=1}^n I(Y_i \leq t, \delta_i = 1)$ . This estimator is an analogue to the estimator commonly used to estimate the cumulative baseline hazard function in the ordinary Cox's regression model. To estimate  $\lambda_0(t)$  itself, a kernel smoothing technique can then be employed here to obtain an estimate via

$$\widehat{\lambda}_0(t) = \int K_{h_\lambda}^\lambda(u-t) d\widehat{\Lambda}_0(u) = \frac{1}{n h_\lambda} \sum_{k=1}^n \frac{K^\lambda((Y_k - t)/h_\lambda) \delta_k}{\widehat{S}_{n,0}^*(Y_k)},$$

where  $K_{h_\lambda}^\lambda(\cdot) = K^\lambda(\cdot/h_\lambda)/h_\lambda$ ,  $K^\lambda(\cdot)$  is a given kernel function and  $h_\lambda$  is a given bandwidth.

## 3 Asymptotic Properties

### 3.1 Consistency and asymptotic normality

Let  $P(t|\mathbf{x}) = E[I(Y \geq t) | \mathbf{X}(t) = \mathbf{x}]$ ,  $Q_0(t) = E[P(t|\mathbf{X}(t)) \lambda(t|\mathbf{X}(t))]$ ,  $\mathbf{Q}_1(t) = E[P(t|\mathbf{X}(t)) \mathbf{X}(t) \lambda(t|\mathbf{X}(t))]$ , and  $\mathbf{Q}_2(t) = E[P(t|\mathbf{X}(t)) \lambda(t|\mathbf{X}(t)) \mathbf{X}(t)^{\otimes 2}]$ . Denote  $\mu_j = \int u^j K(u) du$  and  $\nu_j = \int u^j K^2(u) du$  for  $0 \leq j \leq 2$ .

Now, we impose some technical conditions but they might not be weakest possible. Let  $N(t, \epsilon)$  be an  $\epsilon$ -neighborhood of  $t$ , for  $\epsilon > 0$  and  $t \geq 0$ .

#### Condition A:

(A.1) The kernel function  $K(\cdot)$  is a bounded density with a bounded support  $[-c, c]$ .

(A.2) There exists a random variable  $V$  such that  $\sup_{u \in N(t, \epsilon)} |\mathbf{X}(u)| \leq V$  and

$$E \left[ \exp \left\{ 2 \left( \sup_{u \in N(t, \epsilon)} |\mathbf{a}(u)| + \mathbf{a}'(t) + 3 \right) V \right\} \right] < \infty.$$

(A.3)  $Q_0(u)$ ,  $\mathbf{Q}_1(u)$ , and  $\mathbf{Q}_2(u)$  are continuous in the neighborhood  $N(t, \epsilon)$ .

(A.4) The sequence  $h \rightarrow 0$  and  $n h \rightarrow \infty$  as  $n \rightarrow \infty$ .

(A.5) Assume that the matrix

$$\boldsymbol{\Sigma}(t) = \frac{1}{Q_0(t)} \left[ Q_0(t) \mathbf{Q}_2(t) - \mathbf{Q}_1(t)^{\otimes 2} \right] \quad (7)$$

is positive definite for all  $t \geq 0$ .

We now state the main results on asymptotic consistency and normality of the proposed estimate  $\widehat{\mathbf{a}}(t)$ . The proofs for the following two theorems are relegated to Appendix.

**THEOREM 1.** *Suppose  $P(u|\mathbf{x}) > 0$  for  $u \in N(t, \epsilon)$ . Under conditions (A.1)–(A.4),  $\widehat{\mathbf{a}}(t) \xrightarrow{P} \mathbf{a}(t)$  as  $n \rightarrow \infty$ .*

**THEOREM 2.** *Suppose that  $P(u|\mathbf{x}) > 0$  for  $u \in N(t, \epsilon)$ . Under condition A, we have*

$$\sqrt{nh} \left[ \widehat{\mathbf{a}}(t) - \mathbf{a}(t) - \frac{h^2 \mu_2}{2} \mathbf{a}''(t) + o_p(h^2) \right] \xrightarrow{D} N \left\{ 0, \nu_0 \boldsymbol{\Sigma}^{-1}(t) \right\},$$

as  $n \rightarrow \infty$ .

*Remark 1.* As a consequence of Theorem 2, the theoretical optimal bandwidth for  $a_j(\cdot)$ , which minimizes the asymptotic weighted mean integrated squared error,

$$\int \left[ \frac{1}{4} h^4 \mu_2^2 \left\{ a_j''(t) \right\}^2 + \frac{\nu_0}{nh} \sigma_{jj}(t) \right] w(t) dt,$$

where  $\sigma_{jj}(t)$  is the  $j$ th diagonal element of  $\boldsymbol{\Sigma}^{-1}(t)$ , is given by

$$h_{opt,j} = \left[ \frac{\nu_0 \int \sigma_{jj}(t) w(t) dt}{\mu_2^2 \int \left\{ a_j''(t) \right\}^2 w(t) dt} \right]^{1/5} n^{-1/5}.$$

*Remark 2.* To compare our results with those in Murphy and Sen (1991) based on the sieve method and Martinussen, Scheike, and Skovgaard (2000) by using one-step approach, we define the smoothing estimator of the coefficient function  $a_j(\cdot)$  based on  $\widehat{B}_j(\cdot)$ , a sieve estimator in Murphy and Sen (1991) or one-step estimator in Martinussen, Scheike, and Skovgaard (2000),

$$\widetilde{a}_j(t) = \int h^{-1} K_h'(u-t) \widehat{B}_j(u) du.$$

If  $\widehat{B}_j(\cdot)$  is the sieve estimator in Murphy and Sen (1991), it can be showed easily from Theorem 2 of Murphy and Sen (1991) and Theorem 2 that our local linear estimator  $\widehat{a}_j(\cdot)$  and the sieve estimator  $\widetilde{a}_j(\cdot)$  share the exact same asymptotic behavior. However, it is interesting to note that if  $\widehat{B}_j(\cdot)$  is the one-step estimator in Martinussen, Scheike, and Skovgaard (2000), the asymptotic bias of  $\widehat{a}_j(\cdot)$  is same as that for  $\widetilde{a}_j(\cdot)$ , but the asymptotic variance is different because  $\widetilde{a}_j(\cdot)$  is based on a modified likelihood.

### 3.2 Estimate of variance

Observe from the first equality of (6) that

$$-\ell''(\boldsymbol{\beta}) = \sum_{k=1}^n \int_0^\infty \frac{K_h(u-t)}{G^2(u, \boldsymbol{\beta})} \left[ G(u, \boldsymbol{\beta}) \sum_{i=1}^n Y_i(u) G_i(u, \boldsymbol{\beta}) \tilde{\mathbf{X}}_i(u, u-t) \tilde{\mathbf{X}}_i(u, u-t)^T - \left( \sum_{i=1}^n Y_i(u) G_i(u, \boldsymbol{\beta}) \tilde{\mathbf{X}}_i(u, u-t) \right)^{\otimes 2} \right] dN_k(u)$$

It is easy to see that the upper left  $p \times p$  block matrix of  $-\ell''(\boldsymbol{\beta})/n$  is equal to

$$\begin{aligned} & \frac{1}{n} \sum_{k=1}^n \int_0^\infty \frac{K_h(u-t)}{G^2(u, \boldsymbol{\beta})} \left[ G(u, \boldsymbol{\beta}) \sum_{i=1}^n Y_i(u) G_i(u, \boldsymbol{\beta}) \mathbf{X}_i(u) \mathbf{X}_i(u)^T - \left\{ \sum_{i=1}^n Y_i(u) G_i(u, \boldsymbol{\beta}) \mathbf{X}_i(u) \right\}^{\otimes 2} \right] dN_k(u) \\ &= \boldsymbol{\Sigma}_1(\mathbf{a}(t)) + O_p(h), \end{aligned}$$

where  $\boldsymbol{\Sigma}_1(\mathbf{a}(t)) = \int K_h(u-t) V(u, \mathbf{a}(t)) d\bar{N}(u)$  with

$$V(u, \mathbf{a}(t)) = \frac{G^{(0)}(u, \mathbf{a}(t)) G^{(2)}(u, \mathbf{a}(t)) - \{G^{(1)}(u, \mathbf{a}(t))\}^{\otimes 2}}{\{G^{(0)}(u, \mathbf{a}(t))\}^2}$$

and

$$G^{(j)}(u, \mathbf{a}(t)) = \sum_{i=1}^n I(Y_i \geq u) \exp\{\mathbf{a}^T(t) \mathbf{X}_i(u)\} \mathbf{X}_i(u)^{\otimes j}.$$

Let  $\widehat{\boldsymbol{\Sigma}}(t) = \boldsymbol{\Sigma}_1(\widehat{\mathbf{a}}(t))$ . We show in the Appendix that  $\widehat{\boldsymbol{\Sigma}}(t)$  is a consistent estimate of the asymptotic covariance  $\boldsymbol{\Sigma}(t)$  given in (7).

## 4 Numerical Examples

In this section, we conduct a numerical study via two simulated examples and one real dataset. The study shows that our procedure for the Cox's regression model with time-dependent coefficients is reliable and useful. For the simulation study, the performance of  $\widehat{a}_j(\cdot)$  is assessed via the mean absolute deviation (MAD)

$$\text{MAD}_j = n_0^{-1} \sum_{k=1}^{n_0} |\widehat{a}_j(u_k) - a_j(u_k)|, \quad j = 1, \dots, p,$$

where  $\{u_k, k = 1, \dots, n_0\}$  are the grid points at which the functions  $a_j(\cdot)$  ( $1 \leq j \leq p$ ) are estimated. Similarly, the performance of the joint estimator  $\widehat{\mathbf{a}}(t)$  is evaluated by

$$\text{MAD} = \sum_{j=1}^p \text{MAD}_j.$$

The Epanechnikov kernel  $K(u) = 0.75(1 - u^2)_+$  is used for all the examples.

## 4.1 Simulated examples

We use two models for the simulation study of finite sample performance of the proposed local partial maximum likelihood estimate, one model with one covariate and the other with two covariates. For the first model (Model I), we consider

$$\lambda(t | X_1) = \lambda_0(t) \exp(a_1(t) X_1),$$

where  $\lambda_0(t) = t^{-1/2}/2$ ,  $a_1(t) = t^{1/2}$ , and the covariate  $X_1$  takes values 0 and 1 with equal probability. Suppose that  $X_1$  is the treatment indicator of two treatments. Then  $a_1(t)$  is the logarithm of the relative risk of one treatment over the other. The second model (Model II) considered is

$$\lambda(t | X_1, X_2) = \exp(a_1(t) X_1 + a_2(t) X_2),$$

where  $\lambda_0(t) = 1$ ,  $a_1(t) = t$ , and  $a_2(t) = 1/2$ . The covariate  $X_1$  is generated from uniform  $(-1, 1)$  and  $X_2$  is from  $N(0, 1)$  independent of  $X_1$ .

Several different sizes of bandwidth are employed to reflect the different levels of smoothness. The censoring times are generated from the uniform distribution  $(0, c)$ , where the parameter  $c$  is adjusted to give approximately 30% of censoring in the situations where censoring is considered. In both models, we restrict to the estimation of the time-varying coefficients on the time interval  $[0, 1]$ . Under 30% censoring, there are about 5% observed failure times greater than 1.0 for Model I and about 11% for Model II. The numerical study is carried out using Fortran 77. The random numbers are generated using RANLIB from the public domain, compiled and written by Barry W. Brown and James Lovato, Department of Biomathematics, The University of Texas. The local partial maximum likelihood estimate is obtained using the Newton-Raphson method.

We carry out the experiment on over all performance (MAD) of the proposed estimation procedure for Model I with the grid points  $0.005 + 0.1k$ ,  $k = 0, 1, \dots, 9$ . We calculate the average MAD and standard deviation of MAD based on 100 repeated samples. Three different sample sizes ( $n = 400, 600, 800$ ), two different censoring patterns (0%, 30%), and several different choices of bandwidth are selected for the calculation. When there is no censoring, bandwidths are  $h = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ , and  $h = 0.6, 0.7, 0.8$  are used for the situation when there is 30% censoring. The simulation results for MAD are summarized in Table 1.

Table 1 about here

Figure 1 gives the estimate of the time-varying coefficient  $a_1(t)$  based on a typical sample from Model I, along with its 95% pointwise confidence band for the sample size  $n = 600, 800$ , bandwidth  $h = 0.3, 0.4, 0.6$  and a 30% censoring. Figure 2 displays the estimate of the time-varying coefficients

$a_1(t)$  and  $a_2(t)$  based on a typical sample for Model II, along with their 95% pointwise confidence bands for the sample size  $n = 600$ , bandwidth  $h = 0.3, 0.4, 0.6$  and a 30% censoring. The estimation results for  $n = 800$  are plotted in Figure 3. All the plots are based on the estimations at 200 grid points with the grid points  $t = 0.005k, k = 1, \dots, 200$ .

Figure 1 to figure 3 about here

## 4.2 A real example

Now, we illustrate our method by applying the procedure to analyze a data set from the study of gastric cancer patients (Stablein, Carter, and Novak, 1981). The data consist of the survival times of 90 patients equally divided in two treatment groups, one group treated with chemotherapy alone and the other one treated with the combined treatment of both chemotherapy and radiation. The survival times in days are observed except for 10 patients whose survival times are censored. Carter, Wampler, and Stablein (1983) fitted the Cox's proportional hazards model with two covariates, treatment indicator  $z_1$  (1 for the combined treatment and 0 for chemotherapy alone) and  $t \cdot z_1$  after changes in treatment effects had been observed. Their modeling using time-changing covariate is equivalent to our modeling of time-varying coefficient in the Cox's regression model with  $a_1(t) = \theta_1 + \theta_2 t$ . Carter *et al.* (1983) calculated the estimates  $\hat{\theta}_1 = 1.2711$  and  $\hat{\theta}_2 = -0.0794$  for  $t$  measured in months. Our nonparametric estimation of  $a_1(t)$  with its 95% pointwise confidence bands are evaluated at 200 grid points  $t = 0.005 + 7.5k, k = 10, \dots, 199$  and are plotted in Figure 4. The bandwidth is taken to be  $h = 400$ . At 1492 days, there are 4 censored survival times with  $z_1 = 1$  and 2 censored and 2 observed with  $z_1 = 0$ . Evidently,  $\hat{a}_1(t)$  (or the logarithm of the risk ratio between the two treatments) changes over time, with  $\hat{a}_1(t) > 0$  for  $t < 420$  days and  $\hat{a}_1(t) < 0$  for  $t > 420$  days. Our 95% pointwise confidence bands for  $a_1(t)$  seems to be consistent with the linear assumption for  $a_1(t)$ . However, our nonparametric estimation for  $a_1(t)$  becomes further above the estimate obtained by Carter *et al.* (1983) using the linear assumption and is close to the estimate of Gamerman (1991) who analyzed the data using a dynamic Bayesian approach; see the graph in Gamerman (1991, p. 72). Notice that the confidence band in Figure 4 is quite wide after about 750 days. This is caused by few number of observations in the region. There are 25 observed survival/censoring times greater than 750, of which 10 are censored. Since there are much more failure/censoring times before 750 days, these early failure/censoring times contribute a lot more to the estimates of  $\theta_1$  and  $\theta_2$ . Hence, the estimate of  $a_1(t)$  under the linear assumption reflects the "global" effect and may undermine the true survival experience in later days, e.g., after 750 days. On the other hand, the nonparametric local partial likelihood method places more weight on the failure/censoring times in the neighborhood of  $t$  at which  $a_1(t)$  is estimated. Therefore, it is more likely to reflect the true survival experience.

Figure 4 about here

## Appendix: Proofs

We use the same notation as in Sections 2 and 3. Let

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & h\mathbf{I}_p \end{pmatrix}$$

and  $\tilde{\mathbf{U}}_i(u, u-t) = \mathbf{H}^{-1} \tilde{\mathbf{X}}_i(u, u-t)$ . We introduce the following notation. Let

$$S_{n,0}(\boldsymbol{\alpha}, u) = n^{-1} \sum_{i=1}^n Y_i(u) \exp \left( \tilde{\mathbf{X}}_i(u, u-t)^T \boldsymbol{\beta} + \tilde{\mathbf{U}}_i(u, u-t)^T \boldsymbol{\alpha} \right)$$

and

$$S_0(\boldsymbol{\alpha}, u) = E \left[ P(u | \mathbf{X}(u)) \exp \left( \tilde{\mathbf{X}}(u, u-t)^T \boldsymbol{\beta} + \tilde{\mathbf{U}}(u, u-t)^T \boldsymbol{\alpha} \right) \right].$$

For  $j = 0$  and 1, set

$$S_{n,j}^*(u) = n^{-1} \sum_{i=1}^n Y_i(u) \exp \left( \mathbf{X}_i(u)^T \mathbf{a}(u) \right) \tilde{\mathbf{U}}_i^j(u, u-t)$$

and

$$S_j^*(u) = E \left[ P(u | \mathbf{X}(u)) \exp \left( \mathbf{X}(u)^T \mathbf{a}(u) \right) \tilde{\mathbf{U}}(u, u-t)^{\otimes j} \right].$$

For  $0 \leq j \leq 2$ , put

$$S_{n,j}(u) = n^{-1} \sum_{i=1}^n Y_i(u) \exp \left( \tilde{\mathbf{X}}_i(u, u-t)^T \boldsymbol{\beta} \right) \tilde{\mathbf{U}}_i(u, u-t)^{\otimes j}$$

and

$$S_j(u) = E \left[ P(u | \mathbf{X}(u)) \exp \left( \tilde{\mathbf{X}}(u, u-t)^T \boldsymbol{\beta} \right) \tilde{\mathbf{U}}(u, u-t)^{\otimes j} \right].$$

Note that  $S_{n,1}^*(u)$ ,  $S_1^*(u)$ ,  $S_{n,1}(u)$  and  $S_1(u)$  are  $2p$ -vectors,  $S_{n,2}(u)$  and  $S_2(u)$  are  $2p \times 2p$  matrices, and the rest are scalar.

Before proceeding with the proof of Theorem 1, we first state a simple lemma that is used throughout this section. The proof is straightforward, and omitted. For detail, see the proof of Lemma 6.1 in Masry and Tjøstheim (1997).

LEMMA 1. *Let*

$$c_n(u) = n^{-1} \sum_{i=1}^n Y_i(u) g(u, \mathbf{X}_i(u)) \quad \text{and} \quad c(u) = E [P(u | \mathbf{X}(u)) g(u, \mathbf{X}(u))].$$

*If  $\sup_{u \in N(t, \varepsilon)} E [g^2(u, \mathbf{X}(u))] < \infty$ , then*

$$\sup_{u \in N(t, \varepsilon)} |c_n(u) - c(u)| = O_p \left( n^{-1/2} \right).$$

## Proof of Theorem 1

Let  $\tilde{\beta}$  be the running parameter in (5). For any fixed  $\beta$  (the true value), let  $\hat{\beta}$  be the maximum likelihood estimator maximizing (5). Let  $\alpha = \mathbf{H}(\tilde{\beta} - \beta)$  and  $\hat{\alpha} = \mathbf{H}(\hat{\beta} - \beta)$ . Then, by (5),  $\hat{\alpha}$  maximizes

$$\begin{aligned} l_n(\alpha) &= \int_0^\infty K_h(u-t) n^{-1} \sum_{i=1}^n \left[ \tilde{\mathbf{X}}_i(u, u-t)^T \beta + \tilde{\mathbf{U}}_i(u, u-t)^T \alpha \right] dN_i(u) \\ &\quad - \int_0^\tau K_h(u-t) \log \{n S_{n,0}(\alpha, u)\} d\bar{N}(u) \end{aligned}$$

with respect to  $\alpha$ . To accomplish the proof of Theorem 1, we prove a somewhat more general result. To this end, for each  $\tau > 0$ , let

$$\begin{aligned} l_n(\alpha, \tau) &= \int_0^\tau K_h(u-t) n^{-1} \sum_{i=1}^n \left[ \tilde{\mathbf{X}}_i(u, u-t)^T \beta + \tilde{\mathbf{U}}_i(u, u-t)^T \alpha \right] dN_i(u) \\ &\quad - \int_0^\tau K_h(u-t) \log \{n S_{n,0}(\alpha, u)\} d\bar{N}(u), \end{aligned}$$

Our case corresponds to that of  $\tau = \infty$ . It is easy to see that

$$\begin{aligned} l_n(\alpha, \tau) - l_n(\mathbf{0}, \tau) &= \int_0^\tau K_h(u-t) n^{-1} \sum_{i=1}^n \tilde{\mathbf{U}}_i(u, u-t)^T \alpha dN_i(u) \\ &\quad - \int_0^\tau K_h(u-t) \log \left\{ \frac{S_{n,0}(\alpha, u)}{S_{n,0}(\mathbf{0}, u)} \right\} d\bar{N}(u). \end{aligned} \quad (\text{A.1})$$

Let the filtration  $\mathcal{F}_{nt}$  be the statistical information accruing during the time  $[0, t]$ , namely,

$$\mathcal{F}_{nt} = \sigma\{\mathbf{X}_i(u), N_i(u), Y_i(u), i = 1, \dots, n, 0 \leq u \leq t\}.$$

Then, under the independent censoring scheme,

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) \lambda(u | \mathbf{X}_i(u)) du \quad (\text{A.2})$$

is an  $\mathcal{F}_{nt}$ -martingale. An substitution of (A.2) into (A.1) gives

$$l_n(\alpha, \tau) - l_n(\mathbf{0}, \tau) = A_n(\alpha, \tau) + X_n(\alpha, \tau), \quad (\text{A.3})$$

where

$$A_n(\alpha, \tau) = \int_0^\tau K_h(u-t) \left[ S_{n,1}^*(u)^T \alpha - \log \left\{ \frac{S_{n,0}(\alpha, u)}{S_{n,0}(\mathbf{0}, u)} \right\} S_{n,0}^*(u) \right] \lambda_0(u) du \quad (\text{A.4})$$

and

$$X_n(\alpha, \tau) = \int_0^\tau K_h(u-t) n^{-1} \sum_{i=1}^n \left[ \tilde{\mathbf{U}}_i(u, u-t)^T \alpha - \log \left\{ \frac{S_{n,0}(\alpha, u)}{S_{n,0}(\mathbf{0}, u)} \right\} \right] dM_i(u). \quad (\text{A.5})$$

Under condition A, by Lemma 1, we have

$$\begin{aligned}
A_n(\boldsymbol{\alpha}, \tau) &= \int_0^\tau K_h(u-t) \left[ S_1^*(u)^T \boldsymbol{\alpha} - \log \left\{ \frac{S_0(\boldsymbol{\alpha}, u)}{S_0(u)} \right\} S_0^*(u) \right] \lambda_0(u) du + o_p(1) \\
&= \mathbf{Q}_1(t)^T \otimes (1, \mu_1) \boldsymbol{\alpha} - Q_0(t) \int \log \left\{ \frac{S(\boldsymbol{\alpha}, t, v)}{Q_0(t)} \right\} K(v) dv + o_p(1), \\
&\equiv A(\boldsymbol{\alpha}, \tau) + o_p(1),
\end{aligned} \tag{A.6}$$

where

$$S(\boldsymbol{\alpha}, t, v) = E \left[ P(t | \mathbf{X}(t)) \lambda(t | \mathbf{X}(t)) \exp \left( (\tilde{\mathbf{X}}(t, v))^T \boldsymbol{\alpha} \right) \right].$$

Using the facts that  $E(Y^2Z) - E(YZ)^2 = E[(Y - E(YZ))^2Z] \geq 0$  for  $Z \geq 0$  and that a matrix  $\mathbf{B}$  being positive definite is equivalent to  $\mathbf{a}^T \mathbf{B} \mathbf{a} \geq 0$  for any column vector  $\mathbf{a}$  of appropriate dimension, it is not difficult to check that  $A(\boldsymbol{\alpha}, \tau)$  is strictly concave, with a maximum at the point  $\boldsymbol{\alpha} = \mathbf{0}$ . The process  $X_n(\boldsymbol{\alpha}, \cdot)$  is a locally square integrable martingale with the predictable variation process

$$\begin{aligned}
C_n(v) &\equiv \langle X_n(\boldsymbol{\alpha}, \cdot), X_n(\boldsymbol{\alpha}, \cdot) \rangle(v) \\
&= n^{-2} \sum_{i=1}^n \int_0^v K_h^2(u-t) \left[ \tilde{\mathbf{U}}_i(u, u-t)^T \boldsymbol{\alpha} - \log \left\{ \frac{S_{n,0}(\boldsymbol{\alpha}, u)}{S_{n,0}(\mathbf{0}, u)} \right\} \right]^2 Y_i(u) \lambda(u | \mathbf{X}_i(u)) du.
\end{aligned}$$

By condition A and Lemma 1, one can show that for any  $0 \leq v \leq \tau$ ,

$$E[X_n(\boldsymbol{\alpha}, v)]^2 = E[C_n(v)] = O((nh)^{-1}) = o(1).$$

This, in conjunction with (A.3) and (A.6), implies that

$$l_n(\boldsymbol{\alpha}, \tau) - l_n(\mathbf{0}, \tau) = A(\boldsymbol{\alpha}, \tau) + o_p(1).$$

Since  $\hat{\boldsymbol{\alpha}}$  maximizes the concave function  $l_n(\boldsymbol{\alpha}, \tau) - l_n(\mathbf{0}, \tau)$ , by the concavity lemma in Appendix II of Andersen and Gill (1982), we have

$$\hat{\boldsymbol{\alpha}} = \mathbf{H} \left( \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \xrightarrow{P} \mathbf{0}.$$

This completes the proof of theorem. □

## Proof of Theorem 2

Let  $\gamma_n = (nh)^{-1/2}$ . Now, we redefine  $\boldsymbol{\alpha} = \gamma_n^{-1} \mathbf{H} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ . Then  $\tilde{\boldsymbol{\beta}} = \gamma_n \mathbf{H}^{-1} \boldsymbol{\alpha} + \boldsymbol{\beta}$ . By following the similar steps as in the proof of Theorem 1, we have

$$l_n(\gamma_n \boldsymbol{\alpha}) = n^{-1} \ell(\gamma_n \mathbf{H}^{-1} \boldsymbol{\alpha} + \boldsymbol{\beta})$$

and

$$l_n(\gamma_n \boldsymbol{\alpha}, \tau) - l_n(\mathbf{0}, \tau) = A_n(\gamma_n \boldsymbol{\alpha}, \tau) + X_n(\gamma_n \boldsymbol{\alpha}, \tau),$$

where  $A_n(\cdot, \tau)$  and  $X_n(\cdot, \tau)$  are defined in (A.4) and (A.5), respectively. By the Taylor's expansion at  $\boldsymbol{\alpha} = \mathbf{0}$ , it follows that

$$\log \left\{ \frac{S_{n,0}(\gamma_n \boldsymbol{\alpha}, u)}{S_{n,0}(\mathbf{0}, u)} \right\} = \frac{S_{n,1}(u)^T \gamma_n \boldsymbol{\alpha}}{S_{n,0}(\mathbf{0}, u)} + \frac{1}{2} \gamma_n^2 \boldsymbol{\alpha}^T \left[ \frac{\mathbf{S}_{n,2}(u)}{S_{n,0}(\mathbf{0}, u)} - \frac{S_{n,1}(u)^{\otimes 2}}{S_{n,0}^2(\mathbf{0}, u)} \right] \boldsymbol{\alpha} + o_p(\gamma_n^2). \quad (\text{A.7})$$

Under condition A, by Lemma 1, for  $|u - t| < ch$ , (A.7) becomes

$$\log \left\{ \frac{S_{n,0}(\gamma_n \boldsymbol{\alpha}, u)}{S_{n,0}(\mathbf{0}, u)} \right\} = \frac{S_1(u)^T \gamma_n \boldsymbol{\alpha}}{S_0(u)} + \frac{1}{2} \gamma_n^2 \boldsymbol{\alpha}^T \left[ \frac{\mathbf{S}_2(u)}{S_0(u)} - \frac{S_1(u)^{\otimes 2}}{S_0^2(u)} \right] \boldsymbol{\alpha} + o_p(\gamma_n^2) \quad (\text{A.8})$$

as  $n \rightarrow \infty$ . Substituting (A.8) into  $A_n(\gamma_n \boldsymbol{\alpha}, \tau)$  given in (A.4) and applying Lemma 1 to  $S_{n,j}^*(u)$  for  $j = 0$  and 1, we have

$$A_n(\gamma_n \boldsymbol{\alpha}, \tau) = \gamma_n A_{n,1}(\tau)^T \boldsymbol{\alpha} - \frac{1}{2} \gamma_n^2 \boldsymbol{\alpha}^T F_{n,1}(\tau) \boldsymbol{\alpha} + o_p(\gamma_n^2),$$

where

$$A_{n,1}(\tau) = \int_0^\tau K_h(u - t) \left[ S_1^*(u) - \frac{S_1(u)}{S_0(u)} S_0^*(u) \right] \lambda_0(u) du$$

and

$$F_{n,1}(\tau) = \int_0^\tau K_h(u - t) \left[ \frac{\mathbf{S}_2(u)}{S_0(u)} - \frac{S_1(u)^{\otimes 2}}{S_0^2(u)} \right] S_0^*(u) \lambda_0(u) du.$$

It follows from condition A and Theorem 1 in Sun (1984) that, for each  $\tau > 0$ , as  $n \rightarrow \infty$ ,

$$F_{n,1}(\tau) - \boldsymbol{\Sigma}(t) \otimes \boldsymbol{\Omega} = o_p(1),$$

where  $\boldsymbol{\Sigma}(t)$  is defined in (7) and  $\boldsymbol{\Omega} = \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix}$ . Therefore,

$$A_n(\gamma_n \boldsymbol{\alpha}, \tau) = \gamma_n A_{n,1}(\tau)^T \boldsymbol{\alpha} - \frac{1}{2} \gamma_n^2 \boldsymbol{\alpha}^T \boldsymbol{\Sigma}(t) \otimes \boldsymbol{\Omega} \boldsymbol{\alpha} + o_p(\gamma_n^2). \quad (\text{A.9})$$

Similarly, substituting (A.8) into  $X_n(\gamma_n \boldsymbol{\alpha}, \tau)$  given in (A.5), we have

$$X_n(\gamma_n \boldsymbol{\alpha}, \tau) = \gamma_n X_{n,1}(\tau)^T \boldsymbol{\alpha} - \frac{1}{2} \gamma_n^2 \boldsymbol{\alpha}^T F_{n,2}(\tau) \boldsymbol{\alpha} + o_p(\gamma_n^2),$$

where

$$X_{n,1}(\tau) = \int_0^\tau K_h(u - t) n^{-1} \sum_{i=1}^n \left[ \tilde{\mathbf{U}}_i(u, u - t) - \frac{S_{n,1}(u)}{S_{n,0}(u)} \right] dM_i(u)$$

and

$$F_{n,2}(\tau) = \int_0^\tau K_h(u - t) \left[ \frac{\mathbf{S}_2(u)}{S_0(u)} - \frac{S_1(u)^{\otimes 2}}{S_0^2(u)} \right] d\bar{M}(u)$$

with  $\bar{M}(t) = (1/n) \sum_{i=1}^n M_i(t)$ . By considering the second moment of  $F_{n,2}(\tau)$  and some simple analysis, we have

$$F_{n,2}(\tau) = O_p(\gamma_n).$$

Therefore,

$$X_n(\gamma_n \boldsymbol{\alpha}, \tau) = \gamma_n X_{n,1}(\tau)^T \boldsymbol{\alpha} + o_p(\gamma_n^2).$$

This, in conjunction with (A.3) and (A.9), implies that

$$l_n(\gamma_n \boldsymbol{\alpha}, \tau) - l_n(\mathbf{0}, \tau) = [A_{n,1}(\tau) + X_{n,1}(\tau)]^T \gamma_n \boldsymbol{\alpha} - \frac{1}{2} \gamma_n^2 \boldsymbol{\alpha}^T \boldsymbol{\Sigma}(t) \otimes \boldsymbol{\Omega} \boldsymbol{\alpha} + o_p(\gamma_n^2).$$

Now, let  $\hat{\boldsymbol{\alpha}} = \gamma_n^{-1} \mathbf{H}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ . Then  $\hat{\boldsymbol{\alpha}}$  maximizes  $l_n(\gamma_n \boldsymbol{\alpha}, \tau)$  with respect to  $\boldsymbol{\alpha}$ . By the quadratic approximation lemma (see, e.g., Fan and Gijbels 1996, p. 210), we obtain

$$\hat{\boldsymbol{\alpha}} = \gamma_n^{-1} (\boldsymbol{\Sigma}(t) \otimes \boldsymbol{\Omega})^{-1} [A_{n,1}(\tau) + X_{n,1}(\tau)] + o_p(1). \quad (\text{A.10})$$

Define

$$\tilde{S}_1(u) = E \left[ P(u | \mathbf{X}(u)) \exp \left( \tilde{\mathbf{X}}(u, u-t)^T \boldsymbol{\beta} \right) \mathbf{X}(u) \right]$$

and

$$\tilde{S}_{n,1}(u) = n^{-1} \sum_{i=1}^n Y_i(u) \exp \left( \tilde{\mathbf{X}}_i(u, u-t)^T \boldsymbol{\beta} \right) \mathbf{X}_i(u).$$

Let

$$A_{n,1}^*(\tau) = \int_0^\tau K_h(u-t) \left[ \mathbf{Q}_1(u) - \tilde{S}_1(u) \lambda_0(u) \frac{S_0^*(u)}{S_0(u)} \right] du \quad (\text{A.11})$$

and

$$X_{n,1}^*(\tau) = \int_0^\tau K_h(u-t) n^{-1} \sum_{i=1}^n \left[ \mathbf{X}_i(u) - \frac{\tilde{S}_{n,1}(u)}{S_{n,0}(u)} \right] dM_i(u).$$

It is clear that  $A_{n,1}(\tau)^T = (A_{n,1}^*(\tau)^T, \mu_1 A_{n,1}^*(\tau)^T)$  and  $X_{n,1}(\tau)^T = (X_{n,1}^*(\tau)^T, \mu_1 X_{n,1}^*(\tau)^T)$ . Since  $(\boldsymbol{\Sigma}(t) \otimes \boldsymbol{\Omega})^{-1} = \boldsymbol{\Sigma}(t)^{-1} \otimes \boldsymbol{\Omega}^{-1}$ , the first  $p$ -components of (A.10) yields

$$\gamma_n^{-1} (\hat{\mathbf{a}}(t) - \mathbf{a}(t)) = \gamma_n^{-1} \boldsymbol{\Sigma}^{-1}(t) [A_{n,1}^*(\tau) + X_{n,1}^*(\tau)] + o_p(1). \quad (\text{A.12})$$

Note that  $\tilde{\mathbf{X}}(u, u-t)^T \boldsymbol{\beta} = \mathbf{a}(t)^T \mathbf{X}(u) + (u-t) \mathbf{a}'(t)^T \mathbf{X}(u)$  and

$$\mathbf{Q}_1(u) - \tilde{S}_1(u) \lambda_0(u) = E \left[ P(u | \mathbf{X}(u)) \lambda_0(u) \mathbf{X}(u) \left( \exp(\mathbf{a}(u)^T \mathbf{X}(u)) - \exp(\tilde{\mathbf{X}}(u, u-t)^T \boldsymbol{\beta}) \right) \right].$$

For  $|u-t| < ch$ ,  $\mathbf{a}(u)^T \mathbf{X}(u) = \mathbf{a}(t)^T \mathbf{X}(u) + (u-t) \mathbf{a}'(t)^T \mathbf{X}(u) + \frac{1}{2} (u-t)^2 \mathbf{a}''(t)^T \mathbf{X}(u) + o_p((u-t)^2)$ .

Thus, by condition A,

$$\begin{aligned} & \mathbf{Q}_1(u) - \tilde{S}_1(u) \lambda_0(u) \\ &= E \left[ P(u | \mathbf{X}(u)) \lambda_0(u) \exp \left( \mathbf{a}(t)^T \mathbf{X}(u) + (u-t) \mathbf{a}'(t)^T \mathbf{X}(u) \right) \frac{1}{2} (u-t)^2 \mathbf{X}(u) \mathbf{X}(u)^T \mathbf{a}''(t) \right] \\ & \quad + o((u-t)^2) \\ &= E \left[ P(t | \mathbf{X}(t)) \lambda_0(t) \exp \left( \mathbf{a}(t)^T \mathbf{X}(t) \right) \mathbf{X}(t) \mathbf{X}(t)^T \right] \frac{1}{2} (u-t)^2 \mathbf{a}''(t) + o((u-t)^2) \\ &= \frac{1}{2} (u-t)^2 \mathbf{Q}_2(t) \mathbf{a}''(t) + o(h^2). \end{aligned}$$

Similarly,

$$\begin{aligned} S_0^*(u) - S_0(u) &= E \left[ P(u|\mathbf{X}(u)) \left( \exp(\mathbf{a}(u)^T \mathbf{X}(u)) - \exp(\tilde{\mathbf{X}}(u, u-t)^T \boldsymbol{\beta}) \right) \right] \\ &= \frac{1}{2} (u-t)^2 \mathbf{Q}_1(t)^T \mathbf{a}''(t) / \lambda_0(t) + o(h^2), \end{aligned}$$

and

$$(S_0^*(u) - S_0(u)) / S_0(u) = \frac{1}{2} (u-t)^2 \mathbf{Q}_1(t)^T \mathbf{a}''(t) / Q_0(t) + o(h^2).$$

Therefore,

$$\mathbf{Q}_1(u) - \tilde{S}_1(u) \lambda_0(u) \frac{S_0^*(u)}{S_0(u)} = \frac{1}{2} (u-t)^2 \boldsymbol{\Sigma}(t) \mathbf{a}''(t) + o_p(h^2). \quad (\text{A.13})$$

Plugging (A.13) in the expression for  $A_{n,1}^*(\tau)$  given in (A.11), (A.12) becomes

$$\gamma_n^{-1} \left[ \hat{\mathbf{a}}(t) - \mathbf{a}(t) - \frac{h^2 \mu_2}{2} \mathbf{a}''(t) + o_p(h^2) \right] = \gamma_n^{-1} \boldsymbol{\Sigma}^{-1}(t) X_{n,1}^*(\tau) + o_p(1),$$

so that

$$\sqrt{nh} \left[ \hat{\mathbf{a}}(t) - \mathbf{a}(t) - \frac{h^2 \mu_2}{2} \mathbf{a}''(t) + o_p(h^2) \right] = \boldsymbol{\Sigma}^{-1}(t) \sqrt{nh} X_{n,1}^*(\tau) + o_p(1).$$

The process  $U_n^*(v) = \sqrt{nh} X_{n,1}^*(v)$  is a locally square integrable martingale with the predictable variation process

$$\langle U_n^*, U_n^* \rangle(v) = n^{-1} h \sum_{i=1}^n \int_0^v K_h^2(u-t) \left[ \mathbf{X}_i(u) - \frac{\tilde{S}_{n,1}(u)}{S_{n,0}(u)} \right]^{\otimes 2} Y_i(u) \lambda(u | \mathbf{X}_i(u)) du.$$

By Lemma 1, one can show that

$$\langle U_n^*, U_n^* \rangle(v) = \int K^2(u) du \left[ Q_0(t) \mathbf{Q}_2(t) - \mathbf{Q}_1(t)^{\otimes 2} \right] / Q_0(t) + o_p(1) = \nu_0 \boldsymbol{\Sigma}(t) + o_p(1).$$

Write the  $l$ th element of the vector  $U_n^*(v)$  as

$$\frac{\sqrt{nh}}{n} \sum_{i=1}^n \int_0^t K_h(u-t) H_{n,i,l}(u) dM_i(u).$$

To prove the asymptotic normality, we need to check the Lindeberg condition

$$\sum_{i=1}^n \int_0^v n^{-1} h K_h^2(u-t) H_{n,i,l}^2(u) I \left\{ \sqrt{h/n} K_h(u-t) |H_{n,i,l}(u)| > \varepsilon \right\} Y_i(u) \lambda(u | \mathbf{X}_i(u)) du \xrightarrow{P} 0$$

for all  $\varepsilon > 0$ . The last statement is valid by condition A and Lemma 1. This establishes that

$$\sqrt{nh} X_{n,1}^*(v) \xrightarrow{D} N \{0, \nu_0 \boldsymbol{\Sigma}(t)\}, \quad 0 \leq v \leq \tau.$$

Therefore,

$$\sqrt{nh} \left[ \hat{\mathbf{a}}(t) - \mathbf{a}(t) - \frac{h^2 \mu_2}{2} \mathbf{a}''(t) + o_p(h^2) \right] \xrightarrow{D} N \{0, \nu_0 \boldsymbol{\Sigma}^{-1}(t)\}.$$

The proof of the theorem is complete.  $\square$

## Proof of consistency of $\widehat{\Sigma}(t)$

Let  $g^{(j)}(u, \mathbf{a}(t)) = E \left[ P(u | \mathbf{X}(u)) \exp(\mathbf{a}^T(t) \mathbf{X}(u)) \mathbf{X}(u)^{\otimes j} \right]$  for  $0 \leq j \leq 2$ . Then  $g^{(j)}(t, \mathbf{a}(t)) = \mathbf{Q}_j(t) / \lambda_0(t)$ . Under conditions (A.1)–(A.4), by Lemma 1, there exists a neighborhood  $\mathcal{B}$  of  $\mathbf{a}(t)$  such that the functions  $g^{(j)}(u, \mathbf{a})$ ,  $0 \leq j \leq 2$ , are continuous functions of  $\mathbf{a} \in \mathcal{B}$ , uniformly in  $u \in N(t, \epsilon)$ ,  $g^{(0)}(u, \mathbf{a}(t))$  is bounded away from zero on  $(u, \mathbf{a}) \in N(t, \epsilon) \times \mathcal{B}$ , and  $g^{(0)}(u, \mathbf{a}(u))$  is continuous on  $u \in N(t, \epsilon)$ . Further, for  $0 \leq j \leq 2$ ,

$$\sup_{\mathbf{a} \in \mathcal{B}, |u-t| < ch} \left\| n^{-1} G^{(j)}(u, \mathbf{a}) - g^{(j)}(u, \mathbf{a}) \right\| \xrightarrow{P} 0$$

as  $n \rightarrow \infty$ . Let

$$v(u, \mathbf{a}(t)) = \frac{g^{(0)}(u, \mathbf{a}(t))g^{(2)}(u, \mathbf{a}(t)) - \{g^{(1)}(u, \mathbf{a}(t))\}^{\otimes 2}}{\{g^{(0)}(u, \mathbf{a}(t))\}^2}.$$

Then, by (7),

$$\begin{aligned} \Sigma(t) &= v(t, \mathbf{a}(t))\lambda_0(t)g^{(0)}(t, \mathbf{a}(t)) \\ &= \int K_h(u-t)v(u, \mathbf{a}(t))\lambda_0(u)g^{(0)}(u, \mathbf{a}(u)) du + o(1). \end{aligned}$$

Now, subtracting and adding some appropriate terms and using triangle inequality, we have

$$\begin{aligned} &\left\| \widehat{\Sigma}(t) - \Sigma(t) \right\| \\ &\leq \left\| \int K_h(u-t)(V(u, \widehat{\mathbf{a}}(t)) - v(u, \widehat{\mathbf{a}}(t))) d\bar{N}(u) \right\| \\ &+ \left\| \int K_h(u-t)(v(u, \widehat{\mathbf{a}}(t)) - v(u, \mathbf{a}(t))) d\bar{N}(u) \right\| \\ &+ \left\| \int K_h(u-t)v(u, \mathbf{a}(t))\{d\bar{N}(u) - \lambda_0(u)n^{-1}G^{(0)}(u, \mathbf{a}(u)) du\} \right\| \\ &+ \left\| \int K_h(u-t)v(u, \mathbf{a}(t))\{n^{-1}G^{(0)}(u, \mathbf{a}(u)) - g^{(0)}(u, \mathbf{a}(u))\}\lambda_0(u) du \right\| \\ &+ \left\| \int K_h(u-t)v(u, \mathbf{a}(t))g^{(0)}(u, \mathbf{a}(u))\lambda_0(u) du - v(t, \mathbf{a}(t))\lambda_0(t)g^{(0)}(t, \mathbf{a}(t)) \right\|. \end{aligned}$$

Following the steps in the last part of the proof of Theorem 3.2 of Andersen and Gill (1982), we have  $\left\| \widehat{\Sigma}(t) - \Sigma(t) \right\| \xrightarrow{P} 0$ .

## References

- Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, **10**, 1100–1120.
- Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
- Bennett, S. (1983a). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, **2**, 273–277.

- Bennett, S. (1983b). Log-logistic regression models for survival data. *Applied Statistics*, **32**, 165–171.
- Carter, W.H., Wampler, G.L., and Stablein, D.M. (1983). *Regression Analysis of Survival Data in Cancer Chemotherapy*. Dekker, New York.
- Cheng, S.C., Wei, L.J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, **82**, 835–845.
- Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.
- Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.
- Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.
- Dabrowska, D.M. and Doksum, K.A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics*, **15**, 1–24.
- Dabrowska, D.M., Doksum, K.A., Feduska, N.J., Husing, R., and Neville, P. (1992). Methods for comparing cumulative hazard functions in a semi-proportional hazard model. *Statistics in Medicine*, **11**, 1465–1476.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.
- Fan, J., Gijbels, I., and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, **25**, 1661–1690.
- Gamerman, D. (1991). Dynamic Bayesian methods for survival data. *Applied Statistics*, **40**, 63–79.
- Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B*, **55**, 757–796.
- Huffer, F.W. and McKeague, I.W. (1991). Weighted least squares estimation for Aalen’s additive risk model. *Journal of the American Statistical Association*, **86**, 114–129.
- Lehmann, E.L. (1983). *Theory of Point Estimation*. Wadsworth & Brooks/Cole, Pacific Grove, California.
- Lin D.Y. and Ying Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.
- Lin, D.Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative intensity for counting processes. *The Annals of Statistics*, **23**, 1712–1734.
- Martinussen, T., Scheike, T.H., and Skovgaard, I.M. (2000). Efficient estimation of fixed and time-varying covariates effects in multiplicative intensity models. *Unpublished manuscript*.
- Marzec, L. and Marzec, P. (1997). On fitting Cox’s regression model with time-dependent coefficients. *Biometrika*, **84**, 901–908.
- Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**, 214–252.

- McKeague, I.W. and Sasieni, P.D. (1994). A partly parametric additive risk model. *Biometrika*, **81**, 501–514.
- Murphy, S.A. (1993). Testing for a time dependent coefficient in Cox’s regression model. *Scandinavian Journal of Statistics*, **20**, 35–50.
- Murphy, S.A. and Sen, P.K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and Their Applications*, **39**, 153–180.
- Pettitt, A.N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society, Series B*, **44**, 234–243.
- Scheike, T.H. and Zhang, M. (2000). An additive-multiplicative Cox-Aalen regression model. *Unpublished manuscript*.
- Stablein, D.M., Carter, W.H., and Novak, J.W. (1981). Analysis of survival data with non-proportional hazard functions. *Controlled Clinical Trials*, **2**, 149–159.
- Sun, Z. (1984). Asymptotic unbiased and strong consistency for density function estimator. *Acta Math Sinica*, **27**, 769–782.
- Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**, 559–567.
- Zucker, D.M and Karr, A.F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics*, **18**, 329–353.

Table 1. Simulation results for MAD based on 100 replicates for Model I.

censoring	$n$	$h$	Av MAD	Std MAD
0%	400	.4	.2067	.1773
		.5	.1713	.1329
		.6	.1704	.1315
		.7	.1551	.1180
		.8	.1526	.1132
	600	.4	.1663	.1287
		.5	.1396	.1091
		.6	.1385	.1127
		.7	.1432	.1061
		.8	.1170	.0881
	800	.4	.1437	.1076
		.5	.1300	.0951
		.6	.1218	.0894
		.7	.0977	.0789
		.8	.1021	.0898
	30%	400	.6	.2378
.7			.2349	.1776
.8			.2395	.1628
600		.6	.1968	.1507
		.7	.1980	.1541
		.8	.1841	.1481
800		.6	.1676	.1376
		.7	.1753	.1383
		.8	.1547	.1178

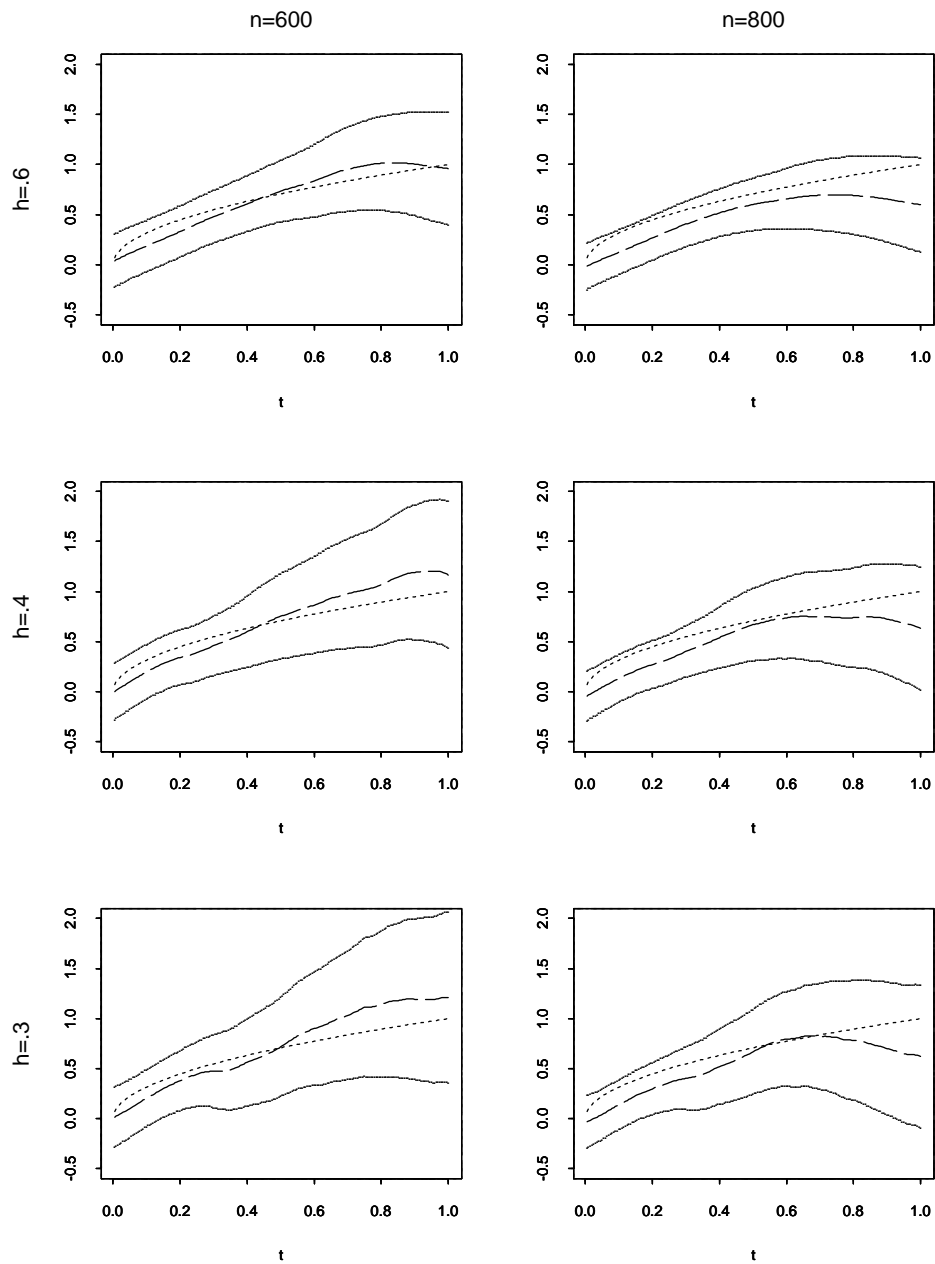


Figure 1. Estimation of  $a_1(t)$  for Model I with  $n = 600, 800$  and 30% censoring for Model I. The dashed line is the estimate of  $a_1(t)$ , the solid lines are the estimates of  $a_1(t)$  plus/minus 1.96 times the estimate of its standard deviation, and the dotted line is the true function.

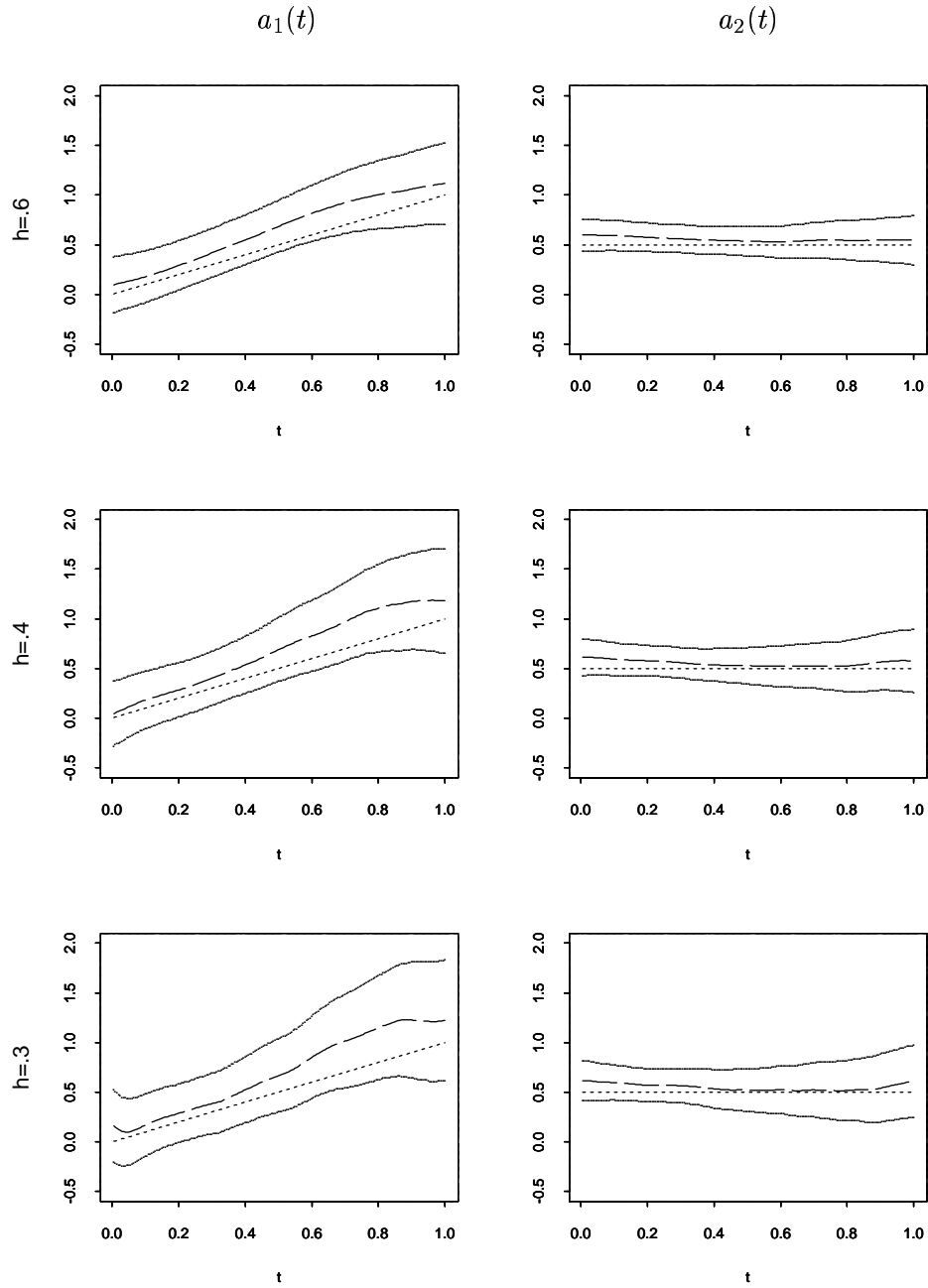


Figure 2. Estimation of  $a_1(t)$  and  $a_2(t)$  for Model II with  $n = 600$  and 30% censoring. The dashed lines are the estimates of  $a_i(t)$ ,  $i = 1, 2$ , the solid lines are the estimates of  $a_i(t)$  plus/minus 1.96 times the estimate of its standard deviation, and the dotted lines are the true functions.

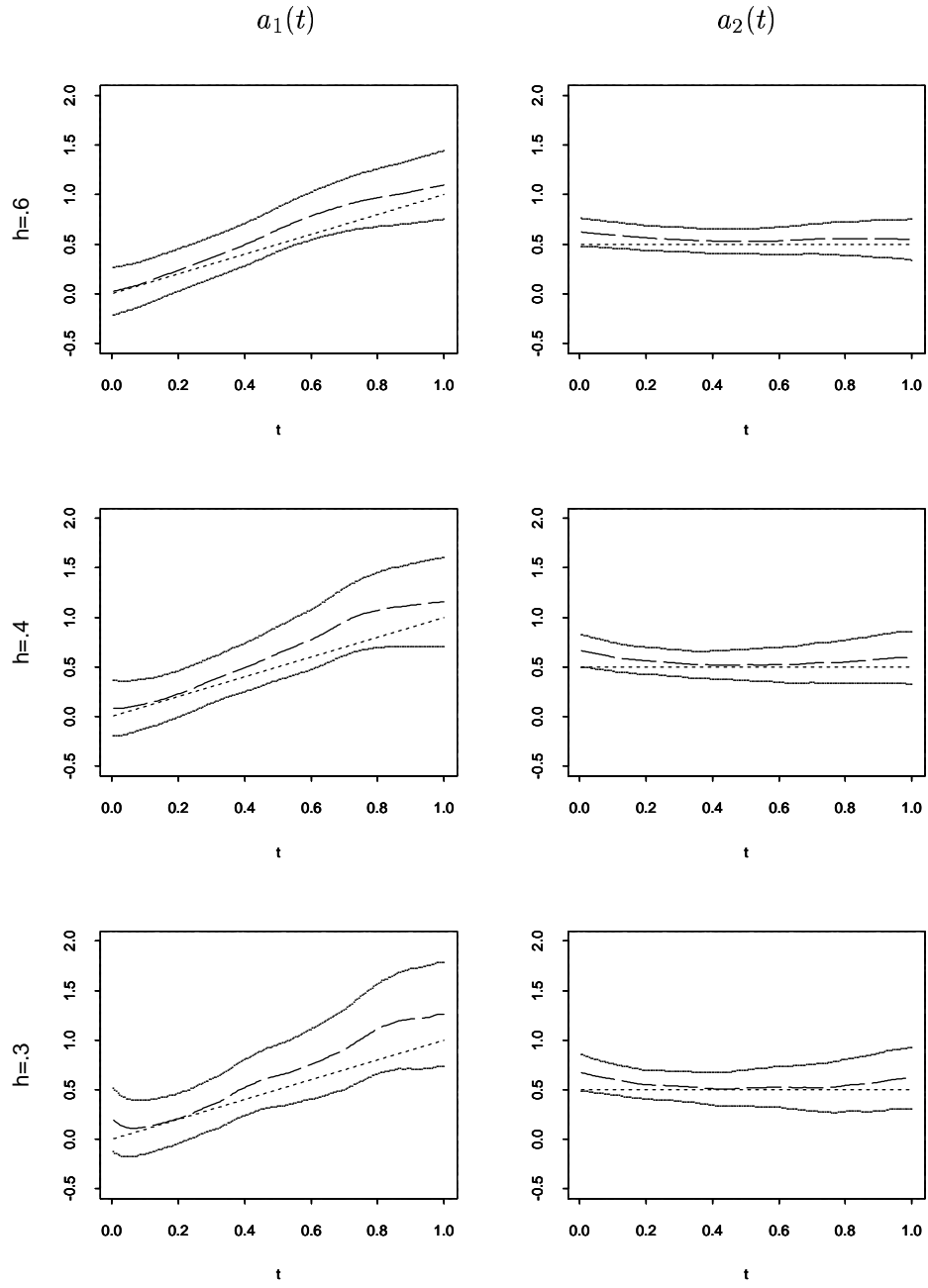


Figure 3. Estimation of  $a_1(t)$  and  $a_2(t)$  for Model II with  $n = 800$ .

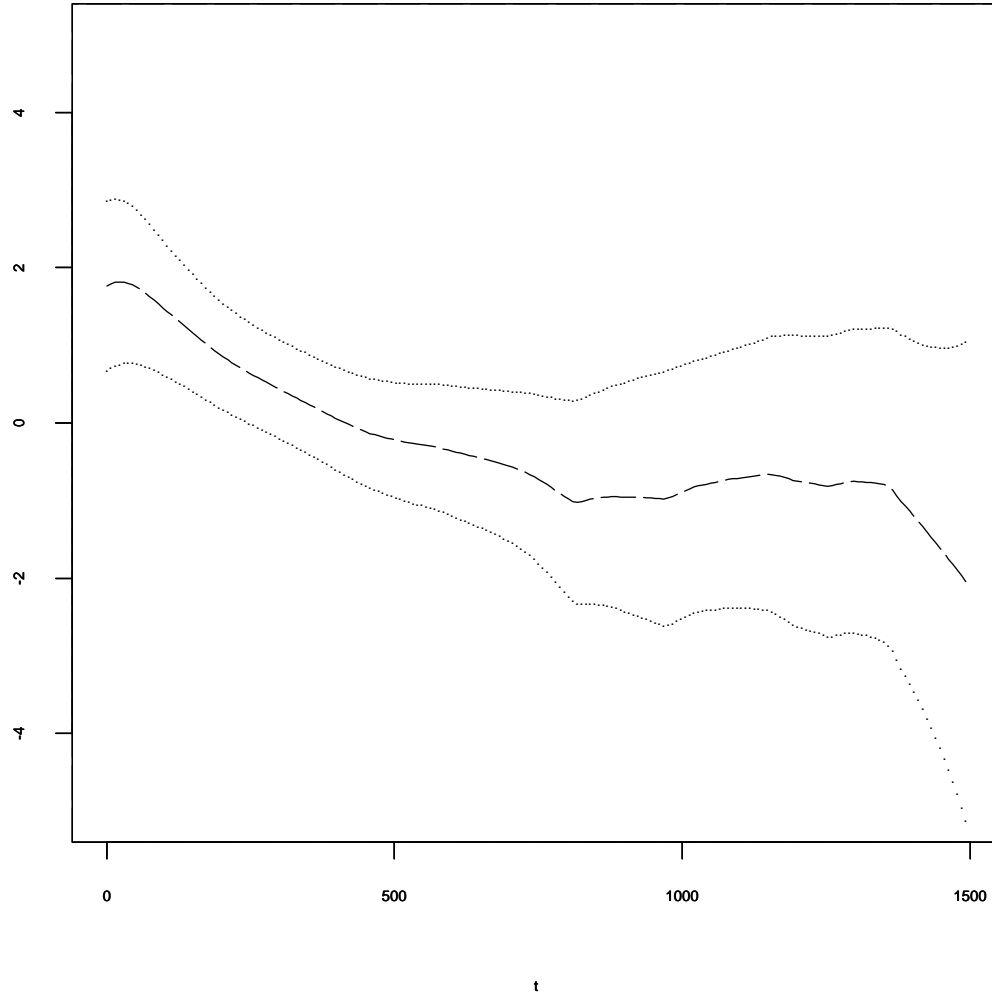


Figure 4. Estimation of  $a_1(t)$  for gastric cancer data with its 95% confidence bands. The dashed line is the estimate of  $a_1(t)$  and the dotted lines are the estimates of  $a_1(t)$  plus/minus 1.96 times the estimate of its standard deviation.